Bot-based collective blocklists in Twitter: The counterpublic moderation of harassment in a networked public space

R. Stuart Geiger
Berkeley Institute for Data Science, University of California, Berkeley, CA, USA
stuart@stuartgeiger.com

Information, Communication, and Society 19(6).

This is an author's accepted manuscript. The official version of this article can be found at http://www.tandfonline.com/doi/abs/10.1080/1369118X.2016.1153700 and a blog post summarizing this article can be found at https://bids.berkeley.edu/news/moderating-harassment-twitter-blockbots

Abstract

This article introduces and discusses bot-based collective blocklists (or blockbots) in Twitter, which have been developed by volunteers to combat harassment in the social networking site. Blockbots support the curation of a shared blocklist of accounts, where subscribers to a blockbot will not receive any notifications or messages from those on the blocklist. Blockbots support counterpublic communities, helping people moderate their own experiences of a site. This article provides an introduction and overview of blockbots and the issues that they raise about networked publics and platform governance, extending an intersecting literature on online harassment, platform governance, and the politics of algorithms. Such projects involve a far more reflective, intentional, transparent, collaborative, and decentralized way of using algorithmic systems to respond to issues like harassment. I argue that blockbots are not just technical solutions but social ones as well, a notable exception to common technologically determinist solutions that often push responsibility for issues like harassment to the individual user. Beyond the case of Twitter, blockbots call our attention to collective, bottom-up modes of computationally assisted moderation that can be deployed by counterpublic groups who want to participate in networked publics where hegemonic and exclusionary practices are increasingly prevalent.

KEYWORDS: Harassment; social media; algorithms; moderation; public sphere; networked publics

Suggested Citation

Geiger, R. Stuart. (2016). "Bot-based collective blocklists in Twitter: The counterpublic moderation of harassment in a networked public space." *Information, Communication, and Society 19(6)*. http://www.stuartgeiger.com/blockbot-ics.pdf

1. Introduction and overview

1.1. Harassment as an issue of platform governance

Harassment has long been an issue in online spaces, particularly gender-based harassment (Dibbell, 1993; Herring, 1999), which is prevalent across many private and public online sites and platforms (Marwick & Miller, 2014; Matias et al., 2015). A 2014 Pew survey of Americans found that 73% of adult American internet users had witnessed harassment online, and 40% had personally been harassed. (Duggan, 2014). Coordinated harassment campaigns are increasingly organized by more-and-less organized groups, who synchronously flood a target's social media feeds (Heron, Belford, & Goker, 2014; Phillips, 2015). Given the chilling effects that harassment has, it is a compelling civil rights issue about the capacity for people to participate in public spaces (Citron, 2014). Harassment is also especially prevalent on the popular social networking and microblogging site Twitter compared to many competing platforms. Twitter, Inc. has generally taken a far more hands-off approach to moderation than other social networking sites, and the design of the platform affords unsolicited interactions in ways that others do not.

In this article, I discuss these institutional and infrastructural dimensions of harassment on Twitter. Lawrence Lessig famously declared ‘code is law' (Lessig, 1999), speaking to the quasi-governmental authority that systems administrators have over their digital domains and those who inhabit it. Staff at Twitter, Inc. are responsible for not only developing and enforcing the site's rules, but also designing the site in a way that affords and constrains particular kinds of activities – just as in other centrally hosted sites like Facebook, YouTube, reddit, World of Warcraft, LinkedIn, or Pinterest. In academic research and popular commentary, there is an increasing concern with such issues of platform governance and ‘the politics of platforms' (e.g., Crawford & Gillespie, 2014; Gillespie, 2010; Morozov, 2013; Pariser, 2012). Platform owner/operators are

responsible for constituting ‘networked publics' (boyd, 2010), which are built and administered in ways that support specific ideas about what it means for people to interact in a shared space. For example, Massanari's analysis of reddit critiques the technological affordances, governance structure, and cultural habitus of the news and discussion site, which she argues intersects in ways that ‘provide fertile ground for anti-feminist and misogynistic activism' (Massanari, 2015, p. 1). When discussing harassment as an issue of platform governance, software is crucial to keep in view, given how the work of moderation, filtering, and gatekeeping is increasingly reliant on highly automated algorithmic systems (Diakopoulos, 2015; Gillespie, 2014; Tufekci, 2014).

However, this article is not about the computationally supported moderation work performed directly by Twitter, Inc. staff. Nor is it about the relationship between this ‘server sovereign’ and governmental agencies – which a long lineage of online harassment scholarship has investigated (Citron, 2014; Marwick & Miller, 2014; Tokunaga, 2010). Instead, I expand the literature on online harassment and platform governance by examining bot-based collective blocklists (or ‘blockbots'), which support a novel form of user-generated, bottom-up, computationally supported moderation work. Blockbots are developed and used by independent, volunteer users of Twitter, who have developed their own computational tools to help individually and collectively moderate their own experiences on Twitter. Functionally, blockbots work similarly to ad blockers: people can curate lists of accounts they do not wish to encounter (individually, in groups, or using algorithmic procedures), and others can subscribe to these blocklists with just a few clicks. They are part of a growing genre of bespoke code (Geiger, 2014), which runs alongside existing platforms, rather than being directly integrated into server-side code.

Blockbots extend the affordances of the site to make the work of responding to harassment more efficient and more communal. Blockbots extend the basic functionality of individual

blocking, in which users can hide specific accounts from their experience of the site. Blocking has long been directly integrated into Twitter's user interfaces, which is necessary because by default, any user on Twitter can send tweets and notifications to any other user. However, individually blocking accounts can be a Sisyphean task for those targeted by harassers, which is why a subscription model has emerged. Some blockbots use lists of harassers curated by a single person, others use community-curated blocklists, and a final set use algorithmically generated blocklists. Blockbots first emerged in Twitter in 2012, and since then, the computational infrastructure has been standardized with the development of the BlockTogether.org service, which makes it easy and efficient for blocklist curators and subscribers to connect. While some blockbot projects are more private, others are open source and highly public, with their developers and maintainers widely publicizing their blocklists, bot code, and blocking procedures. These include Randi Harper's ggautoblocker (http://twitter.com/ggautoblocker), which has over 3000 subscribers as of mid-2015 and has received significant coverage in social and mainstream media.

While blockbots were initially developed to combat harassment in Twitter, they can be (and are) used to filter accounts for a variety of reasons beyond harassment, including incivility, hate speech, trolling, and other related phenomena. It is important to note that just because an account is added to a blocklist does not mean the account has engaged in harassment according to a legal definition of the term (or even at all). This aspect of blockbots has led to controversies about what it means to put an account on a blocklist.

1.2. Overview of the article

This article has three core sections beyond this introduction; in these sections I discuss relevant literature on this topic, then the operation and impacts of blockbots for the groups that use

them, and finally the acts of collective sensemaking that take place through blockbots. In Section 2, I review several related strands of literature around online harassment, platform governance, and the public sphere. It is important to understand harassment as a civil rights issue of governance that touches on what kind of public a site like Twitter ought to be, rather than seeing harassment as a series of isolated, independent events. I introduce Nancy Fraser's concept of 'counterpublics,' which are the non-dominant groups that are often excluded from fully participating in public forums and venues. The concept of counterpublics is useful in helping theorize online harassment and blockbot projects, as Fraser discusses the ways in which non-dominant groups rely on 'parallel discursive arenas' where they can interact on their own terms. Yet blockbots involve a different mode of counterpublic agency, as they are an intentional refusal to listen to those who have been identified as harassers, which is supported through this community-built infrastructure.

In Section 3, I discuss the history and operation of blockbots. I show how they are projects in which counterpublic groups exercise agency over their own experiences on Twitter. Blockbots emerged in the context of a perceived governance gap, when anti-harassment activists and advocates were arguing that Twitter, Inc. was not doing enough to prevent harassment. I discuss how blockbots facilitate disciplines of listening (Crawford,2009), which give subscribers more agency in selectively 'tuning in' to affective publics on Twitter (Papacharissi, 2014). Finally, in the fourth section, I focus on how blockbots serve as sites for collective counterpublic sensemaking about harassment. I discuss cases of redesign and reconfiguration: moments when the functionality of blockbots was changed based on new ideas about what harassment is, how it ought to be identified, and what kinds of experiences a community wanted to have in a privately owned and operated public space. These moments are socio-technical reconfigurations (Suchman, 2007),

which show how blockbots should not be seen as ‘technological solutionist' (Morozov, 2013) projects that shift the burden of responsibility for dealing with harassment from platform owner/operators to isolated individuals. Rather, they are closer to the ‘recursive publics' that Kelty (2008) describes with open source software projects, where a community's development of infrastructure is deeply entwined with their development of values and ideals.

1.3. Methodology

This article is an initial overview and discussion of blockbots in Twitter, intended to introduce this phenomenon and discuss the implications they raise for networked publics. This article's contributions operate at a more theoretical level, as I use empirical cases to discuss particular issues about harassment and platform governance. The cases I discuss are ones I have encountered while conducting an ongoing ethnography of infrastructure (Star, 1999). I have been studying the role of automated software agents in the governance of several networked publics, including Twitter, Wikipedia, and reddit. I focused on these infrastructures as dynamic and relational, ‘emerg[ing] for people in practice, connected to activities and structures' (Bowker, Baker, Millerand, & Ribes, 2010 , p. 99). I sought moments of controversy and breakdown to make structures, norms, discourses, and invisible work (Star & Strauss, 1999) visible and comparable. As much of the work of developing blockbots takes place online, much of my analysis involved analyzing and reconstructing the trace data that members of these communities rely on to coordinate their work (Geiger & Ribes, 2011). I did not directly participate in the development of blockbots in Twitter, but I did subscribe to multiple blockbots, regularly observed public community discussions and use cases about open source blockbots for twelve months, and developed several bots for other non-harassment purposes in Twitter.

I also relied on interviews with the developers of major blockbots to understand and contextualize blockbot development. This included both interviews I conducted personally and transcripts of interviews conducted by journalists, activists, and academics (Fleishman, 2014; Hess, 2014; Schwartz, Lynch, & Harper, 2015), as this issue has gained attention in certain media outlets. This work, combined with my years of experience participating in open source software and free culture projects at various levels, has helped me better understand the work of blockbot development and has better equipped me to follow and interpret the trace data that is accessible on forums, code repositories, Twitter, and other sites. In analyzing this data, I took an iterative and inductive approach to combine many methods, including interviews with blockbot developers, observations of routine and exceptional activity in and around blockbots, archival and historical methods using publicly accessible records, and systems analysis and software studies methods. I have also worked to anonymize the cases of blockbots as much as possible due to the sensitive nature of this topic, except for those individuals who widely publicize their own projects and gave me explicit permission to discuss their bots in my publications.

2. Literature review: harassment, technological determinism, and networked publics

2.1. Harassment as a civil rights issue about participation in networked publics

Online harassment is a complex issue, and I discuss it as an issue about the governance of networked publics, specifically focusing on how privately owned public spaces are moderated – and by whom. Citron, in *Hate crimes in cyberspace*, (2014), defines cyber harassment as 'intentional infliction of substantial emotional distress accomplished by online speech that is persistent enough to amount to a 'course of conduct' rather than an isolated incident,' which is typically carried out through 'threats of violence, privacy invasions, reputation-harming lies, calls

for strangers to physically harm victims, and technological attacks' (p. 3). Citron responds to claims about ‘free speech' by arguing that harassment is a civil rights issue, given how harassment works to inhibit participation. While traditional free speech frameworks seek to minimize restrictions on speech by an established authority, Citron argues that we must also prioritize people's ability to freely express themselves in public without fear or coercion, regardless of the source. If harassment is considered protected under principles of free speech, this trades off with the free speech rights of those targeted. Citron reviews many cases and studies about how harassment works to silence targets into withdrawing from public spaces – in fact, this is frequently the goal of harassers. Such a situation is not unique to online media. Nancy Fraser extensively discusses how chilling effects played out in the coffeehouses and other forums of early modern Europe in her critiques of Habermas's influential account of the public sphere (Fraser, 1990).

2.2. *The Californian ideology, individual vs. collective ethics, and technological solutionism*

Online harassment has many causes, but the current state of harassment online is in part a product of a particular technologically determinist, libertarian mindset that is prominent in Silicon Valley, which Barbrook and Cameron identified as ‘the Californian Ideology' (1996). The modern Internet is the product of both countercultural and libertarian activists, who were concerned with censorship from traditional governments and corporations (Turner, 2006). Activists like John Perry Barlow of the Electronic Frontier Foundation wrote tracts like ‘A Declaration of the Independence of Cyberspace' (1996), celebrating the technological principles that made online interactions ‘immune' to more traditional forms of regulation. Texts expressing this ideology even frequently included pro-inclusion and diversity statements, imagining that the Internet's inherent resistance to state censorship would lead to ‘a world where anyone, anywhere may express his or

her beliefs, no matter how singular, without fear of being coerced into silence or conformity,' as Barlow wrote. Yet this understanding of a mediated public is based on an ideal of the isolated individual, whose classical liberal rights to freedom of expression are to be supported by technology. As Adam critiques in her work on computer ethics (Adam, 2005), when ethics does come on the scene among technologists, it often is discussed in a way that assumes 'individual, rationalistic, rule-based ethical models' (p. 38) like utilitarianism, which align with technologically determinist principles. In contrast, Adam argues for feminist ethics of collectivity and care that focus on the structural inequalities of marginalized groups. She argues that 'despite holding a rhetoric of equality and participation,' the standard utilitarian, individualistic, allegedly meritocratic ethics common among technologists 'often make no challenge to the structures that are causing that inequality in the first place'(Adam, 2000, p. 2).

The individualism issue also intersects with another mindset prevalent in Silicon Valley, which sees these problems as technological ones requiring technological solutions. Evgeny Morozov has recently termed this mindset 'technological solutionism' (Morozov, 2013), and the search for technical solutions to societal problems is also an aspect of the Californian ideology Barbook and Cameron identified. Morozov, Adam, and Barbrook and Cameron – each writing at different times about the same dominant mindset – concur that one of the core problems with such technological solutions is that they are often based in an autonomous individualist mindset. Technology is often used to shift the burden of solving these problems to the individual, frequently assuming that having such a responsibility is empowering. In the issue of harassment, this can be seen in the rise of muting, blocking, flagging, and reporting features, which Crawford and Gillespie (2014) critique on these grounds. They advocate 'a more social mechanism for a more social problem' (p. 12), looking to the open backstage model of Wikipedia, where people debate cases

and policies in public. Ultimately, they conclude that individualistic mechanisms like ‘flags may be structurally insufficient to serve the platforms’ obligations to public discourse, failing to contend with the deeper tensions at work' (p. 15).

2.3. The WAM authorized reporter partnership

One kind of response that takes a more communitarian, discursive, and civic-focused approach to platform governance is the recent pilot project between Twitter, Inc. and the non-profit Women, Action, and the Media (or WAM), analyzed and described by (Matias et al., 2015). In that three-week pilot, WAM was granted an ‘authorized reporter' status, where Twitter users could report harassment to WAM, instead of the default flagging mechanism that sends reports to Twitter, Inc.’s internal processes. WAM reviewers were an intermediary between reporters of harassment and Twitter, Inc.’s somewhat black-boxed team of humans and algorithms that enforce the site's rules. WAM reviewers would evaluate and discuss these reports, sometimes working with the reporter to further document harassment in cases of incomplete or ambiguous reports. If WAM reviewers decided that a report was appropriate to escalate to Twitter, Inc. staff, they would submit it to a specialized ticketing system. The WAM team would interact with Twitter, Inc. staff as necessary on the reporter's behalf, answering questions and advocating if necessary.

In the three-week pilot period, WAM received 811 reports and decided to escalate 161 of them. Of those 161 reports, Twitter, Inc. took action on 55%, suspending 70 accounts, issuing 18 warnings, and deleting 1 account. An independent academic study (Matias et al., 2015) argued that the project was a success in piloting an alternative to the often opaque processes and policies around harassment in major web platforms. They argued that it was important to build a more communal way of responding to harassment, finding that many targets of harassment needed

different kinds of support in making sense of harassment, particularly given how harassers can use sophisticated techniques to overwhelm their targets and mask their own identity. Only 43% of reports came directly from those targeted by harassment, with the majority coming from a target's authorized delegate or a bystander who observed the harassment – showing a key flaw with taking an individualist approach to harassment. However, one issue with this model is that it requires a substantial amount of labor performed by communities or non-profit organizations, which ultimately benefits the for-profit corporation Twitter, Inc. This raises similar ethical concerns that scholars of microwork platforms have raised about the use of crowdworkers (Irani, 2013).

2.4. Theorizing publics and counterpublics

WAM's partnership with Twitter touches on an issue about how non-dominant groups assemble in response to hegemonic venues, where they are often excluded from participating on their own terms. In the next section, I discuss this further by turning to Fraser's (1990) feminist critique of Habermas's account of the public sphere (Habermas, 1989), in which she theorizes how non-dominant groups form 'counterpublics.' Habermas's influential account of the bourgeois public sphere celebrated the coffeehouses, newspapers, and other forums where 'members of the public' could rationally debate socio-political issues and build consensus. His account of early modern publics depicts spaces where anyone can enter, bracketing their own social status to engage with others as equal peers in deliberation. Fraser notes that these were highly exclusionary spaces, particularly for women, persons of color, and members of the working classes. Yet even when marginalized individuals were not officially restricted from participating in these spaces, the fiction that the space was a neutral one and that all participants were equal often served to make subtler forms of exclusion less visible: 'such bracketing usually works to the advantage of

dominant groups in society … the result is the development of powerful informal pressures that marginalize the contributions of members of subordinated groups' (Fraser, 1990, p. 64).

Fraser critiques the idea of a universal public sphere as a neutral space where all who enter are ostensibly equal – something that is frequently claimed, but rarely found in practice. She contrasts the 'counterpublics' in early modern Europe that existed alongside the spaces typically reserved for wealthy white men. Such counterpublics 'contested the exclusionary norms of the bourgeois public, elaborating alternative styles of political behaviour and alternative norms of public speech' (p. 116). Fraser critiques the hegemonic way in which certain public venues for socio-political debate and discussion came to be known as 'the public,' while other kinds of venues and activities that were populated by women, minorities, and working class individuals were excluded from this concept of the public. Subordinated groups had to enter hostile spaces in order to have their discourse be considered part of the public, and Fraser extensively reviews the literature about how dominant groups engage in various practices to silence, intimidate, and chill participation by non-dominant groups. Fraser's feminist critique recasts the bourgeois public sphere as but one of many public spheres – albeit one that was exclusively assumed to represent the population as a whole.

2.5. Do counterpublics need to be parallel discursive arenas?

Traditionally, counterpublics have been understood as 'parallel discursive arenas' (Fraser, 1990, p. 67), separate spaces where members of a subordinated group are able to participate in their own kind of collective sensemaking, opinion formation, and consensus building. Fraser references late twentieth century US feminism as an example of a vibrant counterpublic, with independent publishers, bookstores, conferences, festivals, advocacy

organizations, and other spaces for face-to-face and mediated interaction that ran parallel to more dominant equivalent institutions. As counterpublics are characterized by their lack of a hegemonic claim to represent or speak to the entire population, members must employ alternative tactics to make their concerns and activities visible to 'the public,' while also maintaining a safe space to discuss and understand issues relevant to them on their own terms. With the rise of computer-mediated communication in the 1990s/2000s, many scholars discussed the potential for digitally mediated environments to be counterpublics (Fernback,1997; Papacharissi, 2002; Poster, 2001). In self-organized spaces, marginalized groups could assemble free from the modes of domination that existed in ostensibly 'neutral' spaces. Members of counterpublic spaces could potentially discuss and debate issues according to their own discursive norms, come to common understandings about issues, then engage with more hegemonic media that claims to represent 'the public.' However, as Fernback and Papacharissi note, there are many ways in which such separate online spaces can become disrupted, derailed, and delegitimized by more dominant and hegemonic forces, just as in the counterpublics of early modern Europe.

As blockbots effectively create multiple versions of the same centrally hosted social networking site, they are a different version of Fraser's 'parallel discursive arenas' than the separate digitally mediated environments that are to run alongside more dominant and hegemonic spaces. Rather than creating a separate, alternative discursive space, blockbots are a way in which counterpublic groups exercise agency over their own experiences within a hegemonic discursive environment. In the next sections, I explore blockbots through this lens of counterpublics, discussing how blockbots emerged as a more communal response to the issue of online harassment.

3. Blockbots as a response to harassment campaigns

3.1. The emergence of blockbots

In this section, I discuss the history of blockbots, showing how they were developed to help counterpublic groups participate in a hegemonic networked public on their own terms. Blockbots were initially created around the broader issue of online harassment in Twitter in 2012, although this issue has received substantially more attention in 2014 and 2015 due to a number of more-and-less organized harassment campaigns that unfolded on Twitter – including those aligned with the GamerGate movement (which has opposed feminist video game critics) and multiple shorter cases in which celebrities have been harassed (Chess & Shaw, 2015; Heron et al., 2014; Matias et al., 2015). In 2012, some people who were facing coordinated harassment campaigns for their feminist political stances called for changes to Twitter's policies, enforcement mechanisms, and user interface features, in order to minimize the impact of harassment and take action against identified harassers. In addition to this work petitioning Twitter, Inc., a small group of advocates and activists turned to software-based approaches to help their fellow community members moderate their own experiences on Twitter.

The first blockbot made creative use of features developed largely to support third-party clients (like smartphone applications) – likely an unintended consequence of these Application Programming Interfaces (APIs). These blockbots made it possible for people to collectively curate blocklists of those they identified as harassers, then synchronize these blocklists with subscribers. With a few clicks, a Twitter user could subscribe to a blocklist and automatically no longer see any tweets or notifications from anyone added to that list. Some of these blocklists are curated by single individuals, others by like-minded groups, and a final set are algorithmically generated based on various methods of data collection and analysis. In fact, a core strength of using third-

party bots is that the list of blocked accounts can be created and curated by any given socio-technical system. A blocklist can be curated on an open wiki in which anyone can edit or comment, a vote-based process restricted to a tight-knit group, a ‘team of one’ using Twitter's default blocking interface, an automated agent running predictive models, and so on. Then, using a bot, the blocklist can be automatically synchronized to subscribers using the same computational infrastructure. Today, the BlockTogether.org site (which was launched in August 2014), serves as a centralized clearinghouse for blocklist curators and subscribers. Curators share their blocklists with BlockTogether.org, and the site's bots update subscribers blocklists on their behalf.

Blockbots as a computational entity encapsulate an organization that collectively articulates a list of blockworthy accounts. In fact, blockbots are most compelling in that they help bring together a group of people who oppose a particular understanding of harassment. The history of blockbots cannot just be told in terms of software development release cycles, as they are also about the formation of counter-harassment communities. This capacity for collective action in counter-harassment work is important given the disparities of scale that are associated with online harassment. As many scholars note, a particularly problematic form of harassment takes the form of ‘piling on,' where a large number of people each send a small number of messages, overwhelming the target (Citron, 2014; Matias et al., 2015). The work of harassment can be efficiently distributed and decentralized, with anonymous imageboards serving as one of many key sites for the selection of targets. Some prominent anti-feminist individuals also use Twitter itself to direct their tens of thousands of followers to particular accounts. In such a situation, it only takes a short amount of time and energy to send a single harassing reply. In contrast, the work of responding to harassment is much more difficult to scale, as each of those messages must be dealt with by the recipient. Targets of more-and-less coordinated harassment campaigns are at a distinct

disadvantage even with Twitter's built-in blocking feature. With blockbots, counter-harassment work can be more efficiently distributed and decentralized across a group that shares common understandings about what harassment is.

3.2. The agency to selectively tune into affective publics

Blockbots support a mode of collective action in which people are able to regain some agency over their own experiences on Twitter. Typically, targets of harassment are advised to set their accounts to private, which severely limits their ability to engage in public discourse. In contrast, blockbots let targets of harassment continue to participate in networked publics, but selectively tuned out of the kind of content that would otherwise potentially drive them away from the site. There are numerous accounts by blockbot subscribers who have publicly wrote and tweeted about their use of blockbots like ggautoblocker, noting how their experiences of Twitter dramatically changed after subscribing. The ggautoblocker bot, which uses a social network graph approach to generate a blocklist of individuals in and around the GamerGate movement, has over 3000 subscribers as of mid-2015 (Schwartz et al., 2015). One subscriber of ggautoblocker, who installed the program after an anti-feminist group attacked the hashtag stream of a major conference, exclaimed simply that ‘It works!' Other blockbot subscribers have wrote about how blockbots helped keep Twitter ‘more livable' or more ‘usable' for them when they faced coordinated harassment campaigns. They stated that without the blockbot, they would likely have had no alternative than leaving the site or setting their account to private.

Through this capability of selective filtering, blockbots are a novel way in which counterpublic groups are seeking to refine what Crawford calls ‘disciplines of listening.' She argues that practices of seeking and consuming content in networked publics are an ‘embedded

part of networked engagement' (Crawford, 2009, p. 527). In this view, 'lurking' is as much a complex and multi-faceted mode of participation as submitting content, and blockbots help counterpublic groups participate in Twitter more on their own terms. Similarly, Papacharissi discusses how selective aggregation mechanisms in social media streams support 'affective publics' in which people not only share information and opinion, but also form shared 'structures of feeling':

> Publics assembled out of individuals feeling their way into a particular news stream generated via Twitter engage in practices of rebroadcasting, listening, remixing content, and creatively presenting their views – or fragments of their views – in ways that evolve beyond the conventional deliberative logic of a traditional public sphere. These practices permit people to tune into an issue or a particular problem of the times but also to affectively attune with it, that is, to develop a sense for their own place within this particular structure of feeling. (Papacharissi, 2014, p. 118)

In this way, those who use blockbots do not have to affectively attune with all the harassers who flood their notifications with offensive tweets. If the harassers have already been blocked, they are not visible to the subscriber, and if the harassers have not been blocked, then specific action can against them by adding them to the blocklist.

There is one major caveat about the role of blockbots in affective tuning, which can be seen in hashtag streams, where blockbots do not currently function. This is because it is the responsibility of Twitter clients (the programs or web pages that are used to access the platform) to retrieve tweets from the platform's API, and then filter and display those to the user. Twitter, Inc.'s clients (including TweetDeck) do not filter blocked accounts from search results, including hashtag streams. Tweets from blocked accounts are only filtered from the 'timeline,' the default view of Twitter based on tweets from accounts a user is following. These technical decisions reshape the affordances of the site in powerful ways, which also changes how bespoke tools like blockbots operate. Blockbots do let non-staff developers change the affordances of a privately

owned and operated site like Twitter, but they also are constrained by the design of the systems they seek to change (and the tacit approval of Twitter, Inc. staff, which can block these bots from operating if they so choose). Yet even when Twitter's clients are not programed to filter out blocked accounts when viewing search results, they do still help people moderate their own experiences on Twitter by filtering out unsolicited notifications.

4. **Blockbots are embedded in and emerge from counterpublic communities**

In this next section, I take up blockbots as both communities and computation. I argue that as computational infrastructure for supporting the classification of harassment, blockbots are ongoing accomplishments of collective sensemaking, in which counterpublic groups work to enact ideas about what harassment is and how it ought to be dealt with. I discuss moments of reflection and reconfiguration about blockbots, which show how blockbots often involve a quite different approach than the technologically determinist 'solutionist' mindset prevalent in Silicon Valley.

4.1. Blockbots as communities, not just technologies

As a computational system for classifying harassment, a blockbot's software code enacts a particular understanding of what harassment is and how it ought to be identified. Like all classification systems (and technologies), they are not neutral; they reflect the ways that their designers understand the world (Bowker & Star, 1999). Blockbots are compelling cases for showing how algorithmically supported classification systems are situated within particular contexts, which extend far beyond their source code. In studying the historical development of several different blockbots over time, I have seen how such systems are continually developed and redeveloped as people come to better understand what it even means to ask and answer questions

like ‘Who is and is not a harasser?' and ‘What ought to be done about harassment?' The answers to such questions do not simply require building the right technical infrastructure – this is the ‘solutionist’ belief that there could be one universal system that would finally settle the issue about what is and is not appropriate content. These systems are ongoing accomplishments in sensemaking, just as Bowker and Star note in their studies of classification systems about race, health, and labor. Answers to questions about what kind of behavior ought to be made visible in public spaces emerge out of the lived experiences of many different kinds of people.

In both my archival research and interviews about blockbots, I initially began focusing on this specific kind of automated software agent, but I was continually drawn to the broader projects, communities, and institutions in which those blockbots had meaning and significance. The overwhelming majority of anti-harassment blockbots I encountered were not one-off software development products. They were instead developed out of or into broader projects seeking to formulate responses to online and/or gender-based harassment. For example, the ggautoblocker bot was initially developed by Randi Harper, who recently launched a non-profit organization around online harassment that includes but extends far beyond ggautoblocker – the Online Abuse Prevention Initiative (OAPI). This context is important in understanding how different bot-assisted projects built around curating a collective blocklist operate as counterpublic modes of filtering and gatekeeping. Those involved with blockbots frequently note that these tools are imperfect, incomplete, and continually evolving responses to a complex situation, where they do not have the ability to change all aspects of how Twitter operates. I have continually found that blockbot developers and authorized curators (sometimes called ‘blockers’) regularly revise the code and procedures for curating a bot-supported shared blocklist. They are also frequently seeking better social and technical ways of curating blocklists, identifying harassers, and appealing blocks. These

revisions are compelling cases of socio-technical reconfigurations (Suchman, 2007), as they simultaneously involve changes in more abstract, normative understandings about harassment as well as concrete alterations to a blockbot's source code.

Specific reconfigurations I have observed include: creating an appeals board with a formalized process to review accounts that were allegedly wrongly added to a blocklist; providing support for blockers to document why they added an account to a blocklist; requiring that a second authorized blocker review and approve a new addition to a blocklist, when previously, any authorized blocker could independently add an account to the blocklist; and splitting a single blocklist into a set of multiple lists, based on different understandings of what constituted blockworthyness. These modifications and extensions illustrate how the people who operate such blockbots are actively and continually reflecting on how to best design a social-computational system for moderating their own experiences of Twitter. Far from representing a 'solutionist' mindset that harassment is simply a technical problem to be solved with the right assemblage of algorithms, these cases show how the ostensibly technical work of software development can be a way in which counterpublic groups work out various ideas about what harassment is and what ought to be done about it.

Such changes should not be seen as purely technical solutions, but are instead socio-technical reconfiguration (Suchman, 2007). As Suchman argued with cases about artificial intelligence projects, contexts and existing social artifacts do not determine the configuration of an algorithmically supported system, but they do work as resources that people leverage when seeking to carry out their goals. This shows how an algorithmic system is not only made up software code, but also the shared discourses, practices, internal conflicts, standards, and political and ideological commitments of the people who participated in its design, development, and

deployment. Blockbots are a more communal and collective response to the issue of online harassment, one more in line with the feminist ethics of care that Adam (2000) advocates.

5. Conclusion

5.1. Concerns about fragmentation and automated discrimination

Blockbots can be celebrated as a way for counterpublic groups to moderate their experiences online, but it raises two of the more common fears expressed by those who discuss the Internet and the public sphere. The first fear is the fragmentation of the public sphere into separate, polarized groups, and the second is the use of algorithmic systems as discriminating gatekeepers. Critical scholars must pay close attention to such processes of inclusion and exclusion, because they are the mechanisms in which modes of cultural domination operate (Williams, 1977, p. 125). Scholars and activists who focus on issues of gender, sexual orientation, race and national origin, class, disability, and many other axes of subordination have long critically interrogated the modes of filtering and exclusion that work to erase certain kinds of activity from 'the social,' 'the political,' or 'the public.'

Cultural curation has long been a core mechanism in which domination is reinforced (operating more subtly than more visible acts of formal exclusion and repression), and so it is understandable for blockbots to initially appear suspect by those who are deeply concerned with these issues of social justice. Many scholars both in computer science as well as the social sciences and humanities are concerned about how automated systems for filtering and moderating content can function as invisible gatekeepers, operating similar to the hierarchical editors that controlled content in more traditional media like newspapers, television, and radio. Given the way recommendation and filtering systems work, there is strong potential for such systems to even

influence elections, as a controversial study by staff at Facebook, Inc. suggested (Bond et al., 2012). In response, scholars have sought projects focusing on ‘algorithmic accountability' (Diakopoulos, 2015) and have pushed for more transparency in how such platforms filter and promote content. It makes sense to ask if blockbots raise similar kinds of concerns that algorithmic recommendation and filtering systems do.

With blockbots, such fears of around the biases of algorithmic systems must be understood in their counterpublic context. Blockbots are opt-in, rather than opt-out, making them quite different than the kind of algorithmically supported filtering that exists in the Facebook news feed, for example. Blockbots do not operate according to a top-down model of gatekeeping, as they are built for particular communities to come together and enact a different mode of gatekeeping than is the default in Twitter. Concerns about blockbots as a potentially oppressive mode of filtering must be understood in the larger context of harassment, which works to exclude, repress, and silence public participation, as Fraser (1990), Herring (1999), and Citron (2014) all review. Before blockbots existed, there was already a complex social and technical system that shaped people's experiences of Twitter as a networked public space, which in turn shaped broader understandings of what ‘the public’ believes. That system is the entire ecosystem of Twitter, which includes all the people who silence others from public participation in ways that are not seen as abuse or harassment by Twitter, Inc. staff. While there are concerns about someone being placed on a blocklist by accident or for inappropriate reasons, this is a concern that I found was shared by blockbot developers and curators themselves. The largest and most popular blockbots are run in public, with their blocklists and blocking procedures visible to all, and many have developed accountability and appeals procedures to deal with these concerns.

However, definitions of harassment vary significantly, and blockbots are increasingly used by counterpublic groups to filter activity that is better described as incivility or trolling, rather than harassment. This has led to controversies about what it means to put an account on a blocklist, given that accounts do not have to engage in harassment according to a legal definition of the term in order to be added to a blocklist. Yet in responding to critiques of blockbots as censorship, we must also keep in mind Fraser's critique of 'the public' as a hegemonic way of elevating one of many publics above all others. Part of the privilege of dominant groups is the ability to define the terms of the public by deciding what does and does not belong, as well as defining the current state of the world as the natural default. Blockbots commandeer that privilege to institute a different definition of the public, even if it only has direct effects for those who choose to opt in to the counterpublic group's redefinition. This redefinition of the public calls attention to the different understandings about what a social networking site is and ought to be – and who it ought to be for.

5.2. Technological solutionism

Blockbots are certainly a technology that is deployed to help solve the problem of harassment on Twitter. However, they should not be seen as the kind of top-down technical solution that can be installed to fix the problem 'once and for all' – which, as Barbrook and Cameron as well as Morozov argue, is often accomplished by shifting the burden of responsibility to individual, isolated users. Compared to individual mechanisms, blockbots have emerged as more communal, counterpublic responses to harassment. Like Kelty's discussion of open source software infrastructure (Kelty, 2008), blockbots help form a 'recursive public' in which the ideals of the group are intentionally embedded in the design of the software that supports their activities. Blockbots are a different kind of computationally supported mode of platform governance, which

can be seen in the formation of broader anti-harassment communities around blockbots and the thoughtful reconfigurations in blockbot infrastructure. In fact, despite the utility that blockbots have in helping people shape their own experiences online, they are perhaps even more impactful in that they have provided a catalyst for the development of anti-harassment communities. These groups bring visibility to the issue and develop their own ideas about what kind of a network public Twitter ought to be.

Blockbots provide a concrete alternative to the default affordances of Twitter, showing a different version of a public: one where people have more agency to selectively tune out of harassment, without dropping out of public participation altogether. Their existence has sparked broader conversations about what public discourse online ought to look like, as well as what kind of relationship platform owner/operators ought to have with ‘their' users. The kind of bottom-up, decentralized, community-driven approach exemplified by blockbots stands in opposition to the more traditional top-down, centralized, systems administration approach exemplified by Lessig's ‘code is law' argument and much of the ‘politics of algorithms' literature. Blockbots are as much of a social solution as they are a technological one, and their strength is in their capacity to serve as multiply overlapping sites for collective sensemaking and reflective reconfiguration among counterpublic communities – rather than seeking to deploy a single technological solution that seeks to fix the problem for all users, once and for all. As the issue of online harassment is multifaceted and complex, there is unlikely to be a single solution. However, we should look to projects like blockbots when thinking about what it means for privately owned public spaces ought to be moderated.

Acknowledgements

I would like to thank Jenna Burrell and Paul Duguid, who have helped supervise this work as part of my Ph.D, providing invaluable support, feedback, and mentorship. I am grateful to Nathan Matias for his generous work in helping me investigate and conceptualize this specific topic area. I would also like to thank many other people for helping me in research and revise this work, including: Aaron Halfaker, Amanda Menking, Amy Johnson, Ben Light, Gina Neff, Megan Finn, Nick Doty, Norah Abokhodair, Philip Howard, Randi Harper, Richmond Wong, Samuel Woolley, Whitney Phillips, the members of the UC-Berkeley School of Information seminar on Technology and Delegation, the Center for Media, Data, and Society at Central European University, audience members at my presentation at the Association of Internet Researchers' annual meeting, and the anonymous reviewers.

Funding

This work was supported by a doctoral completion fellowship at UC-Berkeley and a pre-doctoral fellowship at the Center for Media, Data, and Society at the Central European University in Budapest, Hungary.

References

Adam, A. (2000). Gender and computer ethics in the internet age. *Computer Professionals for Social Responsibility*, *18*(1).

Adam, A. (2005). *Gender, ethics, and information technology*. New York, NY: Palgrave Macmillan.

Bond, R. M., Fariss, C. J., Jones, J. J., Kramer, A. D. I., Marlow, C., Settle, J. E., & Fowler, J. H. (2012). A 61-million-person experiment in social influence and political mobilization. *Nature*. doi:10.1038/nature11421

Barbrook, R., & Cameron, A. (1996). The Californian ideology. Science as Culture. doi:10.1080/09505439609526455

Barlow, J.P. (1996) A Declaration of the Independence of Cyberspace. In J. Casimir (Ed.), *Postcards from the Net: An Intrepid Guide to the Wired World Web*, pp. 365-7. Sydney: Allen and Unwin. Retrieved from https://www.eff.org/cyberspace-independence

Bowker, G. C., Baker, K., Millerand, F., & Ribes, D. (2010). Toward information infrastructure studies: Ways of knowing in a networked environment. In *International Handbook of Internet Research* (pp.97–117). Dordrecht: Springer. doi:10.1007/978-1-4020-9789-8_5

Bowker, G. C., & Star, S. L. (1999). *Sorting things out: Classification and its consequences*. Cambridge, MA: MIT Press.

boyd, d. (2010). Social network sites as networked publics: Affordances, dynamics, and implications. In Z. Papacharissi (Ed.), *A Networked Self: Identity, community, and culture on social network sites* (pp. 39–58). New York: Routledge. Retrieved from http://www.danah.org/papers/2010/SNSasNetworkedPublics.pdf

Chess, S., & Shaw, A. (2015). A Conspiracy of fishes, or, how we learned to stop worrying about #Gamergate and Embrace Hegemonic Masculinity. *Journal of Broadcasting & Electronic Media*,*59*(March), 37–41. doi:10.1080/08838151.2014.999917

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

Crawford, K. (2009). Following you: Disciplines of listening in social media. *Continuum: Journal of Media & Cultural Studies*. doi:10.1080/10304310903003270

Crawford, K., & Gillespie, T. (2014). What is a flag for? Social media reporting tools and the vocabulary of complaint. *New Media & Society*. doi:10.1177/1461444814543163

Diakopoulos, N. (2015). Algorithmic accountability: Journalistic investigation of computational power structures. *Digital Journalism*, *3*(3), 398–415. doi:10.1080/21670811.2014.976411

Dibbell, J. (1993). A rape in cyberspace. *The Village Voice*, 42, 36. Retrieved from http://www.villagevoice.com/news/a-rape-in-cyberspace-6401665

Duggan, M. (2014). *Online harassment: Summary of findings*. Retrieved from http://www.pewinternet.org/2014/10/22/online-harassment/

Fernback, J. (1997). The individual within the collective: Virtual ideology and the realization of collective principles. In S. Jones (Ed.), *Virtual culture: Identity and communication in cybersociety* (pp.36–55). London: Sage.

Fleishman, G. (2014, August 11). How collaborative social blocking could bring sanity to social networks. *BoingBoing*. Retrieved from http://boingboing.net/2014/08/11/whack-a-troll-with-collaborati.html

Fraser, N. (1990). Rethinking the public sphere: A contribution to the critique of actually existing democracy. *Social Text*, *26*, 56–80. doi:10.2307/466240

Geiger, R. S. (2014). Bots, bespoke, code and the materiality of software platforms. *Information, Communication & Society*, *17*(3), 342–356. doi:10.1080/1369118X.2013.873069

Geiger, R. S., & Ribes, D. (2011). *Trace ethnography: Following coordination through documentary practices*. In Proc HICSS 2011. IEEE. Retrieved from http://www.stuartgeiger.com/trace-ethnography-hicss-geiger-ribes.pdf

Gillespie, T. (2010). The politics of "platforms". *New Media & Society*, *12*(3), 347–364. doi:10.1177/1461444809342738

Gillespie, T. (2014). The relevance of algorithms. In T. Gillespie, P. Boczkowski, & K. Foot (Eds.), *Media technologies: Essays on communication, materiality, and society* (pp. 167–194). Cambridge, MA:The MIT Press. Retrieved from http://6.asset.soup.io/asset/3911/8870_2ed3.pdf

Habermas, J. (1989). *The structural transformation of the public sphere*. Cambridge, MA: The MIT Press.

Heron, M. J., Belford, P., & Goker, A. (2014). Sexism in the circuitry. *ACM SIGCAS Computers and Society*, *44*(4), 18–29. doi:10.1145/2695577.2695582

Herring, S. C. (1999). The Rhetorical Dynamics of Gender Harassment On-Line. *The Information Society*. doi:10.1080/019722499128466

Hess, A. (2014, October). Twitter won't stop harassment on its platform, so its users are stepping in. *Slate*. Retrieved from http://www.slate.com/blogs/future_tense/2014/08/06/twitter_harassment_user_created_apps_block_together_flaminga_and_the_block.html

Irani, L. (2013). The cultural work of microwork. *New Media & Society*, *17*(5), 720–739. doi:10.1177/1461444813511926

Kelty, C. (2008). *Two bits: The cultural significance of free software*. Durham, NC: Duke University Press.

Lessig, L. (1999). *Code and other laws of cyberspace*. New York, NY: Basic Books.

Marwick, A., & Miller, R. (2014). *Online harassment, defamation, and hateful speech: A primer of the legal landscape*. New York: Fordham Center on Law and Information Policy.

Massanari, A. (2015). #Gamergate and the fappening: How Reddit's algorithm, governance, and culture support toxic technocultures. *New Media & Society*,1461444815608807. doi:10.1177/1461444815608807

Matias, J. N., Johnson, A., Boesel, W. E., Keegan, B., Friedman, J., & DeTar, C. (2015). *Reporting, reviewing, and responding to harassment on Twitter*. Retrieved from http://papers.ssrn.com/abstract=2602018

Morozov, E. (2013). *To save everything click here: The folly of technological solutionism*. New York, NY: PublicAffairs.

Papacharissi, Z. (2002). The virtual sphere: The internet as a public sphere. ***New Media & Society***, *4*(1), 9–27. doi:10.1177/14614440222226244

Papacharissi, Z. (2014). *Affective publics: Sentiment, technology, and politics*. Oxford: Oxford University Press.

Pariser, E. (2012). *The filter bubble. The filter bubble*. New York, NY: Penguin.

Phillips, W. (2015). *This is why we can't have nice things: Mapping the relationship between online trolling and mainstream culture*. Cambridge, MA: The MIT Press.

Poster, M. (2001). *Cyberdemocracy: Internet and the public sphere* (pp. 259–271). Malden, MA: Blackwell Publishing.

Schwartz, R., Lynch, D., & Harper, R. (2015). OAPI. *FLOSS Weekly*, 331. Retrieved from https://twit.tv/shows/floss-weekly/episodes/331

Star, S. L. (1999). The ethnography of infrastructure. *American Behavioral Scientist*, *43*(3), 377–391. doi:10.1177/00027649921955326

Star, S. L., & Strauss, A. (1999). Layers of silence, Arenas of voice: The ecology of visible and invisible work. ***Computer Supported Cooperative Work (CSCW)***, *8*,9–30. doi:10.1023/A:1008651105359

Suchman, L. (2007). *Human-machine reconfigurations: Plans and situated actions*. Cambridge: Cambridge University Press.

Tokunaga, R. S. (2010). Following you home from school: A critical review and synthesis of research on cyberbullying victimization. *Computers in Human Behavior*, *26*(3),277–287. doi:10.1016/j.chb.2009.11.014

Tufekci, Z. (2014). Engineering the public: Big data, surveillance and computational politics. *First Monday*, *19*(7). doi:10.5210/fm.v19i7.4901

Turner, F. (2006). *From counterculture to cyberculture: Stewart brand, the whole earth network, and the rise of digital utopianism*. Chicago, IL: University of Chicago Press.

Williams, R. (1977). *Marxism and literature*. Oxford: Oxford University Press.