International Journal of Communication 10(2016), Feature 5022–5031 1932–8036/2016FEA0002

When Bots Tweet: Toward a Normative Framework for Bots on Social Networking Sites

NATHALIE MARECHAL[1]
University of Southern California, USA

Political actors are using algorithms and automation to sway public opinion, notably through the use of "bot" accounts on social networking sites. This article considers the responsibility of social networking sites and other platforms to respect human rights, such as freedom of expression and privacy. It then proposes a set of standards for chat bots operating on these platforms, building on the existing policies of leading social networking platforms and on the indicators laid out by Ranking Digital Rights. A good normative framework for the use of bots on social networking sites should have three components: bots should clearly be labeled as such, they should not contact other users without consent, and information collected by them should only be used for disclosed purposes.

Keywords: bot, chat bot, social networking services, Twitter, ethics, technology, policy, digital rights

Political actors are using algorithms and automation in efforts to sway public opinion, notably through the use of "bot" accounts on Twitter, Facebook, Reddit, and other social media platforms. Bots are understood to be "amalgamations of code that mimic users and produce content" (Woolley & Howard, 2016, para. 1) or "automated software agents" (Geiger, 2016, p. 1). This is a rather broad category, as any repetitive online task can, in theory at least, be automated. Existing types of bots include malicious botnets—networks of bots controlled at a distance—used for distributed denial of service (DDoS) attacks, research bots that crawl the Web to retrieve data and other information, editing bots such as those used by Wikipedia, and chat bots that can reply to basic queries.

In this article, I consider bots that operate within social networking sites (SNS) such as Facebook, Twitter, and Reddit. Many bots, including custom Python script, scrape otherwise publicly viewable data for research purposes, while others proactively engage in the sites' public discourse, much in the same way that a human user might. As Stuart Geiger (2016) has noted,

[1] The author gratefully acknowledges the feedback from the daylong workshop, Algorithms, Automation and Politics, organized by the European Research Council funded Computational Propaganda project of the Oxford Internet Institute and held as a preconference to the International Communication Association Meeting in Fukuoka, Japan, in June 2016. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author and do not necessarily reflect the views of the European Research Council.

> blockbots support counterpublic communities, helping people moderate their own experiences of a site . . . [by] support[ing] the curation of a shared blocklist of accounts, where subscribers to a blockbot will not receive any notifications or messages from accounts on the blacklist. (p. 1)

Users who feel harassed on social media use blockbots to help regain control of their social media experience. Blockbots help maintain the right of users to exist online free from harassment, without actually preventing the harasser from speaking. Conversely, other bots are designed to actively participate in public discourse by harassing users, retweeting content produced by predetermined users, hijacking hashtags or other curated conversations, or impersonating public figures or institutions (Forelle, Howard, Monroy-Hernandez, & Savage, 2015). How should a SNS deal with the existence of these types of bots on the platform, while respecting users' rights to privacy and freedom of expression?

One of the principal difficulties in elaborating such privacy and free expression norms is that they ought to be universally applicable: It would be unrealistic to expect companies to maintain materially different policies with respect to different governments in varying political contexts. Indeed, this would place private companies in the untenable position of adjudicating the relative merits of political goals and tactics. For example, a central argument against the U.S. and UK governments' demands for so-called backdoors into otherwise encrypted communication channels is that such access would inevitably lead to similar demands for entry from other governments. Some of these governments would then use that access in ways that are incompatible with human rights, including the persecution of human rights defenders and other dissidents.

The norms should also be "substance blind," in the sense that they should not discriminate between different political parties or ideologies—with clearly delineated exceptions in keeping with internationally recognized human rights standards, such as those concerning hate speech and pornography. Nor should these norms prohibit all bot use, as not all bots pose threats to human rights: many are benignly amusing ("Ten Bots That Make," 2014), whereas others are in fact research tools that help human researchers expose algorithmic threats to digital rights (Olson, 2013).

Another important consideration will be the applicability of the proposed norms to human beings. Would it be ethical to apply the same norms restricting bot "speech" to human beings, or is there someone distinctively "nonhuman" about bots that justifies treating them differently? Tim Wu (2013) makes a convincing case that robots do not deserve the same free speech protections as people do, but what if you can't tell whether a particular account belongs to a human or a machine? What if a human user automates some of her account's actions, but not all of them? These questions do not lend themselves to easy answers.

Human Rights and Bot Activity

My argument assumes certain normative stances: that human rights are universal; that companies have a responsibility to respect human rights, as per the UN Guiding Principles on Business and

Human Rights ("the Ruggie Principles"). Next, that technological platforms are not, in fact, neutral spaces where users engage in the free marketplace of ideas, but that their design and policy choices affect the options available to users, inherently privileging certain classes of users over others. Finally, that the use of political bots can often threaten human rights in both the digital and physical realms. In this section, I expand on and provide support for these claims.

Disputes over the universality of human rights have a long history. Today, many governments justify abusive laws and practices—such as restricting online expression and access to information—by citing cultural differences. Respect for human rights can manifest itself in different ways, women's right to nudity versus women's right to modesty, for instance. I reject the claim that one can pick and choose among the rights contained in the Universal Declaration of Human Rights. Indeed, human rights are indivisible and interdependent, and weakening the protections of any one of these rights weakens the others as well (Minkler & Sweeney, 2011).

While the 1948 Declaration was written with nation-states in mind, the 2011 Guiding Principles on Business and Human Rights extend the responsibility for respecting human rights to private firms, including information communication technology companies (Ruggie, 2013). Indeed, these private information intermediaries are the sovereigns of cyberspace (DeNardis, 2015; MacKinnon, 2012). Social networking sites, in particular, are the new public sphere where power and counterpower come face to face (Castells, 2007; Habermas, 1989). The intricate politics of liking, reacting, commenting, friending, and unfriending may have seemed sophomoric at first, better suited to teenage microdramas than to the serious business of politics. But as the U.S. presidential election campaign season heats up, closely intertwined with the high-stakes protest dynamics of the #BlackLivesMatter movement and its counterclaims, #AllLivesMatter and #BlueLivesMatter, the weight of social networking decisions have become apparent. It is safe to say that readers of academic journals who are also active on social media will have pondered the appropriate reaction to an acquaintance's objectionable post. Race, class, gender, and other dimensions of privilege determine who can afford to confront opposite views in the public sphere, especially hateful ones. The resulting conversation is often a far cry from equal participation, though social media users who benefit from privilege tend to assume that all voices in a social media conversation have an equal chance to be heard (Geiger, 2016).

In this regard, the digital public sphere parallels its analog predecessor. As Nancy Fraser (1990) points out, the salons and cafés of post-Enlightenment Europe were far from neutral spaces. The voices of the working classes, of communities of color, and of women were consistently marginalized, even as the dominant discourse maintained the fiction that the space was a neutral one.

Similarly, hegemony in social network platforms is maintained in various ways: harassment, trolling, doxxing, threats, and "mansplaining" all serve to police who has the right to be part of the public. In July 2016, the comedian and *Ghostbusters* actress Leslie Jones was hounded off of Twitter by mobs of racist, misogynist trolls who objected to her appearance in the reboot of the 1980s cult classic. Unable to prevent her inclusion in the film, they silenced her voice on this platform (Rogers, 2016).

Today's counterpublics can talk amongst themselves through secret Facebook pages, moderated sub-reddits, or within communities of protected Twitter accounts, much as their forebearers could mingle among their own kind in parallel discursive arenas (Fraser, 1990). But this does nothing to correct the impression that public opinion largely aligns with the opinions of the most privileged members of society. Block bots allow members of counterpublics to participate in the "mainstream" discourse while shielding themselves from at least some forms of harassment. Block bots appear to be a tool for supporting freedom of expression.

Other bot types, however, seem to be designed for the express purpose of disrupting free expression—for example, by hijacking Twitter hashtags and flooding them with contrary or irrelevant messages. Journalist Erin Gallagher has been studying Mexican botnets, which she calls weaponized censors that silence dissent and cover up crimes. Gallagher identifies five primary uses of political bots: hashtag spamming; making artificial trends; smear campaigns; death-threat campaigns; and political propaganda (Gallagher, 2015). Each tactic infringes on the free expression rights of others in a different way. Hashtag spamming—the practice of affixing a specific hashtag to irrelevant content—renders the hashtag unusable. Artificial trends can bury real trends, thus keeping them off the public and media's radar. Smear campaigns and death threats can both intimidate vocal political opponents and dissuade would-be speakers. The line between propaganda and legitimate political speech is a fine one, of course, and may in some cases sit entirely in the eye of the beholder. Nonetheless, bots can be used to amplify the propagandists' desired message.

The Status Quo: How SNSs Deal With Bots

The following section describes leading platforms' existing policies and analyzes their shortcomings. The four platforms selected—Twitter, Facebook, Reddit, and Telegram—were chosen for their global popularity and openness to bot use. Of the four, Twitter has the most explicit guidelines concerning bot use, Reddit lacks any bot-specific rules, and Telegram has only one: that bots' usernames should end with the word "bot," thus identifying their automated nature to human users. Some of Facebook's guidelines, on the other hand, seem intended to prevent companies and brands from using Messenger to reach users without purchasing Facebook ads or from processing payments without giving Facebook a portion of the revenue. While it is well within Facebook's right to institute such rules, the rules lack a clear connection to digital rights. Other Facebook rules consist of prohibitions against spamming users without prior consent.

Twitter

The microblogging platform's Help Center includes a page dedicated to "Automation Rules and Best Practices" (Twitter, 2016). The following types of bots are prohibited:

- Posting automated links that redirect through landing or ad pages before the final content
- Distributing user content without express consent
- Automatically tweeting to trending topics (also known as hashtag spamming)
- Automated replies or mentions without Twitter's approval

- Automated retweeting, unless Twitter deems this activity to be a "community benefit"
- Automated following and unfollowing
- Automated favoriting

The first rule seems clearly designed to protect Twitter's advertising revenue. Coming from a company that is struggling to generate revenue, this is understandable if irrelevant to human rights. The remaining rules are designed to discourage revenge porn, hashtag spamming, harassment, and creating artificial trends, respectively. However, the enforcement of these rules is haphazard at best. Twitter, like other SNSs, relies chiefly on user reports of rule breaking (Crawford & Gillespie, 2016). Even then, I argue that much activity that would seem like harassment is deemed to meet community standards. Rules without fair enforcement are just anarchy. Jillian C. York (2016) argues that this practice feeds a culture of snitching, akin to the Stasi's network of citizen informants, and that the reliance on flagging serves to reify existing power imbalances.

Facebook

Although bots have not historically been a prevalent part of Facebook communications, this may change. At the April 2016 F8 event, Facebook CEO Mark Zuckerberg unveiled Messenger's new chat bot functionality, which the company's vice president for messaging, David Marcus, said represents the next step in the Internet's evolution, supplanting websites and apps (Hempel, 2016). The Facebook for Developers site (Facebook, 2016, para. 1) listed the following rules for chat bot developers at the time:

- Place any user authentication method in a clear and conspicuous location to ensure people consent to initiating message threads.
- Don't contact people in Messenger unless they've agreed to be contacted by you, or the party to whom you are operating as a service provider.
- Messenger Opt-out: respect all requests (either on Messenger or off) by people to block, discontinue, or otherwise opt-out of your using Messenger to communicate with them.
- After we approve your initial intended use cases for Messenger, don't add additional use cases without submitting them to us for our review and approval. Responding to a customer inquiry is permitted without our prior approval, but ensure your response adheres to our policies (i.e., it must not include ads).
- Don't use Messenger for advertising, marketing, or for sending promotional content of any kind, even if a person opts-in to receiving this content, without our prior written permission.
- Don't request or share individual payment card, financial account numbers or other cardholder data within Messenger.
- Don't include links to sites off Messenger where payment information is collected, without our prior permission.
- Don't use any data obtained from us about the people you reach in Messenger, other than the content of message threads, for any purpose other than as reasonably necessary to support the message types you elect to use.
- Games: Don't use Messenger Platform if your game accepts payment (i.e., games must be free to play and offer no paid content).

- We may limit or remove your access to Messenger if you receive large amounts of negative feedback or violate our policies, as determined by us in our sole discretion.

As with Twitter, many of these policies seem designed to protect Facebook's revenue stream. Again, that is a reasonable enough corporate strategy, but irrelevant to human rights.

Reddit

Although Reddit bots abound, the company does not disclose any rules or guidelines concerning their use (LaCapria, 2014). Presumably, bots are subject to the same rules as human users, including prohibitions on harassment, threats, and nonconsensual disclosure of personal information. Additionally, the moderators of individual sub-Reddits can institute any rules they wish for the portion of the site they manage.

Telegram

The only rule concerning bots on Telegram is that a bot username must end with the word "bot," thus signaling to human users that they are dealing with a chat bot.

Perhaps unsurprisingly, each company's approach to chat bots reflects the broader corporate ethos. Twitter and Facebook are overwhelmingly concerned with protecting their revenue streams while still acknowledging that a poor user experience represents a threat to their business model. Reddit leaves the rulemaking, and responsibility for enforcing any rules, to volunteer moderators of sub-Reddit portions of the platform. Telegram—perennially besieged by accusations that it harbors extremist sympathizers and is openly hostile to the Russian government, among others—merely requests that bots' user names declare their algorithmic nature.

If there is a unifying trend, it is that chat bot policy does not seem to rank highly among these companies' priorities. Nor is this surprising: In recent years, Twitter has struggled to find a sustainable business model, while Reddit has weathered its own share of controversy and executive turnover. Chat bots were entirely absent from the Facebook ecosystem until very recently, and Telegram lacks most of the trappings of a global technology company, including a business model, offices, or a trust and safety team. It is clear, then, that no consensus exists concerning the use of bots on SNSs. What might such a set of best practices look like?

Best Practices for Bot Policies

One of the most evolved normative frameworks for evaluating the behavior and impact of firms and their technologies has been developed by the Ranking Digital Rights (RDR) project. After three years of research and methodology development, in November 2015 the RDR project launched its inaugural Corporate Accountability Index, which assesses the public commitments and disclosures of 16 global technology companies on freedom of expression and privacy. The 31 evaluation indicators are grounded in the Guiding Principles on Business and Human Rights (Ruggie, 2013) and are structured in such a way

that they are broadly applicable to the full gamut of companies and product lines. Indeed, in 2017 the Index will expand to include smartphone software and hardware. Researchers at the Center for Internet and Society (Bangalore) are using these indicators to evaluate providers of free Wi-Fi hotspots. Civil society organizations across Latin America, with support from the Electronic Frontier Foundation, are jointly developing a research and advocacy campaign to pressure Internet service providers in the region to better respect privacy and free expression norms, using a similarly aligned approach.

Although the inaugural Index was published in 2015, there are already indications that the strategy of coupling public benchmarks with company-oriented insider advocacy is effective. For example, Facebook has implemented two recommendations from the 2015 Index in its WhatsApp and Instagram applications: end to end encryption and two-step authentication. Likewise, Facebook's Messenger now offers optional encryption for messages between two mobile applications. Moreover, anecdotal evidence suggests that companies beyond the 16 that were ranked in 2015 are using the Index to improve upon their own performance. Representatives from several companies that were not part of the ranking have told RDR staff that they used the indicators in internal assessments of policies and practices related to digital rights. It is clear that norms based on the RDR indicators have the potential to maximize the benefits from bots on SNSs while limiting their negative effects. What, then, should these norms be?

Ranking Digital Rights Governance indicators apply to a company as a whole, and are not specific to an SNS's bot policy. This set of indicators seeks to measure the company's overall understanding of the role it plays in mediating its users participation in the public sphere, and its commitment to enhancing, rather than restricting, user's freedom of expression and privacy. They should therefore be part of any SNS's self-evaluation process.

The 2015 RDR Index comprised 31 indicators, of which six measured company commitment to human rights, 11 evaluated disclosure related to freedom of expression, and 14 pertained to privacy-related disclosures. The project team has been revising the indicators in anticipation of data collection for the 2017 Index and expanding the methodology to include devices, software, and the companies that make them. A draft revised methodology was published for comment in July 2016, to be finalized by late summer. The proposed norms below take the ongoing revision work into account.

The freedom of expression indicators fall under five overlapping categories: indicators about the terms of service, about censoring content accessed or generated by users, about notifying users of actions involving their accounts or content, about transparency reporting, and user anonymity or pseudonymity. Other indicators pertaining to net neutrality and Internet shutdowns do not apply to SNSs. The privacy indicators fall under six overlapping categories: indicators about privacy policies, about user data, about governmental or other requests for user data, about notifying users of actions involving their accounts or content, about transparency reporting, and data security.

These indicators all apply to SNSs; the 2015 Index evaluated Facebook, Instagram, Ozone, Tumblr, Twitter, and Vkontakte. The indicators make no specific mention of bots. In addition to making a company-wide commitment to human rights, and implementing the specific best practices related to free

expression and privacy outlined above, what norms should SNSs follow to better respect digital rights in this area?

Existing company practices provide a starting point for answering this question. Telegram's requirement that bots end their username with the word "bot" is one way of clearly disclosing to users that an account is automated. Existing practices related to digital rights boil down to disclosure, consent, and respect for the principle of secondary use, which dictates that data collected for one purpose should not be used for another, unrelated purpose:

1. Disclosure: Bot accounts should be clearly identified as such.
2. Consent: Bots should not initiate contact with human users without their consent, including interactions such as liking, favoriting, or retweeting.
3. Secondary use: Bot "owners" should not use information collected about users for purposes other than those disclosed to users at the time of collection.

Currently, none of the SNSs examined in this article meet this standard. Twitter comes closest, but as discussed above, unenforced rules are all but meaningless. The problem of terms of service enforcement is one with which companies, governments, civil society actors, academics, and users are currently grappling, and bots are but one relatively small part of the discussion. Nevertheless, these three principles of disclosure, consent, and secondary use are a solid starting point.

These principles also offer applicability beyond the narrow issue of chat bot policy. As emphasized by Ranking Digital Rights, companies should disclose their privacy policies, terms of service, and other documentation of the platform's rules to all potential users, in a way that is easy for users who are not telecommunications lawyers to understand. For platforms like SNSs, whose affordances allow for chat bots, these documents should outline what types of bots are permitted (or not) and the circumstances under which bots might be blocked or censored, including the process leading to such a determination. Companies should make clear to users how they can submit a complaint about a particular bot, how to appeal any complaints made against them, and the process for resolving such disputes. After the fact, companies should disclose, on a regular basis, the number and types of complaints received about bots (including the number of complaints received from government entities), and the resolution of these complaints. Likewise, companies should disclose the process through which governments or other third parties can request data about a bot or its human creator, its process for responding to such requests, and aggregate figures describing the outcome of such requests.

When civil society actors demand more transparency from technology companies, corporate officials often respond that accountability exercises require extensive tracking tools, as well as human and technical resources. Some individuals, particularly on the engineering side, argue that as private corporations, their primary responsibility is to make money for their shareholders, not save the world. The Ruggie Principles are more clear-cut with respect to sectors like the diamond industry and manufacturing, where labor issues come into play. The Ruggie Principles nonetheless demand that technology companies respect human rights, notably free expression and privacy. While bots have many more applications that

raise human rights concerns, rights-respecting policies about the use of chat bots on social media are a place to start.

References

Castells, M. (2007). Communication, power and counter-power. *International Journal of Communication*, *1*, 238–266.

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

DeNardis, L. (2015). The global war for Internet governance. New Haven, CT: Yale University Press.

Facebook. (2016, July 1). Facebook for developers—Platform guidelines—Messenger. Retrieved from https://developers.facebook.com/docs/messenger-platform/guidelines

Forelle, M. C., Howard, P. N., Monroy-Hernandez, A., & Savage, S. (2015). Political bots and the manipulation of public opinion in Venezuela. *SSRN Electronic Journal*. doi:10.2139/ssrn.2635800

Fraser, N. (1990). Rethinking the public sphere: A contribution to the critique of actually existing democracy. In C. J. Calhoun (Ed.), *Habermas and the public sphere* (pp. 109–142). Cambridge, MA: MIT Press.

Gallagher, E. (2015, August). *Mexican botnet dirty wars.* Presented at the Chaos Communication Camp 2015, Zehdenick, Germany. Retrieved from https://events.ccc.de/camp/2015/Fahrplan/events/6795.html

Geiger, R. S. (2016). Bot-based collective blocklists in Twitter: The counterpublic moderation of harassment in a networked public space. *Information, Communication & Society, 19*(6), 787–803. doi:10.1080/1369118X.2016.1153700

Habermas, J. (1989). *The structural transformation of the public sphere: an inquiry into a category of bourgeois society*. Cambridge, MA: MIT Press.

Hempel, J. (2016, April 12). Facebook believes messenger will anchor a post-app Internet. *Wired.com.* Retrieved from http://www.wired.com/2016/04/facebook-believes-messenger-will-anchor-post-app-internet/

LaCapria, K. (2014, March 14). Reddit bots ranked! Seven of our favorite fakeish Reddit accounts. *Social News Daily*. Retrieved from http://socialnewsdaily.com/26682/reddit-bots-ranked-seven-of-our-favorite-fakeish-reddit-accounts/

MacKinnon, R. (2012). *Consent of the networked: The world-wide struggle for Internet freedom*. New York, NY: Basic Books.

Minkler, L., & Sweeney, S. (2011). On the indivisibility and interdependence of basic rights in developing countries. *Human Rights Quarterly, 33*(2), 351–396. doi:10.1353/hrq.2011.0017

Olson, P. (2013, December 2). This landmark study could reveal how the Web discriminates against you. *Forbes.com*. Retrieved from http://www.forbes.com/sites/parmyolson/2013/12/02/this-landmark-study-could-reveal-how-the-web-discriminates-against-you/#37c484b04fa5

Ranking Digital Rights. (2015). *Corporate accountability index 2015 research indicators.* Washington, DC: Author. Retrieved from https://rankingdigitalrights.org/wp-content/uploads/2015/06/RDR-2015-CAI-Indicators.pdf

Rogers, K. (2016, July 19). Leslie Jones, star of "Ghostbusters," becomes a target of online trolls. *NYTimes.com.* Retrieved from http://www.nytimes.com/2016/07/20/movies/leslie-jones-star-of-ghostbusters-becomes-a-target-of-online-trolls.html?_r=0

Ruggie, J. G. (2013). *Just business: Multinational corporations and human rights*. New York, NY: W. W. Norton.

Ten bots that make Twitter a better place. (2014, August 22). *Gadgets 360*. Retrieved from http://gadgets.ndtv.com/social-networking/features/ten-bots-that-make-twitter-a-better-place-579944

Twitter. (2016, April 7). Automation rules and best practices. Retrieved from https://support.twitter.com/articles/76915#

Woolley, S., & Howard, P. N. (2016). Social media, revolution, and the rise of the political bot. In P. Robinson, P. Seib, & R. Frohlich (Eds.), *Routledge handbook of media, conflict, and security*. New York, NY: Routledge.

Wu, T. (2013). Machine speech. *University of Pennsylvania Law Review*, *161*, 1495–1533.

York, J. C. (2016, July 14). Facebook and Twitter are getting rich by building a culture of snitching. Quartz. Retrieved from http://qz.com/731347/facebook-and-twitter-are-getting-rich-by-building-a-culture-of-snitching/