TO: Mark Zuckerberg, Chief Executive Officer of Facebook
FROM: Sarah Chekfa
RE: Recommendations on the Subject of Platform Speech Regulation
DATE: March 30, 2018
SECTION: Palashi Vaghela (206)

I. Introduction

This policy memo will provide an overview of important speech problems confronted by social media platforms, and propose potential policy solutions accordingly. I will provide crucial background information into the issues of hate speech and fake news, outline the key positions taken by Facebook in confronting both these issues, and ultimately propose a policy recommendation that I think is best considering the current political climate and need for resolution of these two dominating problems.

Section II contextualizes the issues of speech facing social media platforms, detailing the essential meanings and implications of both hate speech and fake news. Section III details key policies Facebook has introduced to tackle both these issues, analyzing their impact and dissecting their depth. Section IV will suggest what I view as the best possible current solution to the issues of hate speech and fake news. In response to the issue of hate speech, I will suggest that Facebook strengthen online accountability and standardize hate speech moderation; in response to fake news I will suggest that Facebook invest in automatic hoax detection technologies. I conclude the policy memo in Section V, and in Section VII I cite the sources I used to research this policy memo.

II. Contextualizing the Issue: Speech Problems Facing Social Media Platforms

In this section, I will provide key background detailing fundamental speech problems facing social media platforms in the current networked digital age. Understanding the foundations and prevalence of these problems is crucial to understanding the current techno-political conversation on how best to resolve them. I divide this contextual section into three subsections, in which I will describe two of the most important speech problems: A. hate speech, and B. fake news.

A. Hate Speech

Hate speech is difficult to define in any context because there is no one, internationally accepted for it. In the context of social media platforms, however, it can be broadly conceived to entail any online communication that "denigrates people on the basis of their membership to a particular group" (Defy Hate Now). It typically takes the form of text (such as comments and statuses), images, and videos. Different platforms—such as Facebook, Instagram, Twitter, Youtube, and Reddit—hold differing, sometimes conflicting perspectives on what exactly constitutes hate speech. It is often difficult to monitor and regulate hate speech because the line between merely offensive speech and hate speech varies depending on whom one consults. The lack of accountability, stemming from the sense of anonymity and distance made possible by the very nature of online communication, can be considered a critical facilitator of hate speech on social platforms. Tech companies have been heavily criticized for not doing enough to combat hate speech. In some countries, the federal government has intervened in an attempt to alleviate the issue: in January, Germany, for example, put into effect a law that fines social media companies like Facebook and Twitter $60 million for failing to remove hateful content on their platforms within 24 hours (TechCo).

B. Fake News

Fake news, a term popularized by Donald Trump in the 2016 presidential election, comprises online "news stories that are false…[in which] the story itself is fabricated, with no verifiable facts, sources or quotes" (University of Michigan). Fake news contributes to a greater political macrocosm founded upon misinformation and disinformation (Lomas). A large percentage of fake news takes the form of propaganda, produced to intentionally mislead its audience; other instances of fake news are designed to be clickbait, written to elicit profit via the number of people who click on the stories. Fake news articles are often "designed to provoke an emotional response" in their readers to induce them to share the news; other times, they are spread by "bots," computer algorithms engineered to pretend to be real people while automatically proliferating such duplicitous content (University of Michigan). Their viral nature is precisely what makes fake news so dangerous—social platforms make it possible to share such facetious stories quickly and virtually effortlessly. Experts are divided upon the degree to which such false stories contributed to the election of Donald Trump (Lewis), with Facebook itself announcing that 126 million people in the United States may have been exposed to fake stories produced by Russian-government-backed agents during the election cycle (Romm, Wagner). Some argue that the business models and algorithms which form the structural base of these social platforms incline them towards the proliferation of fake news (Abeshouse).

III. Assessing and Analyzing the Implications of Facebook's Current Policies

In this section, I will review key policies Facebook has implemented to address the issues of A. hate speech and B. fake news, respectively. I will also assess the implications and impact of these policies, analyzing how they work to resolve the free speech problems they seek to address.

A. What Facebook Has Done to Reduce Hate Speech

Facebook defines hate speech as "an attack, such as a degrading generalization or slur, targeting a 'protected category' of people, including one based on sex, race, ethnicity, religious affiliation, national origin, sexual orientation, gender identity, and serious disability or disease." Such posts considered "hate speech" are cause for removal from the site (Carlsen, Haque).

In Europe, in 2016, Facebook signed a Code of Conduct on illegal hate speech, promising to review the majority of reported hate speech within 24 hours, and remove them within that time frame should they be verified as hate speech. And while the Commission lauded them for their "steady progress" in removing hate speech, other studies have found that Facebook is failing to adequately monitor such hateful speech (Lomas). While Facebook may seem to posit an all-encompassing definition of hate speech in its community standards, in practice these standards are unequally and misguidedly enforced on the platform. ProPublica conducted a study, finding that this uneven, flawed enforcement of hate speech causes hateful posts to stay up and innocent posts discussing racism or sexism to be taken down, erroneously flagging them as hate speech (Tobin, et al.).

In response to these issues, Facebook has promised to increase the number of staff in charge of safety and content moderation issues on the platform in 2018. But because Facebook has close to 2 billion users and receives "millions of reports" flagging hateful content every week, many believe this is not enough. The company has historically relied mainly on algorithms to maintain community standards, yet it has promised to move in a more human direction. Zuckerberg recently announced that he wants to "make it simpler to report problems...faster for reviewers to determine which posts violate our standards and easier for them to contact law enforcement if someone needs help" (Lunden).

B. What Facebook Has Done to Fight Fake News

In 2016, Mark Zuckerberg considered "crazy" the very notion that fake news facilitated Donald Trump's victory in the presidential election (Wagner). But after steady public outrage, in early 2017 Facebook pioneered the "Disputed Flags" tool, which marked fake news articles as "disputed." In order for this label to be attached to a story, however, a fake story had to go through an extended process. Either Facebook's users had to report the story, or Facebook's software had to detect it was fake. Then, Facebook would send the story to third party organizations, such as Snopes and Politifact, who had agreed to provide it with free fact-checking. And finally, if two of those fact-checkers verify its falsity, the "Disputed Flags" label was attached. Critics argued that this process was slow—a story could go days without being labelled as disputed; by the time it is finally labeled as "disputed," thousands of readers could have already been exposed to it. The initiative was also criticized for being too soft on the very real issue of fake news—after all, "disputed" "makes [a fake news article] sound like a bar debate about the NBA's MVP," not the indisputably fictitious fabrication that it is, argues Peter Kafka of *Recode* (Kafka).

At the end of December 2017, Facebook halted this policy entirely, replacing it with the Related Articles tool (Lyons). After a year of testing, they realized that Disputed Flags actually "buried critical information that explained the inaccuracies, and could backfire by entrenching a person's false beliefs" (Ong). The new Related Articles tool provides people with more context about the fake story, reducing the likelihood that it is shared. "Using language that is unbiased and non-judgmental helps us to build products that speak to people with diverse perspectives," Facebook states (Smith). The process is also much faster, requiring only one fact checker rather than two.

IV. Recommendations: How Can Facebook Better Address Speech Issues?

In this section, I will propose clear policy recommendations for Facebook to address the issues of A. hate speech and B. fake news, respectively. I propose that hate speech be tackled by strengthening online accountability and standardizing hate speech moderation; and that fake news be confronted with automatic hoax detection.

A. Addressing Hate Speech

I. Strengthen Online Accountability

While Facebook should continue to hire more individuals to serve on its safety and community standards team, there is definitely more that can be done about the problem of hate speech. Facebook can work on strengthening online accountability through "real-name registration"—i.e., requiring that individuals provide their true identity when registering with Facebook. In this fashion, Facebook can ensure that individuals are held accountable for what they post and share on the platform. Since "the likelihood that people will engage in worse behavior [increases] if they believe their actions are anonymous and not likely to be made public," moving towards verifying online identities will make people less likely to post offensive comments that threaten the safety of marginalized, protected minorities (West).

II. Standardize Hate Speech Moderation

Facebook should also adopt a more standardized, equal approach to hate speech moderation. It should consult with social justice experts and sociological researchers to update its training manuals to update its classifications of certain demographics and social groups and ensure that minorities and other marginalized individuals are protected at all times. By reviewing how categories of people (i.e. race, social class, sex, gender identity, religion, etc.) are classified and protected, and updating community standards to reflect the very real controversies and threats confronted by marginalized people in the world, Facebook moves towards a more inclusive community (Carlsen, Haque).

B. Addressing Fake News

I. Automatic Hoax Detection

While Facebook should continue the use of its Related Articles to fight against fake news, there is much more that can be done to tackle the issue. In order to expand the fight against fake news on its platform, Facebook should "invest in technology to find fake news and identify it for users through algorithms and crowdsourcing" (West). When the identification of fake news becomes automated utilizing "public interest algorithms," the public can be further protected from misinformation and democracy is better preserved, says former FCC Commissioner Tom Wheeler. Such automatic hoax detection ensures that fake news is quickly pinpointed and labelled accordingly so as to minimize its dangerous impact on the people (Warrington).

V. Conclusion

The issues of hate speech and fake news are complicated ones that plague our digital society. The enduring nature and debated meaning of free speech has ensured that these issues are subject to immense controversy, especially given the heavily polarized political climate to which America is privy. It is crucial that Facebook invest in automatic hoax detection, standardize its hate speech moderation standards, and strengthen online accountability in order to confront the issues of fake news and hate speech. But even the adoption of these policies will not forever stave off these issues as social platforms develop and the free speech politics defining us as a people continues to change. Fake news and hate speech posit conversations that will never quite end, but these are the first steps Facebook must take to attempt to curtail them on its platform. This will bring us closer to realizing our goals for a free, democratic society.

VI. Bibliography

Below I cite the sources I utilized to pen this policy memo.

Abeshouse, Bob. “Troll Factories, Bots and Fake News: Inside the Wild West of Social Media.” | *Al Jazeera*, Al Jazeera, 8 Feb. 2018, www.aljazeera.com/blogs/americas/2018/02/troll-factories-bots-fake-news-wild-west-social-media-180207061815575.html.

Carlsen, Audrey, and Fahima Haque. “What Does Facebook Consider Hate Speech? Take Our Quiz.” *The New York Times*, The New York Times, 13 Oct. 2017, www.nytimes.com/interactive/2017/10/13/technology/facebook-hate-speech-quiz.html.

“INTRODUCTION TO HATE SPEECH ON SOCIAL MEDIA.” *Defy Hate Now*.

Kafka, Peter. “Facebook Has Started to Flag Fake News Stories.” *Recode*, Recode, 4 Mar. 2017, www.recode.net/2017/3/4/14816254/facebook-fake-news-disputed-trump-snopes-politifact-seattle-tribune.

Lewis, Helen. “Did Fake News on Facebook Swing the US Election?” *New Statesman*, www.newstatesman.com/world/2016/11/did-fake-news-facebook-swing-us-election.

Lomas, Natasha. “Fake News Is an Existential Crisis for Social Media .” *TechCrunch*, TechCrunch, 22 Mar. 2018, techcrunch.com/2018/02/18/fake-news-is-an-existential-crisis-for-social-media/.

Lomas, Natasha. “Facebook, Twitter, YouTube Praised for ‘Steady Progress’ Quashing Illegal Hate Speech in Europe.” *TechCrunch*, TechCrunch, 22 Jan. 2018, techcrunch.com/2018/01/19/facebook-twitter-youtube-praised-for-steady-progress-quashing-illegal-hate-speech-in-europe/.

Lunden, Ingrid. “Facebook to Add 3,000 to Team Reviewing Posts with Hate Speech, Crimes, and Other Harming Posts.” *TechCrunch*, TechCrunch, 3 May 2017, beta.techcrunch.com/2017/05/03/facebook-to-hire-3000-to-review-posts-with-hate-speech-crimes-and-other-harming-posts/?_ga=2.82536019.418731031.1522428847-260663223.1520030071.

Lyons, Tessa. “News Feed FYI: Replacing Disputed Flags with Related Articles.” *Facebook Newsroom*, newsroom.fb.com/news/2017/12/news-feed-fyi-updates-in-our-fight-against-misinformation/.

Ong, Thuy. “Facebook Found a Better Way to Fight Fake News.” *The Verge*, The Verge, 21 Dec. 2017, www.theverge.com/2017/12/21/16804912/facebook-disputed-flags-misinformation-newsfeed-fake-news.

“Research Guides:” *Library Research Guides*, guides.lib.umich.edu/fakenews.

Romm, Tony, and Kurt Wagner. “Facebook Says 126 Million People in the U.S. May Have Seen Posts Produced by Russian-Backed Agents.” *Recode*, Recode, 30 Oct. 2017, www.recode.net/2017/10/30/16571598/read-full-testimony-facebook-twitter-google-congress-russia-election-fake-news.

Smith, Jeff. “Designing Against Misinformation – Facebook Design – Medium.” *Medium*, Facebook Design, 20 Dec. 2017, medium.com/facebook-design/designing-against-misinformation-e5846b3aa1e2.

“The Tumultuous Relationship Between Social Media and Hate Speech.” *TechCo*, 10 Jan. 2018, tech.co/relationship-social-media-hate-speech-2018-01.

Tobin, Ariana, et al. “Facebook's Uneven Enforcement of Hate Speech Rules Allows Vile Posts to Stay Up.” *ProPublica*, www.propublica.org/article/facebook-enforcement-hate-speech-rules-mistakes.

Wagner, Kurt. “Mark Zuckerberg Says It's 'Crazy' to Think Fake News Stories Got Donald Trump Elected.” *Recode*, Recode, 11 Nov. 2016, www.recode.net/2016/11/11/13596792/facebook-fake-news-mark-zuckerberg-donald-trump.

Warrington, Anna. “Public Interest Algorithms Have Been Proposed to Tackle the Problems Created by Social Media Algorithms.” *Futures Centre*, 2 Nov. 2017, thefuturescentre.org/signals-of-change/205047/public-interest-algorithms-have-been-proposed-tackle-problems-created.

West, Darrell M. “How to Combat Fake News and Disinformation.” *Brookings*, Brookings, 18 Dec. 2017, www.brookings.edu/research/how-to-combat-fake-news-and-disinformation/.