To: Vijaya Gadde, Head of Legal, Policy and Trust at Twitter, Inc.
From: Nicholas Araya
Re: Recommendations for Improved Content Moderation
Date: 10 May 2020
Section: 212 (Samir Passi)

This policy memo discusses Twitter, Inc.'s[1] efforts to improve its content moderation; it outlines two content moderation problems that the company currently faces, addresses the company's current management of these problems, and recommends additional policies to remedy such issues. The content moderation issues which this memo focuses on are as follows:

- Content Moderation Measures' Lack of Transparency and Accessibility; and
- Inconsistent Response Time to Abuse and/or Harassment Reports.

This memo suggests the following policies:

- Increased Transparency, Disclosure, and Accessibility around Content Moderation; and
- Prioritization of Abuse and/or Harassment Reports.

The rest of this memo is thus organized: Section I provides key background information regarding Twitter's attempts to improve its content moderation. Section II discusses the aforementioned content moderation issues, evaluating Twitter's attempts to confront the aforementioned issues and revealing any shortcomings of these attempts. Section III further elaborates on the aforementioned policy suggestions, explaining their utility and addressing any counter-arguments. Section IV concludes the memo and discusses the implications of Twitter's current approach to content moderation and the implementation of my proposals.

Section I: Background

The purpose of this section is to contextualize Twitter's content moderation efforts. In this section, I discuss the history of Twitter's content moderation efforts, its response to criticism on its perspectives on content moderation, and the evolution of its policies regarding free speech, revealing shortcomings and points of controversy.

Twitter's regulation of users' content has evolved a great deal since the service was first created in 2006. Like other online platforms, and protected from liability for third party users' posts by Section 230 of the Communications Decency Act of 1996, Twitter sought to be "the free speech wing of the free speech party" (Halliday). In its efforts to create an inclusive online space, Twitter took an all-encompassing stance on the content that it would allow its users to create since the company "long believed that good speech can counteract bad speech" (Cox & Koebler), and stated that it would "'not censor user content except in limited circumstances'" (Leetaru 2020). However, individuals and other media platforms have criticized Twitter's inconsistent, loose, and obscure methods of content moderation, but also the potentially extreme implications Twitter's ability to "wield absolute authority over what every government on earth

[1]This memo will use the terms "Twitter, Inc." and "Twitter" interchangeably.

may say to their citizens in the online world that has become the defacto modern town square" (Leetaru, "The Remarkable Reversal").

In recent years, Twitter has attempted to clarify and methodize its content moderation procedures in response to the widespread public criticism it has received. For example, In December of 2017 Twitter announced that it would be enforcing new and more clearly outlined rules regarding hate speech; these new regulations included "broadening its hateful conduct policy and rules against abusive behavior to include those accounts that abuse or threaten others through their profile information, like their username, display name, or profile bio." in order to better combat and censor hate speech (Perez 2017). Additionally, the company In October of 2018, Twitter chose to specify whether deleted content was deleted because it violated content regulations in order to make "it more transparent to users whether a deleted tweet was deleted by the user or because Twitter took an action" (Perez 2018).

Yet, some of Twitter's most recent efforts aimed at determining who should be de-platformed, such as its research on how white supremacists use twitter to disseminate information have also been construed a poor choice which evidences matter "'the callousness with which they've approached this issue on their platform'" (Cox & Koebler). These claims of Twitter's insensitivity towards content are also reflected in assertions that Twitter does not take reported instances of abuse and threats seriously (Perez 2017). In addition to attempting to better moderate content through more specific and clear regulations, in September of 2019, Twitter announced that it would begin integrating AI focused specifically on religious hate speech. However, critics "see the narrower scope as a retreat" or means by which Twitter attempts to appease and attenuate public concern while evading accountability (Harrison).

Section II: The Problems

In this section, I further explain Twitter, Inc.'s key speech issues in the context of content moderation, and how Twitter has responded to public criticism.

A. Content Moderation Measures' Lack of Transparency and Accessibility

Twitter has begun specifying and disclosing certain elements of its content moderation process, but until recently this lack of transparency caused widespread public distrust and skepticism over the company's efforts to moderate and promote free speech. In 2018, "when asked whether Twitter would consider releasing its full set of guides, manuals, documentation, tutorials, training materials and all other materials given to its reviewers or a justification for why it believes this material cannot be released, the company did not respond" (Leetaru, "Is Twitter Really Censoring Free Speech?"). Similarly, in 2019 "Twitter declined to comment on how many moderators it employs and how many more, if any, it would add" and did not disclose if new content moderation projects were underway (Harrison). The lack of information available on the internal workings of Twitter's speech moderation indirectly hinders free speech on the platform through attempting to control the discourse thereof. Because this information is private, people lack the knowledge to discuss specific shortcomings in Twitter's approach to content moderation, and instead can only propound that it is inherently suspect of Twitter to withhold this information, and is likely a calculated effort to evade accountability.

Moreover, "Even simple questions like the percent of Twitter's accounts that are bots and how much of its content is automatically generated are complete unknowns" (Leetaru, "Is Twitter Really Censoring Free Speech?"). The lack of available statistics on this topic further

serves to hinder free speech by undermining people's trust in the platform, thereby causing them to feel less secure in their usage of the app. In response to these concerns, Twitter has created an interactive, biannual Transparency Report (Twitter Transparency Report) which thoroughly outlines its policies and rules. However, this report still lacks content moderation statistics. Furthermore, Twitter employees have attempted to justify the inconsistencies in its content moderation, asserting that "'as hateful conduct and online abuse continue to evolve, our efforts to combat this behavior must also evolve; our work in this space will never be done'" and that "'it's like, now this is abuse, and now this is abuse, and now *this* is abuse—when what seems like abuse to the outside world is much more straightforward'" (Kosoff); by coming forward with this information about the complications of content moderation, Twitter attempts to regain the public's trust and be more transparent.

B. Inconsistent Response Time to Abuse and/or Harassment Reports

Despite the carefully outlined guidelines on reporting harassment and abuse (Twitter Rules and policies: Abusive behavior), there exist multiple accounts that "Twitter is slow or unresponsive to harassment reports until they're picked up by the media" and gain public attention (Warzel). These instances of harassment and abuse usually are towards women who are then encouraged by Twitter to report these instances; however "despite having policies that explicitly state that hateful conduct and abuse will not be tolerated on the platform, Twitter appears to be inadequately enforcing these policies when women report violence and abuse", and Twitter's inaction can cause "a level of mistrust and lack of confidence in the company's reporting process" (Amnesty International). Multiple claims of Twitter's disregard and dismissal of instances of harassment or abuse not only can cause distrust of the platform and disincentivize people from speaking freely about controversial topics for fear of harassment, but may also negatively impact victims' mental health. In her Forbes article, technology and digital culture reporter Fruzsina Eordogh reflects on her "at least half a dozen Twitter mobbings", contending that over the years, "Twitter has made *some* progress on this issue", and has developed settings to help users deal with harassment, likely in response to public criticism (Eordogh). Further, in November of 2016 Twitter "rolled out a keyword filter and a mute tool for conversation threads, as well as a 'hateful conduct' report option" Warzel. It also seems that this issue is somewhat tied to the aforementioned issue on Twitter's lack of content moderation transparency: because statistics on content moderators are unavailable, it is unclear if Twitter fails to respond appropriately to harassment because it lacks content moderators, or if the company fails to respond because it has other greater priorities.

Section III: Policy Recommendations

In this section, I suggest policy recommendations for the aforementioned problems to improve content moderation.

A. Increased Transparency, Disclosure, and Accessibility around Content Moderation

Increasing transparency and discussion around Twitter's currently somewhat opaque content moderation practices will aid in the promotion and protection of speech on the platform by empowering users to learn more about the company's regulatory processes and statistics. When users and the public have access to this information, such individuals are then able to give better feedback and criticize flaws in a manner more conducive to progress, for at the moment the

discourse on Twitter's content regulation is marked by suspicion due to the company's oblique and vague treatment of this regulatory information. Furthermore, increased openness about Twitter's beliefs on free speech and content moderation could also help illuminate the service's fickle nature regarding content regulation, and convey to the public that this changeability or inconsistency is more calculated than disorganization, or apathy. According to Danielle Citron, a Twitter trust and safety partner, addressing harassment is a slow process because content moderators care deeply about getting each case right, "'They really do have their users' speech issues in mind in a way that's very holistic'" (Kosoff). Transparency may thus help promote a better understanding of Twitter's values and understanding of content, in addition to promoting discussion of fair content creation and regulation. One counter-argument to this suggestion might be that Twitter is unable to disclose certain user statistics to the public because it would require Twitter to break its privacy policy to its users. However, should this be the case, when asked by a news source to comment on the company's secrecy concerning a certain element of content moderation or user data, Twitter could explain that it would be a breach of its privacy policy to disclose, in a manner that would not construe the company as suspicious as it would appear if it failed to comment.

B. Prioritization of Harassment and/or Abuse Reports

Due to limited data available to the public, it is unclear what prevents Twitter from addressing reports of abuse and/or harassment in a more timely manner. It is possible that these reports are something that Twitter assays to consider holistically, thus leading to slow assessment of these reports. However, it is more likely moderators are overwhelmed by the number of harassment/abuse reports, and thus it becomes difficult for moderators to promptly handle— evidenced by a Twitter engineer's apology regarding a harassment case, in which he references "tweets for harassment reports falling through the cracks of the company's reporting system" (Warzel). Increasing the response rate to harassment/abuse reports will require hiring more content moderators, obtaining reliable technology using artificial intelligence, or perhaps a more robust reporting system with better resources for people who are victims of online abuse/harassment. Although it may be costly, it's important to better address this problem because this issue of speech can have serious mental health consequences, and cause users to engage less with Twitter, or feel uncomfortable contributing to the "global, public conversation" that Twitter aims to foster (Twitter Topics; Company).

Section IV: Conclusion

Although upon closer inspection, Twitter's at times confusing content moderation practices are rooted in the company's desire to approach each new situation holistically, on the surface Twitter's moderation efforts are ridiculed by most media as blasé about extreme and offensive types of speech, but also abusive of its power. This criticism, as well as shortcomings in Twitter's moderation practices such as the slow response time to abuse and harassment reports, have caused users to distrust Twitter. To increase clarity surrounding Twitter's content moderation practices and address these above-referenced issues of transparency and approach to abuse and harassment reports, the memo recommends that Twitter prioritize response to abuse and harassment reports and better online tools for addressing harassment, as well as a much greater degree of transparency about content moderation and user information.

Works Cited

Eordogh, Fruzsina. “Twitter's Anti-Harassment Tools, Reviewed.” *Forbes*, Forbes Magazine, 11 Mar. 2019, Retrieved 05 May 2020, from www.forbes.com/sites/fruzsinaeordogh/2019/03/11/twitters-anti-harassment-tools-reviewed/#5ab694211e13.

Halliday, Josh. “Twitter's Tony Wang: 'We Are the Free Speech Wing of the Free Speech Party'.” The Guardian, Guardian News and Media, 22 Mar. 2012, Retrieved 05 May 2020, from www.theguardian.com/media/2012/mar/22/twitter-tony-wang-free-speech.

Harrison, Sara. “Twitter and Instagram Unveil New Ways to Combat Hate-Again.” *Wired*, Conde Nast, 11 July 2019, Retrieved 05 May 2020, from www.wired.com/story/twitter-instagram-unveil-new-ways-combat-hate-again/.

Koebler, Jason, and Joseph Cox. “Twitter Has Started Researching Whether White Supremacists Belong on Twitter.” Vice, 29 May 2019, Retrieved 05 May 2020, from www.vice.com/en_us/article/ywy5nx/twitter-researching-white-supremacism-nationalism-ban-deplatform.

Kosoff, Maya. “‘Just an Ass-Backward Tech Company’: How Twitter Lost the Internet War.” *Vanity Fair*, Hive, 19 Feb. 2018, Retrieved 05 May 2020, from www.vanityfair.com/news/2018/02/how-twitter-lost-the-internet-war.

Leetaru, Kalev. “The Remarkable Reversal: How Companies Now Censor Governments.” *Forbes*, Forbes Magazine, 11 Jan. 2018, Retrieved 05 May 2020, from www.forbes.com/sites/kalevleetaru/2018/01/09/the-remarkable-reversal-how-companies-now-censor-governments/#13ff8ac66f97.

---. “Is Twitter Really Censoring Free Speech?” *Forbes*, Forbes Magazine, 12 Jan. 2018, Retrieved 05 May 2020, from www.forbes.com/sites/kalevleetaru/2018/01/12/is-twitter-really-censoring-free-speech/#75ba19fc65f5.

---. “Twitter 'Misinformation' Demo App Stirs Free Speech Questions.” RealClearPolitics, 26 Feb. 2020, Retrieved 05 May 2020, from www.realclearpolitics.com/articles/2020/02/26/twitter_misinformation_demo_app_stirs_free_speech_questions_142496.html.

Perez, Sarah. “Twitter Today Starts Enforcing New Rules around Violence and Hate.” *TechCrunch*, TechCrunch, 18 Dec. 2017, Retrieved 05 May 2020, from techcrunch.com/2017/12/18/twitter-today-starts-enforcing-new-rules-around-violence-and-hate/.

---. “Twitter Makes It Easier to See Enforcement Taken on Reported Tweets.” *TechCrunch*, TechCrunch, 17 Oct. 2018, Retrieved 05 May 2020, from techcrunch.com/2018/10/17/twitter-makes-it-easier-to-see-enforcement-taken-on-reported-tweets/.

“Toxic Twitter - The Reporting Process.” *Amnesty International*, Mar. 2018, Retrieved 05 May 2020, from www.amnesty.org/en/latest/research/2018/03/online-violence-against-women-chapter-4/.

“Transparency Report.” Twitter, Twitter, Retrieved 05 May 2020, from transparency.twitter.com/en.html.

Warzel, Charlie. “Twitter Is Still Dismissing Harassment Reports And Frustrating Victims.” *BuzzFeed News*, BuzzFeed News, 18 July 2017, Retrieved 05 May 2020, from www.buzzfeednews.com/article/charliewarzel/twitter-is-still-dismissing-harassment-reports-and.

“World Leaders on Twitter.” *Twitter*, Twitter, 5 Jan. 2018, Retrieved 05 May 2020, from blog.twitter.com/en_us/topics/company/2018/world-leaders-and-twitter.html.