Spring 4-27-2022

Breaking Things: Origins and Consequences of Racialized Hate Speech on Facebook

Katherine Herrick
Macalester College, katie.ann.herrick@gmail.com

Breaking Things:

Origins and Consequences of Racialized Hate Speech on Facebook

Katie Herrick

Professor James von Geldern, International Studies

April 27, 2022

ABSTRACT

This thesis seeks to bring attention to the ways in which the effects of hate speech--specifically racialized hate speech--transcends digital platforms. It will begin by connecting the phenomenon of racialized hate speech on Facebook to specific psychological tendencies that the company consciously amplifies for its own financial benefit. The first chapter interrogates the common narrative that violent rhetoric indicates a flaw in the platform's design, instead arguing that proliferation of such content is encouraged by Facebook's algorithm. From there, the second chapter examines what happens when a technology giant leverages human psychology for corporate greed. A true worst-case scenario, the Rohingya genocide in Myanmar elucidates Facebook's negligent behavior and illustrates the consequences of failing to proactively mitigate hate speech. Finally, the third chapter discusses existing and proposed efforts to regulate Facebook and similar platforms. As an issue that encompasses ethical dilemmas, policy predicaments, and business implications, reducing the prevalence of racialized hate speech on Facebook poses challenges for all regulatory actors, from the United Nations, to sovereign states, to the corporation itself. In the end, the most effective means of protecting human rights on digital networks may not rest upon the United Nations, nor individual nations, nor private corporations, but upon social media users themselves.

ACKNOWLEDGEMENTS

This thesis could not have happened without the support and advice of countless mentors. I would like to extend my immense (and somewhat surprised) gratitude to Professor Gilbert Rodman for agreeing to read a very long paper, written by a student he does not know, who attends a college he does not teach at, simply because I sent him a nice email. I would also like to thank Professor Nadya Nedelsky. She was one of the first to read a chapter of this thesis, and I am so grateful for her thoughtful advice, unwavering support, and overall brilliance; I am lucky to have had the opportunity to learn from her. Thank you to Professor James von Geldern, for being an outstanding academic advisor and life-changing personal mentor. I cannot imagine what my time at Macalester would have looked like without JVG in it. I also want to express my unending admiration for the incredible Janessa Cervantes, whose unmatched organizational skills and perpetually sunny disposition made the thesis process both far easier and far more enjoyable.

In addition to my thesis panel, I would be remiss if I did not thank both of my parents. Thank you, Mom and Dad, for instilling within me a love of learning and the work ethic required to turn that love into a 100-page paper. And thank you for dealing with my stress, uncertainty, and messy drafts throughout the process—you are the most patient and enthusiastic cheerleaders a daughter could ask for! Perhaps most of all, I want to thank Grandpa. Growing up on his stories of reporting on Vietnam, the Chicago riots, environmental degradation, the JFK assassination, and corrupt policing showed me the way that words can speak truth to power and be used to create real change. Though he was never able to read this thesis, I hope he would be proud of it. You were right, Grandpa; it's in my genes.

TABLE OF CONTENTS

INTRODUCTION

I was born at the turn of the millennium, in the midst of Y2K paranoia. The rapid change of technology has defined my entire generation, for within our short lifetimes we have witnessed the invention of everything from wifi, to smartphones, to cryptocurrency. Though these transformations have certainly changed the physical objects in our lives, they have also fundamentally altered our social relationships. Technology enables us to transcend the boundaries of time and space; people can interact with individuals from other continents, and even from distant points in their past.

This hyperconnectivity exponentially increased with the advent of social media. In 2012, a piercing shriek shattered the tranquility of my childhood home, drawing my family into the room where my older sister sat glued to our old PC monitor. She had recently made an account on a new website called 'Facebook,' and received a highly unexpected friend request. My family had a running joke for years about one Federico Vargas,[1] a kindergarten classmate of my sister who committed the unforgivable sin of eating her coloring crayons on the first day of school. Twelve years and hundreds of miles later, Federico and Amy had found each other on Facebook. They chatted back and forth for a bit, and though he claimed not to remember the incriminating incident, he did promise her a consolatory box of crayons.

It was wholesome, serendipitous interactions such as this one that defined Facebook in its early years. Thousands, then millions, and now billions of people flock to the platform every day for everything from news to entertainment, relegating digitized social interaction from a novelty to a normality. In fact, I, as someone with virtually no social media footprint, regularly receive

1. Federico's name has been changed to respect his privacy.

strange looks, surprised exclamations, and occasionally scornful derision when people discover I do not and have never had an account on Facebook, Instagram, Twitter, TikTok, and the like.

I am no luddite—I recognize the potential for positive change that Facebook and its kin possess. Around the time my middle school classmates began making social media profiles, pro-democracy protesters throughout the Middle East and North Africa used those same platforms to organize the Arab Spring. A study by the University of Washington finds that social media in general and Facebook in particular facilitated and enhanced the scope, reach, and efficacy of the movement. The platform's algorithms and broad network allowed messages of justice to spread quickly and widely; "evidence suggests that social media carried a cascade of messages about freedom and democracy across North Africa and the Middle East, and helped raise expectations for the success of political uprising."[2] Videos depicting protests went viral, and opposition Facebook pages gained thousands upon thousands of followers. Social justice activists around the world learned from the Arab Spring's strategy, utilizing social media's virality and decentralization to demand transparency and hold those in power accountable. These tactics are not deployed only in 'distant lands,' either. I have called the Twin Cities home for four years, a period strongly shaped by protest movements like Black Lives Matter and Stop Line 3. Organizers routinely use Facebook and Instagram not only to raise awareness for their causes, but also to capture injustices, communicate actions, and share safety tips. Modern social justice campaigns cannot be divorced from social media platforms.

---

2. Catherine O'Donnell, "New Study Quantifies Use of Social Media in Arab Spring," *University of Washington News,* September 12, 2011, https://www.washington.edu/news/2011/09/12/new-study-quantifies-use-of-social-media-in-arab-spring/.

However, the higher the pedestal, the harder the fall. Over the last five years, Facebook has weathered scandal after scandal, drawing criticism for everything from its mismanagement of user data, to the proliferation of misinformation, to damning allegations of human rights abuse. The company's tumble from grace gained real momentum in 2018 with the revelation of the Cambridge Analytica scandal. To give Donald Trump an edge in the 2016 US presidential election, the titular political consulting firm conspired with Facebook to mine the data of over 50 million users without their knowledge or permission.[3] It raised an alarming question: how much would Facebook sacrifice at the altar of profit?

Just a few years later, whistleblower Francis Haugen leaked internal corporate documents to the Wall Street Journal detailing even more damning transgressions. The Facebook Files, as they came to be known, testify that the company regularly values company growth over user wellbeing.[4] Executives make conscious and repeated decisions to maximize revenue even when doing so comes at the expense of truth, equality, and safety on their platform, especially for those who possess marginalized and/or minority identities. Among other issues, the documents highlighted the prevalence and effects of hate speech on Facebook.

Of course, you don't need to read secret company memorandums to realize that the platform hosts a truly frightening amount of bigotry and vitriol—anyone with a Facebook account knows that. The depth and extent of that content's impact, however, may not be

---

3. Matthew Rosenberg, Nicholas Confessore, and Carole Cadwalladr. "How Trump Consultants Exploited the Facebook Data of Millions." *The New York Times,* March 17, 2018. https://www.nytimes.com/2018/03/17/us/politics/cambridge-analytica-trump-campaign.html.

4. Lima, Cristiano. "A Whistleblower's Power: Key Takeaways from the Facebook Papers." *The Washington Post,* October 26, 2021. https://www.washingtonpost.com/technology/2021/10/25/what-are-the-facebook-papers/.

immediately apparent to the average user. Prejudicial language that originates on online platforms can nevertheless precipitate offline consequences. We as a society are only just beginning to grasp the tangible implications of digital social dynamics and what they mean for personal safety and collective wellbeing. After all, not every Facebook user is as well-intentioned as Federico Vargas.

This project began as an attempt to compile an encyclopedic account of Facebook's sordid relationship with human rights, everything from privacy to censorship to hate speech. However, I soon realized that such an endeavor could consume an entire career; the tech giant's litany of transgressions stretches so long that any singular attempt to catalog them would be doomed to obsolescence or incompletion or both. So, I chose to narrow my scope and focus specifically on hate speech. Yet even that proved too vast a topic, encompassing discrimination against every identity, characteristic, or affiliation a person could possibly possess. Ultimately, current events guided this thesis to its final subject. The last few years have seen a spate of attacks against people of Asian descent as a misinformed backlash against the COVID-19 pandemic; rampant police brutality against Black folks and harsh crackdowns against those who dare speak out against it; state-sponsored desecration of Indigenous lives and land. These injustices feel particularly acute in the Twin Cities, but repeat and reverberate across the world. Racist violence both feeds off and feeds into racist rhetoric online. World leaders employ racially-coded dog whistles in official posts and supremacist groups flourish, finding refuge in the murky realm of cyberspace and utilizing social media platforms to espouse bigoted disinformation. Nowhere do such antagonists seem to feel more at home than Facebook.

Yet common discourse often dissociates physical action, virtual speech, and the social media platforms that link the two. This thesis seeks to bring attention to the ways in which the

effects of hate speech—specifically racialized[5] hate speech—transcend digital platforms. As with all discussions regarding bigoted or inflammatory language, the issue of hate speech on Facebook complicates the ostensibly universal right to free expression. Should individuals who express racist, misogynist, or otherwise prejudicial views be protected from censorship or punishment? A landmark decision in this legal quandary has its origins in the Twin Cities, in fact. The 1992 Supreme Court case *R.A.V. v. St. Paul* concerned an incident in which a young man named Robert A. Viktora burned a cross in the yard of a Black family who lived nearby. However, the Supreme Court was not interested in the legal or moral significance of cross-burning; rather, they took issue with the St. Paul ordinance that Viktora was on trial for violating. The City of St. Paul forbade placing swastikas or a burning cross anywhere "in an attempt to arouse anger or alarm on the basis of race, color, creed, or religion," a law the Supreme Court struck down as unconstitutionally content-based.[6] According to Justice Scalia's opinion, the ordinance impermissibly discriminated against Viktora's speech solely because of the sentiment it expressed. Subsequent related decisions carved out narrow exemptions in the *R.A.V.* ruling, allowing content-based censorship for "exceptionally virulent" forms of discriminatory expression or speech that poses a "true threat."[7]

---

5. For the purposes of this paper, racialization refers to the process by which 'racial meaning' is assigned to a person, practice, relationship, or social group. It draws on the work of Bianca Gonzalez-Sobrino, particularly her 2018 article entitled "Exploring the Mechanisms of Racialization beyond the Black-White Binary" (https://doi.org/10.1080/01419870.2018.1444781). Following her approach, this paper also refers to racialization beyond the constructed black-white binary so common in the United States, and recognizes that 'race' is not a valid scientific category, nor a static and immutable identity; rather, it is a process imbricated with power and privilege. Used here, racialized hate speech refers to language that expresses prejudice or abuse towards an individual or collection of individuals based on their (perceived) racial or ethnic differences.

6. David A. May, "R.A.V. v. St. Paul (1992)," *The First Amendment Encyclopedia,* accessed March 26, 2022, https://www.mtsu.edu/first-amendment/article/270/r-a-v-v-st-paul.

7. Robert A. Kahn, "Virginia v. Black (2003)," *The First Amendment Encyclopedia,* accessed March 26, 2022, https://mtsu.edu/first-amendment/article/271/virginia-v-black.

Yet these categories remain largely ambiguous, particularly in the virtual world. Does a tweet announcing that "when the looting starts, the shooting starts" represent a true threat?[8] Is a Facebook group for white supremacists an exceptionally virulent form of discriminatory expression? Would a social media company's policy banning racial slurs comprise an unconstitutionally content-based limitation of free speech? *Chaplinsky v. New Hampshire* established legal precedent for restricting so-called "fighting words" meant to incite violence, a designation that would certainly seem to apply to racialized hate speech on digital platforms.[9] However, just seven years later, the 1949 *Terminiello v. Chicago* case constricted the definition of fighting words to encompass only expressions that create a "clear and present danger."[10] This suggests that fighting words cannot exist on social media. After all, if only 'sticks and stones' can break bones, what threat could virtual words possibly pose?

As this paper will demonstrate, hate speech on social media can incite real violence, with consequences that stretch far beyond the ethereality of cyberspace. Facebook's vision of an inextricably interconnected world—the very vision that secured the company's social, commercial, and technological dominance—may endanger the safety and wellbeing of marginalized groups. Yet heavy-handed regulation of social media content could easily engender the same problems; censorship seldom benefits minorities. With corporations like Facebook[11]

---

8. Donald Trump, Twitter post, May 28, 2020. Due to the suspension of Trump's Twitter account, the original tweet is unavailable.

9. "Fighting Words." *Legal Information Institute,* accessed March 26, 2022. https://www.law.cornell.edu/wex/fighting_words#:~:text=Fighting%20words%20are%20words%20meant,immediate%20breach%20of%20the%20peace.

10. Ibid.

11. On October 28, 2021 CEO Mark Zuckerberg announced that Facebook, the company known for its titular platform as well as Messenger, Instagram, and WhatsApp, would be rebranded as 'Meta.' This change happened in the midst of writing and is not reflected throughout the paper. Instead, 'Facebook' will refer here to both the platform and the larger corporation.

already entrenched within both the global economy and the global psyche, humanity is left retroactively scrambling to determine how to design, resource, and regulate social media so as to maximize its benefits while mitigating its harms. This thesis seeks to establish the salience of these issues, exploring how the tangible impacts of Facebook's underlying architecture cause violence to racial(ized) minorities across the world. It will begin by connecting the phenomenon of racialized hate speech on Facebook to specific psychological tendencies that the company consciously amplifies for its own financial benefit. The first chapter interrogates the common narrative that violent rhetoric indicates a flaw in the platform's design, instead arguing that proliferation of such content is encouraged by Facebook's algorithm. From there, the second chapter examines what happens when a technology giant leverages human psychology for corporate greed. A true worst-case scenario, the Rohingya genocide in Myanmar elucidates Facebook's negligent behavior and illustrates the consequences of failing to proactively mitigate hate speech. Finally, the third chapter discusses existing and proposed efforts to oversee Facebook and similar platforms. As an issue that encompasses ethical dilemmas, policy predicaments, and business implications, reducing the prevalence of racialized hate speech on Facebook poses challenges for all regulatory actors, from the United Nations, to sovereign states, to the corporation itself. In the end, the most effective means of protecting human rights on digital networks may not rest upon the United Nations, nor individual nations, nor private corporations, but upon social media users themselves.

INTRODUCTION BIBLIOGRAPHY

"Fighting Words." *Legal Information Institute,* accessed March 26, 2022. https://www.law.cornell.edu/wex/fighting_words#:~:text=Fighting%20words%20are%20words%20meant,immediate%20breach%20of%20the%20peace.

Kahn, Robert A. "Virginia v. Black (2003)." *The First Amendment Encyclopedia,* accessed March 26, 2022. https://mtsu.edu/first-amendment/article/271/virginia-v-black.

Lima, Cristiano. "A Whistleblower's Power: Key Takeaways from the Facebook Papers." *The Washington Post,* October 26, 2021. https://www.washingtonpost.com/technology/2021/10/25/what-are-the-facebook-papers/.

May, David A. "R.A.V. v. St. Paul (1992)." *The First Amendment Encyclopedia,* accessed March 26, 2022. https://www.mtsu.edu/first-amendment/article/270/r-a-v-v-st-paul.

O'Donnell, Catherine. "New Study Quantifies Use of Social Media in Arab Spring." *University of Washington News,* September 12, 2011. https://www.washington.edu/news/2011/09/12/new-study-quantifies-use-of-social-media-in-arab-spring/.

Rosenberg, Matthew, Nicholas Confessore, and Carole Cadwalladr. "How Trump Consultants Exploited the Facebook Data of Millions." *The New York Times,* March 17, 2018. https://www.nytimes.com/2018/03/17/us/politics/cambridge-analytica-trump-campaign.html.

Trump, Donald. Twitter post, May 28, 2020.

CHAPTER ONE:
ORIGINS OF RACIALIZED HATE SPEECH ON FACEBOOK

Few successful decisions ever begin with a Tuesday night bender in a college dorm room. Yet those are the humble origins of Facebook, which began when founder and CEO Mark Zuckerberg decided to create a forum for his fellow Harvard students to rate the attractiveness of their female classmates.[1] The result, FaceMash, materialized at a time when "a new kind of Internet [was] emerging — one more about connecting people to people than people to Websites."[2] Sure enough, Zuckerberg's creation underwent rapid expansion before going public in 2012, less than a decade after its conception. At $104 billion, Facebook boasted the largest technology IPO in history, and the third-largest IPO of any industry, ever.[3] Much of the company's appeal stemmed from its competitive streak and innovative edge, characteristics that define it to this day.

But the very qualities that enabled Facebook's evolution from nihility to ubiquity possess darker sides as well. Moral frameworks cannot keep up with the rate of technological advancement, leaving society to reactively grapple with unforeseen consequences. In the last several years especially, Facebook has come under intense scrutiny for the ethical implications of its business practices. The Cambridge Analytica scandal that came to light in 2018 opened the world's eyes to the cooptation of user data, and censorship issues in nations like Thailand, Russia, and China have raised questions of complicity with authoritarian regimes. However, in

---

1. Sheera Frenkel and Cecilia Kang, *An Ugly Truth,* (New York: Harper Collins, 2021).

2. Claire Hoffman, "The Battle for Facebook," *Rolling Stone,* September 15, 2010, https://www.rollingstone.com/culture/culture-news/the-battle-for-facebook-242989/l.

3. Julianne Pepitone, "Facebook Trading Sets Record IPO Volume," CNN Money, May 18, 2012, https://money.cnn.com/2012/05/18/technology/facebook-ipo-trading/index.htm.

addition to violating human rights such as those of privacy and expression, Facebook has also served as a platform where hate speech can flourish.

HATE SPEECH

Though a seemingly straightforward concept, hate speech as a legal category actually remains quite ambiguous. In fact, the United Nations Special Rapporteur on free expression observed in a 2019 report that no definition for it exists within conventional international law.[4] The United States fails to define hate speech as well, though the Federal Bureau of Investigation does delineate hate *crimes.*[5] As the federal webpage on the subject reminds its viewers, "hate itself is not a crime—and the FBI is mindful of protecting freedom of speech and other civil liberties."[6] In lieu of a legal consensus, Facebook has decided its own corporate definition. Hate speech, according to Facebook's Community Standards, consists of a "direct attack against people…on the basis of…protected characteristics: race, ethnicity, national origin, disability, religious affiliation, caste, sexual orientation, sex, gender identity, and serious disease."[7] Yet this black-and-white explanation belies vast swaths of gray area subject to seemingly arbitrary regulation. A ProPublica investigation revealed that Facebook "gives users broader latitude when they write about 'subsets' of protected categories. White men are considered a group because

4. United Nations General Assembly, *Promotion and Protection of the Right to Freedom of Opinion and Expression,* A/74/486, October 9, 2019, https://www.ohchr.org/Documents/Issues/Opinion/A_74_486.pdf.

5. Federal Bureau of Investigation, "Hate Crimes," accessed September 18, 2021, https://www.fbi.gov/investigate/civil-rights/hate-crimes.

6. Federal Bureau of Investigation, "Hate Crimes."

7. Meta, "Hate Speech," accessed May 22, 2021, https://www.facebook.com/communitystandards/hate_speech.

both traits are protected, while female drivers and black children…[are] not protected."[8] Amidst this opacity, hate speech flourishes on the platform. Approximately 5 million posts are flagged as hate speech every single day, and that only counts content that violates Facebook's own narrowly-defined Community Standards.[9] The real number, asserts the Anti-Defamation League, is likely far higher.

What, then, allows for this proliferation? According to its mission statement, Facebook strives to "give people the power to build community and bring the world closer together."[10] Derogatory, intimidating, and exclusive language antithesizes that mission; by the company's own admission, "people [can] use their voice and connect more freely when they don't feel attacked on the basis of who they are."[11] If Facebook is committed to connecting the world, why does its platform foster virulent hate speech that divides rather than unites? In this section, I will argue that violations of human rights—particularly inflammatory hate speech targeting racial minorities—do not constitute a failure of Facebook's technology. Rather, it is a consequence of the company's very business model, which intentionally preys upon human psychology in order to maximize profits. In short, racialized hate speech is a predictable, inevitable byproduct of Facebook's success.

---

8. Julia Angwin and Hannes Grassegger, "Facebook's Secret Censorship Rules Protect White Men From Hate Speech But Not Black Children," *ProPublica,* June 28, 2017, https://www.propublica.org/article/facebook-hate-speech-censorship-internal-documents-algorithms.

9. Anti-Defamation League, "Facebook's Hate Speech Problem is Even Bigger Than We Thought," accessed September 18, 2021, https://www.adl.org/blog/facebooks-hate-speech-problem-is-even-bigger-than-we-thought.

10. Meta, "Our Mission," accessed March 15, 2022, https://about.facebook.com/company-info/.

11. Meta, "Hate Speech."

ATTENTION ECONOMY

Facebook exists in a fundamentally different ecosystem than did its media ancestors. People today face an inundation of information to sift through, bombarded from all sides by largely irrelevant content. In such an environment, attention is in short supply. Economist and cognitive psychologist Herbert Simon first observed this phenomenon in 1971 when he wrote, "The wealth of information means a dearth of something else: a scarcity of whatever it is that information consumes. What information consumes is rather obvious: it consumes the attention of its recipients. Hence a wealth of information creates a poverty of attention."[12] This understanding of consumers' very attention as a scarce resource held profound economic implications, particularly at the turn of the century as the Internet made more information more accessible than ever before. In 1997, Goldhaber argued that "If the Web and the Net can be viewed as spaces in which we will increasingly live our lives, the economic laws we will live under have to be natural to this new space."[13] These new economic laws treat attention as capital, a valuable commodity that can be traded, bought, and sold.

Perhaps no company has capitalized on the commodification of attention better than Facebook. Myllylahti asserts that "user attention has become a currency on the platform," a resource that Facebook possesses in abundance and trades with other corporations for more traditional tender. This constitutes the cornerstone of the platform's business model. Facebook's services, after all, remain free to use; rather than paying cash to post, like, and share, users

12. Herbert A. Simon, et al., "Designing Organizations for an Information-Rich World," *Computers, Communication, and the Public Interest* (1971): 40.

13. Michael H. Goldhaber, "The Attention Economy and the Net," *First Monday* 2, no. 4 (April 1997): https://doi.org/10.5210/fm.v2i4.519.

"exchange their time [and] attention…for access to Facebook's services."[14] Major corporations from Coca-Cola to Walmart to Microsoft will pay for that time and attention—known in the industry as "eyeballs,"—forking over billions of dollars for the right to post their advertisements on Facebook pages. The more eyeballs a page receives, the more valuable it is to the company trying to hold users' attention—and the more lucrative it is for Facebook. This attention economy, notes technologist Tobias Rose-Stockwell, bears a single-minded focus on the metric of 'engagement,' defined as the means "by which companies evaluate the number of clicks, likes, shares, and comments associated with their content."[15] A Facebook page with a high degree of engagement by definition has a high number of people viewing it, and therefore a large audience for potential advertisers. Such a page represents a veritable jackpot for Facebook, as companies will spend more money on advertisements that more people will see. It is the virtual equivalent of renting a billboard on the freeway versus one on a country backroad. Facebook, therefore, designs its algorithms to boost content that will maximize user engagement—and corporate profits.

NEGATIVITY BIAS

The type of content most likely to go viral and maximize ad revenue is not, however, as innocuous as cat videos or reaction memes. Instead, posts containing controversial opinions, inflammatory rhetoric, or divisive language receive significantly more engagement, capturing user attention in a way that more positive posts cannot. Known as negativity bias, this proclivity is well-documented in psychological studies. Fiske and Ohira both demonstrate that stimuli

14. Frenkel and Kang, *An Ugly Truth,* 3.

15. Tobias Rose-Stockwell, "This is How Your Fear and Rage are Being Shared for Profit," *Quartz,* July 28, 2017, https://qz.com/1039910/how-facebooks-news-feed-algorithm-sells-our-fear-and-outrage-for-profit/.

associated with negative words or attributes prompt higher levels and duration of cognitive processes, including attention.[16] There is "a tendency for negative events to result in a greater mobilization of an organism's psychological, cognitive, emotional, and social responses."[17] Negativity bias possesses significant evolutionary utility; organisms that devote more mental resources to information regarding a potential threat are more able to fight or flee if that threat manifests. The zebra that pays attention to a suspicious movement in the grass will likely outlive the zebra that only pays attention to eating the grass. As Berggren notes, "threat stimuli are not just prioritized when competing for attention, but also able to 'hold' attention post-capture."[18] Thus, the bias enables negative stimuli to capture attention more readily as well as retain it.

Though humans have left the savannah and fears of predation in our evolutionary past, our brains remain hardwired to latch onto information and stimuli that evokes threat-induced emotional responses. A team of Finnish psychologists demonstrated how the negativity bias influences what media people prioritize in situations with multiple sources competing for their attention. Negative tweets, they found, "draw more gaze dwell time and are recognized better

16. S. T. Fiske, "Attention and Weight in Person Perception: The Impact of Negative and Extreme Behavior," *Journal of Personality and Social Psychology,* 38 no. 6, (1980): 889–906, https://doi.org/10.1037/0022-3514.38.6.889; Hideki Ohira, "Eyeblink Activity in a Word-Naming Task as a Function of Semantic Priming and Cognitive Load," Perceptual and Motor Skills 82, no. 3 (June 1996): 835–42, https://doi.org/10.2466/pms.1996.82.3.835.

17. Tiffany A. Ito et al., "Negative Information Weighs More Heavily on the Brain: The Negativity Bias in Evaluative Categorizations," *Journal of Personality and Social Psychology* 75, no. 4 (1998): 1, https://doi.org/10.1037/0022-3514.75.4.887.

18. Nick Berggren, "Rapid Attentional Biases to Threat-Associated Visual Features: The Roles of Anxiety and Visual Working Memory Access," *Emotion*, (June 2020): 2, http://dx.doi.org/10.1037/emo0000761.

than positive tweets."[19] This suggests that people will direct more thought and attention to negative information even in online environments that pose no real, physical threat to them.

NEWS FEED ALGORITHM

Of course, the Finnish study's findings apply beyond Twitter. The psychological gravitation towards fear-, disgust-, or rage-inducing content is not lost on Facebook, which recognizes that it can manipulate this adaptive biological mechanism to make money. It does this primarily through its News Feed, a brilliant and secretive algorithmic masterpiece that serves customized content to every user. As Christopher Mims of the Wall Street Journal explains, "Every time one of Facebook's two billion monthly users opens the Facebook app, a personalization algorithm sorts through all the posts that a person could theoretically see, and dishes up the fraction it thinks she or he would like to see first."[20] If a person likes a lot of posts about home gardening, for instance, their News Feed would likely contain more content about tomatoes, soil, or plants than about tacos, skateboarding, or pianos. The feature is meant to increase time spent on the platform by keeping its users happy and engaged.

Engineers at Facebook constantly make tweaks to the News Feed algorithm in order to further optimize it, but occasionally their changes cause massive ripple effects. In 2015, the company altered its entire metric of user activity, moving away from the assumption that more clicks indicated higher engagement. They switched instead to a model that stressed time spent on

---

19. Jari Kätsyri et al., "Negativity Bias in Media Multitasking: The Effects of Negative Social Media Messages on Attention to Television News Broadcasts," *PLoS ONE* 11, no. 5 (2016): 16, https://doi.org/10.1371/journal.pone.0153712.

20. Christopher Mims, "How Facebook's Master Algorithm Powers the Social Network," *The Wall Street Journal*, October 22, 2017, https://www.proquest.com/newspapers/how-facebooks-master-algorithm-powers-social/docview/1953638742/se-2?accountid=12205.

a post or page, believing it would mitigate the prevalence of clickbait articles.[21] Spam sites, however, would have been preferable to the actual consequences of the algorithmic shift. Primed by their neurological biases, people tended to spend more time on sensationalized and salacious content, prompting their News Feeds to display more of that ilk in a self-perpetuating cycle of negativity. Facebook users were indeed more engaged than ever, but they were also angrier and more fearful than ever.

After a particularly dismal fiscal year for the company, the News Feed underwent another overhaul in 2018. Engineers reconfigured the algorithm to emphasize what they called Meaningful Social Interaction (MSI), a formulaic of both intimacy and engagement, quantified and compiled into a tidy score. Intimacy measures closeness, or how connected one user is through another as indicated by mutual friends and interaction with one another on the platform. Engagement, on the other hand, tracks activity such as likes, comments, and shares.[22] A post that received, say, 3000 likes and 500 comments from a person with 150 mutual friends would receive a greater MSI score than a post from a stranger that generated only a handful of likes, and therefore show up in a more prioritized position on one's News Feed.

But what happens when these measures are at odds with each other? Which deserves algorithmic privilege: an uncommented post from a close friend or a viral post from a random user? Facebook's leaders decided that in the contest between intimacy and engagement, the latter wins out. The negativity bias reared its ugly head here as well, for the content that receives the most engagement is also likely to be the most inflammatory, triggering users' predilection to

---

21. Frenkel and Kang, *An Ugly Truth*, 185.

22. Jeff Horowitz, interview with Ryan Knutson and Keach Hagey, *The Facebook Files, Part 4: The Outrage Algorithm,* podcast audio, September 18, 2021, https://www.wsj.com/podcasts/the-journal/the-facebook-files-part-4-the-outrage-algorithm/e619fbb7-43b0-485b-877f-18a98ffa773f.

favor threatening stimuli with their attention. Sure enough, "people inside of Facebook started to notice that [the News Feed] was effectively highlighting the very worst kind of content[:] stuff that was divisive, really negative, and just represented the worst parts of humanity."[23] This effect was exacerbated by another aspect of the News Feed update. MSI does not merely retroactively suggest popular content, it actively anticipates what will be popular in the future and boosts it to get the ball rolling. Known as Downstream MSI, this program uses artificial intelligence (AI) to identify posts that possess trending potential and then bumps them up towards the top of people's News Feeds to attract a critical mass of attention, initiating a self-fulfilling prophecy of virality. Years of data have taught the AI that the content most likely to capture attention is negative, suggestive of some kind of threat or conflict. Media reporter Keach Hagey summarizes the repercussions of this algorithm:

> "[Downstream MSI is] a really different way of filtering what you see. It's not based on what you would actually most like to see or what's most relevant to you or what's highest quality. It's what will get the most comments. And the result of that, it turns out that what gets the most comments is really divisive, outrageous stuff, especially stuff that provokes political anger."[24]

In applying sophisticated technology to base biological impulses, Facebook has given renewed life to the old journalist adage "if it bleeds, it reads."

ETHICAL FAILURE, CORPORATE SUCCESS

The unspoken corollary to that adage, of course, is "if it reads, it makes us money." Facebook designs its algorithms to amplify popular or trending content, attracting ever-greater attention to a particular page or post and making it a hot commodity in the attention economy. Competing for this scarce resource, companies will pay Facebook lots of money to capitalize on

23. Ibid.

24. Ibid.

that captivated attention and run advertisements. Due to humans' psychological negativity bias, controversial, inflammatory, and/or derogatory content garners more attention; thus, such content is actually worth more money. As Sheera Frenkel and Cecilia Kang write in their exposé of Facebook, "the platform is built upon a fundamental possibly irreconcilable dichotomy: its purported mission to advance society by connecting people while also profiting off of them."[25] Facebook possesses a financial incentive to promote hate speech—or, at the very least, to not intervene in instances of hate speech on its platform. Taking down or suppressing hate speech, therefore, may run counter to Facebook's entire purpose as a corporation: to maximize profits.

The contradiction between monetary success and ethical integrity has not escaped the attention of Facebook's leadership. When the company implemented its Meaningful Social Interaction update, it convened a Civic Team to assess the new algorithm's impact on misinformation and political speech. According to internal documents obtained by the Wall Street Journal, the team discovered that MSI actively promoted the platform's most radical, toxic content. To further complicate matters, the worse the content, the more often it was shared. "So if a thing's been reshared 20 times in a row," explains the Journal's lead reporter, Jeff Horowitz, "it's going to be 10x or more likely to contain nudity, violence, hate speech, misinformation, than a thing that has just been not reshared at all."[26] Confronted with this damning revelation, founder and CEO Mark Zuckerberg agreed to scale back the MSI algorithm—but only in very particular places or pertaining to very particular content. The Civic Team compiled a long list of potential solutions ranging from diminishing platform speed, to capping daily group invitations, to

25. Frenkel and Kang, *An Ugly Truth,* 300.

26. Horowitz, *The Outrage Algorithm*.

removing the reshare button entirely. Ultimately, Facebook declined to adopt any of these measures; corporate leadership deemed them too detrimental to business.[27]

Samidh Chakrabarti, leader of the since-disbanded Civic Team, has argued that valuing all user engagement regardless of its substance will "Invariably amplify…sensationalism, hate and other societal harms…. [It] is so predictable that it is perhaps a natural law of social networks."[28] Yet despite this inevitability, Facebook has chosen again and again to protect its bottom line at the cost of human security. In defending this decision, COO Sheryl Sandberg and other spokespeople repeatedly appeal to the sanctity of freedom of expression. "More people being able to share their experiences and perspectives has always been necessary to build a more inclusive society," Zuckerberg maintained in a much-derided speech at Georgetown University.[29] This rhetoric buttresses Facebook's perennial claim that it serves merely as a platform, not a publisher. According to its own self-mythology, the company simply stores and provides information in an objective, unbiased manner without intermediating between content and users. Zuckerberg even attempted to make the bold and entirely implausible argument that Facebook is "not a media company."[30] In reality, of course, the platform constantly shapes, influences, and defines the experience of everyone engaging with it. As international human rights lawyer Jenny Domino notes, "Though platforms seem to only 'facilitate' expression, there

---

27. Ibid.

28. Ibid.

29. Molla, Rani, "Mark Zuckerberg Said a Lot of Nothing in His Big Speech," *Vox*, October 17, 2019, https://www.vox.com/2019/10/17/20919505/mark-zuckerberg-georgetown-free-speech-facebook.

30. Jeffrey Herbst, "The Algorithm is an Editor; Google, Facebook and Other Tech Companies Say They Aren't News Organizations, but the Claim is Becoming Increasingly Implausible," *Wall Street Journal (Online),* April 13, 2016, https://www.proquest.com/newspapers/algorithm-is-editor-google-facebook-other-tech/docview/1780643099/se-2?accountid=12205.

is nothing neutral about the curating, filtering, and 'orchestrating' of posted content that they take on."[31] Facebook constantly makes decisions about allowable content based on what will yield the greatest profit irrespective of that content's impact on its users.

COMMUNITY STANDARDS

It justifies these decisions with a convoluted web of capricious rules and arbitrary guidelines known as Community Standards, which everyone with a Facebook account must consent to. They function as the platform's human rights doctrine—albeit an ad hoc, reactive, inconsistently enforced doctrine. As discussed previously, Facebook defines hate speech as a "direct attack against people…on the basis of…protected characteristics: race, ethnicity, national origin, disability, religious affiliation, caste, sexual orientation, sex, gender identity, and serious disease."[32] Direct attacks themselves, according to the Community Standards, consist of "violent or dehumanizing speech, harmful stereotypes, statements of inferiority, expressions of contempt, disgust or dismissal, cursing and calls for exclusion or segregation."[33] Content that fits within this broad definition falls into one of three tiers of escalating severity. The third tier includes calls for segregation and/or exclusion on political, economic, or social grounds, as well as slurs, or "words that are inherently offensive and used as insulting labels."[34]

Tier two is broader still, encompassing pronouncements of a protected group's physical inferiority, such as generalizations regarding hygiene (i.e. dirty, filthy) or attractiveness (i.e.

---

31. Jenny Domino, "How Facebook is Reconfiguring Freedom of Speech in Situations of Mass Atrocity: Lessons from Myanmar and the Philippines," *Opinio Juris*, January 1, 2019, http://opiniojuris.org/2019/01/01/how-facebook-is-reconfiguring-freedom-of-speech-in-situations-of-mass-atrocity-lessons-from-myanmar-and-the-philippines/.

32. Meta, "Hate Speech."

33. Ibid.

34. Ibid.

ugly, hideous); mental inferiority, such as generalizations regarding intellect (i.e. stupid, idiots), education (i.e. illiterate, uneducated), or mental health (i.e. crazy, insane, retarded); moral inferiority, such as generalizations regarding negative character traits (i.e. coward, liar), or sexual habits (i.e. slut, pervert); group superiority (i.e. men are superior to women); group deviations from a norm (i.e. freak, abnormal). Within the same tier also falls "self-admission to intolerance" (i.e. homophobic, racist), expression of hatred (i.e. despise, hate), and assertion that a protected group should be dismissed or not exist at all. Suggestions that a group causes sickness, repulsion, or distaste (i.e. vile, disgusting, vomit) are also prohibited under this tier, as is cursing. Acknowledging the social relativity of curse words, Facebook defines it as "Referring to the target as genitalia or anus, including but not limited to: cunt, dick, asshole. Profane terms or phrases with the intent to insult, including but not limited to: fuck, bitch, motherfucker. Terms or phrases calling for engagement in sexual activity, or contact with genitalia, anus, feces or urine, including but not limited to: suck my dick, kiss my ass, eat shit."[35] Finally, the first tier forbids speech that is violent or dehumanizing, especially generalizations about or comparisons to insects, animals perceived as intellectually/physically inferior, filth, feces, bacteria, criminals, sexual predators, and denials of existence. The Community Standards make special note of "designated dehumanizing comparisons, generalizations, or behavioral statements," referring to stereotypes historically used to deride particular groups. These include blackface, comparisons of Black people to apes, Jewish people manipulating governmental or financial institutions, associations between Muslims and pigs, and women equated with property or objects.[36]

35. Ibid.

36. Ibid.

Content found to violate any of these three tiers faces removal by algorithmic or human moderators. However, both man and machine have proven fallible, routinely taking down acceptable material or missing instances of hate speech altogether. Yet the fundamental issue with Facebook's Community Standards lies not with their enforcement, but the very philosophy that undergirds them. "Facebook adopts a post-racial, race-blind approach that does not consider history and material differences, while its main focus is on enforcement, data, and efficiency."[37] In doing so, argue Siapera and Viejo-Otera, the platform encourages users to develop their own race-blind behaviors, socializing them into a system of regulation that essentializes racism as "what Facebook defines as racist hate speech."[38] Decontextualizing hate speech out of any historical framework (re)produces white supremacy and commits substantial harm against racialized minorities. Facebook's commitment to free speech and "fundamental equality" creates a calculus of absolute equivalence, treating a page or post that abuses others identically to the page or post that defends others. Thus, a user who spouts racist rhetoric that does not impinge upon the narrow Community Standards receives the exact same privileges and liberties as a user who advocates for racial justice.

This position has faced pushback from within the company itself. According to leaked audio files from an internal Facebook Q&A session, one person asked, "According to your policies "men are trash" is considered tier-one hate speech. So what that means is that our classifiers are able to automatically delete most of the posts or comments that have this phrase in it. [Why?]" In response, Zuckerberg contended "We don't think that [Facebook] should be in the

37. Eugenia Siapera and Paloma Viejo-Otero, "Governing Hate: Facebook and Digital Racism," Television & New Media 22, no. 2 (February 2021): 112—30, https://doi.org/10.1177/1527476420982232, 112.

38. Ibid, 117.

business of assessing which group has been disadvantaged or oppressed."[39] It follows, therefore, that all races deserve equal protection on the platform…and it is one short step from 'all races matter' to 'All Lives Matter.' "In these terms, Facebook applies a post-racial understanding of race and racism, which is essentially a denial of the existing reality of racism."[40] By attempting to mitigate hate speech as a procedural issue, Facebook strips protected categories such as race from all historical or sociopolitical context and in so doing erases past racisms, permits current racisms, and encourages the proliferation of future ones.

The race-blind approach to content moderation not only reifies structures of white supremacy, but actively inhibits racial justice. On her blog, anti-racist educator DiDi Delgado relates her rocky relationship with Facebook and her repeated bans from the platform, an experience shared by countless other Black activists. In an incisive article entitled "Mark Zuckerberg Hates Black People," Delgado explains that she has "lost count of how many Black organizers have had their Facebook accounts temporarily or permanently banned for posting content that even *remotely* challenges white supremacy."[41] Many activists have received bans simply for reposting screenshots of hate speech directed towards them. Often, the initial posts themselves remained on the platform. Facebook banned another organizer for using the phrase "Dear White People" in a post, and, in an instance of almost laughable irony, banned yet another for their criticism of Facebook's racist banning practices.[42] All of these situations reflect the fact

39. Casey Newton, "Why You Can't Say 'Men are Trash' on Facebook," *The Verge*, October 3, 2019, https://www.theverge.com/interface/2019/10/3/20895119/facebook-men-are-trash-hate-speech-zuckerberg-leaked-audio.

40. Siapera and Viejo-Otera, "Governing Hate," 125.

41. DiDi Delgado, Mark Zuckerberg Hates Black People," *Medium*, May 17, 2017, https://thedididelgado.medium.com/mark-zuckerberg-hates-black-people-ae65426e3d2a.

42. Ibid.

that Facebook does not consider past or present oppression when designing its policies. In the eyes of both machine and human content moderators, white people and Black people possess identical histories, experiences, and privileges.

Sometimes, in fact, white people warrant additional privileges when it comes to hate speech. As mentioned previously, Facebook's classified censorship guidelines differentiate between protected and unprotected subsets of the protected categories delineated in its definition of hate speech. One post from a member of the US House of Representatives called for his followers to identify, hunt, and kill radicalized Muslims. "Kill them all. For the sake of all that is good and righteous. Kill them all."[43] Despite this explicit call for violence, Facebook permitted his post to remain on the platform because it specifically targeted *radicalized* Muslims: not a protected subset. A ProPublica investigation into the company's content moderation protocols unearthed a particular training document used to familiarize reviewers with hate speech procedures, including this principle of subcategorization. The slide asked viewers to determine which of the following groups should receive protection from hate speech: female drivers, black children, white men. Correct answer? White men.[44] Both race and gender fall under the designation of protected categories, but neither drivers nor age do. Thus, female drivers and black children—though victims of historical persecution and even violence—do not merit protection from hate speech.

However, even this seemingly arbitrary demarcation of social groups as protected or unprotected is subject to the whims of Facebook executives; the Community Standards that all

43. Peter Holley, "'Kill Them. Kill Them All': GOP Congressman Calls for War against Radical Islamists," *The Washington Post,* June 5, 2017, https://www.washingtonpost.com/news/acts-of-faith/wp/2017/06/05/kill-them-kill-them-all-gop-congressman-calls-for-holy-war-against-radical-islam/.

44. Angwin and Grassegger, "Facebook's Secret Censorship Rules."

users must consent to do not, it seems, apply to everyone equally. Political leaders, influential public figures, and other social elites can operate with broader latitude as their speech is considered important political discourse—even when it employs offensive or exclusionary rhetoric. Posts from American president Donald Trump regarding his so-called Muslim ban remained visible on the platform even when they clearly committed a third-tier violation of "explicit exclusion, which means things like expelling certain groups or saying they are not allowed."[45] This particular case illustrates a much broader trend. Leaked documents from Facebook "suggest that, at least in some instances, the company's hate-speech rules tend to favor elites and governments over grassroots activists and racial minorities."[46] Prominent politicians therefore possess tacit permission to disseminate prejudice, bigotry, and hatred on the platform. Zuckerberg and Facebook spokespeople defend this practice by appealing to free speech and the company's mission to "bring the world closer together."[47] Removal of harmful content is discouraged in the name of connection and the global marketplace of ideas. As comedian, actor, and activist Sacha Baron Cohen put it, "This is not about limiting anyone's free speech. This is about giving people, including some of the most reprehensible people on earth, the biggest platform in history to reach a third of the planet."[48] By adopting a race-blind approach that indulges contemporary oppression while ignoring historic oppression, Facebook "merely repeats

45. Meta, "Hate Speech."

46. Angwin and Grassegger, "Facebook's Secret Censorship Rules."

47. Meta, "Our Mission."

48. Sacha Baron Cohen, "Keynote Address," Transcript of Speech delivered at Anti-Defamation League Never is Now Summit, November 21, 2019, https://www.adl.org/news/article/sacha-baron-cohens-keynote-address-at-adls-2019-never-is-now-summit-on-anti-semitism.

the same unfair treatment to which racialized people have been subjected."[49] In spite of its exhaustive Community Standards, hate speech continues to flourish on the platform.

VIRTUAL HATE, REAL VIOLENCE

Upon first glance, this issue appears to possess limited implications; hate speech, while certainly undesirable, is only words, after all, and when posted on Facebook those words are not even verbalized. When measured against the countless physical atrocities committed each day, what harm can cyberhatred really cause? According to an ever-expanding body of literature, a great deal. As even Facebook acknowledges, online instances of hate speech "are often linked with offline violence."[50] This phenomenon owes itself to the power of speech acts, as well as to a variety of qualities peculiar to social media.

The generative capabilities of speech possess a rich and multidisciplinary scholarly ancestry, drawing upon the work of J.L Austin, Jacques Derrida, and Judith Butler. Initially introduced by Austin, speech act theory articulates the ability of "performative utterances" to not merely describe existing realities, but to actively (re)produce them. He refers to this ability—the power of words to do something rather than simply to say something—as "illocutionary force."[51] While Austin sees the performativity of speech as dependent upon its utterance by the correct subject in the correct circumstance to the correct audience, Derrida generalizes the theory, connecting it instead to the iterability of speech. "Key for Derrida…is the iterability, or repeatability, of the [speech]; it is this reiterative structure, the fact that the [speech] is the same yet also differs and defers…that marks its force (and its power of signification) …. Because the

---

49. Siapera and Viejo-Otera, "Governing Hate," 127.

50. Meta, "Hate Speech."

51. J. L. Austin, *How to Do Things with Words,* Cambridge: Harvard University Press, 1975.

[speech] must be repeated in order to signify, it is always both tied to and divorced from its context of utterance. This separation…provides the performative's force."[52] Both iterability and decontextualization—central to Derrida's conceptualization of speech acts—is magnified on social media platforms. Facebook allows a single post to circulate the globe in mere moments, reaching millions of viewers who may lack the historical, social, or cultural background needed to understand it or verify its veracity. In this way, speech acts on Facebook hold even greater power to generate meaning and shape realities.

Judith Butler examines the ways in which individuals and collectivities can use that power to oppress, harm, and violate others. In verbalizing racial slurs, the speaker "is thus citing that slur, making linguistic community with a history of speakers" and therefore "accumulates the force of authority through the repetition or citation of a…set of practices."[53] Invoking the historicity of a particular derogation reenacts social trauma and reifies subjugation. Observing this same phenomenon, Nobel laureate Toni Morrison notes that "Oppressive languages does more than represent violence; it is violence."[54] Thus, hateful speech acts construct a subject, imbue it with meaning, and then harm it by indexing a historical lineage of trauma. Facebook allows language to overcome its usual spatial and temporal limits, circulating and iterating hate speech—and the violence it creates—to an unprecedented degree.

This violence is compounded by social media's ability to stimulate certain psychological proclivities in ways that hijack critical reasoning capacities. Information-rich environments like

---

52. Amy Hollywood, "Performativity, Citationality, Ritualization," *History of Religions* 42, no. 2 (2002): 104, http://www.jstor.org/stable/3176407.

53. Judith Butler, *Excitable Speech: A Politics of the Performative*, London: Routledge, 1997, 51-52.

54. Toni Morrison, "Nobel Lecture," Transcript of Speech delivered at the Nobel Prize in Literature, December 7, 1993, https://www.nobelprize.org/prizes/literature/1993/morrison/lecture/.

the modern media ecosystem place attention at a premium, driving individuals to resort to cognitive shortcuts such as heuristics when making decisions. Brown and Beruvides define heuristics as "methods that use principles of effort-reduction and simplification that allow decision makers to process information in a less effortful manner."[55] Facebook users tend to depend upon a series of these heuristics when assessing the reliability of the posts they view. In particular, "accuracy is now equated with popularity."[56] Because the platform's algorithms amplify sensationalized or negative content, heuristics-based reasoning indicates that hate speech constitutes a legitimate, trustworthy source of information.

Facebook also tends to restrict the array of information sources that users access. In order to increase engagement on the platform, Facebook's personalization algorithms prioritize displaying content similar to previously viewed content. If a user has displayed a preference for cat videos, their feed will show them more cat videos, which makes the user view even more of them and prompts the algorithm to double down on furry feline footage. Of course, this principle applies to more nefarious content as well. Social media algorithms that create positive feedback loops can facilitate radicalization, exposing users to incrementally more extreme viewpoints with every post they share, video they view, or comment they like. This process simultaneously restricts the range of content an individual can readily see. Müller and Schwarz observe that "preferential selection may limit the spectrum of information people absorb and create 'echo chambers,' which reinforce similar ideas."[57] To return to our hypothetical cat-obsessed user,

---

55. Peter J. Brown and Mario G. Beruvides, "The Heuristic-Based Framework for Attitude Certainty: How Technology and the Attention Economy Are Systematically Eroding Systematic Thinking," *The Psychologist-Manager Journal* 23, no. 2 (May 2020): 80-81. https://doi:10.1037/mgr0000107.

56. Ibid, 82.

57. Karsten Müller and Carlo Schwarz, "Fanning the Flames of Hate: Social Media and Hate Crime," *Social Science Research Network* (June 2020): 2 https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3082972.

because they have demonstrated a clear preference for cat content, their online milieu will consist of others who share their viewpoints. Some of those other users may not only love cats, but think that cats are superior to dogs. Due to Facebook's algorithm, more and more of this pro-cat content will appear on our user's feed, and the more of it they engage with, the more their algorithm will yield similar yet marginally more extreme content. Before long, this user may not only be pro-cat, but also rabidly anti-dog. They may begin to express sentiments of hatred towards dogs, towards dog owners, or even towards any non-feline pet. Enabling this process of radicalization, their personalized newsfeed will continue displaying content that aligns with the user's pre-existing opinions, further reinforcing them. This phenomenon is known as the 'filter bubble.' Filter bubbles serve to "[isolate] people from a diversity of viewpoints and content," allowing them to exist within a subjective reality that fits comfortably within their personal framework of biases and assumptions about the world.[58]

The real danger arises when users then attempt to warp the real world to mirror that subjective reality. Facebook hate speech, as it turns out, does not remain online; rather, it engenders offline hatred, influences real actions, and causes physical violence. Research suggests that "social media can act as a propagation mechanism for violent crimes" not only by providing a platform for collective mobilization, but functioning as a channel for persuasion.[59] In one study of the correlation between anti-refugee hate speech on Facebook and hate crimes targeting refugees in Germany, vitriolic sentiment on the social media platform "persuades potential perpetrators that refugees may be dangerous or undeserving, which may then push some people

58. Brown and Beruvides, "The Heuristic-Based Framework for Attitude Certainty," 89.

59. Müller and Schwarz, "Fanning the Flames of Hate," 1.

over the edge."[60] Individuals who may not commit hate crimes in other circumstances find themselves compelled to do so by the extremist rhetoric they are exposed to on Facebook. This rhetoric receives amplification from the platform's algorithms, designed to prey upon humans' innate predisposition towards negative information in order to maximize user engagement. Hate speech and its ilk are thus rendered lucrative real estate for digital advertisers, who line Facebook's pockets with billions of dollars in order to pitch their product to individuals swamped by so much extraneous information that they can no longer devote their full cognitive capabilities to processing it. The social media corporations, in short, profit off of hatred. Facebook's very business model facilitates the further marginalization of minority groups, fosters inciteful rhetoric, and even enables genocide.

60. Ibid, 37-38.

CHAPTER ONE BIBLIOGRAPHY

Angwin, Julia and Hannes Grassegger. "Facebook's Secret Censorship Rules Protect White Men from Hate Speech But Not Black Children." *ProPublica,* June 28, 2017. https://www.propublica.org/article/facebook-hate-speech-censorship-internal-documents-algorithms.

Anti-Defamation League. "Facebook's Hate Speech Problem is Even Bigger Than We Thought." Accessed September 18, 2021. https://www.adl.org/blog/facebooks-hate-speech-problem-is-even-bigger-than-we-thought.

Austin, J. L. *How to Do Things with Words.* Cambridge: Harvard University Press, 1975.

Baron Cohen, Sacha. "Keynote Address." Transcript of Speech delivered at Anti-Defamation League Never is Now Summit, November 21, 2019. https://www.adl.org/news/article/sacha-baron-cohens-keynote-address-at-adls-2019-never-is-now-summit-on-anti-semitism.

Berggren, Nick. "Rapid Attentional Biases to Threat-Associated Visual Features: The Roles of Anxiety and Visual Working Memory Access." *Emotion*, (June 2020): http://dx.doi.org/10.1037/emo0000761.

Brown, Peter J., and Mario G. Beruvides. "The Heuristic-Based Framework for Attitude Certainty: How Technology and the Attention Economy Are Systematically Eroding Systematic Thinking." *The Psychologist-Manager Journal* 23, no. 2 (May 2020): 76–94. https://doi:10.1037/mgr0000107.

Butler, Judith. *Excitable Speech: A Politics of the Performative*. London: Routledge, 1997.

Delgado, DiDi. "Mark Zuckerberg Hates Black People." *Medium*, May 17, 2017. https://thedididelgado.medium.com/mark-zuckerberg-hates-black-people-ae65426e3d2a.

Domino, Jenny. "How Facebook is Reconfiguring Freedom of Speech in Situations of Mass Atrocity: Lessons from Myanmar and the Philippines." *Opinio Juris*, January 1, 2019. http://opiniojuris.org/2019/01/01/how-facebook-is-reconfiguring-freedom-of-speech-in-situations-of-mass-atrocity-lessons-from-myanmar-and-the-philippines/.

Federal Bureau of Investigation. "Hate Crimes." Accessed September 18, 2021. https://www.fbi.gov/investigate/civil-rights/hate-crimes.

Fiske, S. T. "Attention and Weight in Person Perception: The Impact of Negative and Extreme Behavior." *Journal of Personality and Social Psychology,* 38 no. 6, (1980): 889–906. https://doi.org/10.1037/0022-3514.38.6.889.

Frenkel, Sheera and Cecilia Kang. *An Ugly Truth.* New York: Harper Collins, 2021.

Goldhaber, Michael H. "The Attention Economy and the Net." *First Monday* 2, no. 4 (April 1997): https://doi.org/10.5210/fm.v2i4.519.

Herbst, Jeffrey. "The Algorithm is an Editor; Google, Facebook and Other Tech Companies Say They Aren't News Organizations, but the Claim is Becoming Increasingly Implausible." *Wall Street Journal (Online),* April 13, 2016. https://www.proquest.com/newspapers/algorithm-is-editor-google-facebook-other-tech/docview/1780643099/se-2?accountid=12205.

Hoffman, Claire. "The Battle for Facebook." *Rolling Stone*, September 15, 2010. https://www.rollingstone.com/culture/culture-news/the-battle-for-facebook-242989/l.

Holley, Peter. "'Kill Them. Kill Them All': GOP Congressman Calls for War against Radical Islamists." *The Washington Post,* June 5, 2017. https://www.washingtonpost.com/news/acts-of-faith/wp/2017/06/05/kill-them-kill-them-all-gop-congressman-calls-for-holy-war-against-radical-islam/.

Hollywood, Amy. "Performativity, Citationality, Ritualization." *History of Religions* 42, no. 2 (2002): 93–115. http://www.jstor.org/stable/3176407.

Horowitz, Jeff. Interview with Ryan Knutson and Keach Hagey. *The Facebook Files, Part 4: The Outrage Algorithm.* Podcast audio. September 18, 2021. https://www.wsj.com/podcasts/the-journal/the-facebook-files-part-4-the-outrage-algorithm/e619fbb7-43b0-485b-877f-18a98ffa773f.

Ito, Tiffany A., Jeff T. Larsen, N. Kyle Smith, and John T. Cacioppo. "Negative Information Weighs More Heavily on the Brain: The Negativity Bias in Evaluative Categorizations." *Journal of Personality and Social Psychology* 75, no. 4 (1998): 887–900. https://doi.org/10.1037/0022-3514.75.4.887.

Kätsyri, Jari, Teemu Kinnunen, Kenta Kusumoto, Pirkko Oittinen, and Niklas Ravaja. "Negativity Bias in Media Multitasking: The Effects of Negative Social Media Messages on Attention to Television News Broadcasts." *PLoS ONE* 11, no. 5 (2016): https://doi.org/10.1371/journal.pone.0153712.

Mims, Christopher. "How Facebook's Master Algorithm Powers the Social Network; The Algorithm Behind Facebook's News Feed, a 'Modular Layered Cake, Extracts Meaning from Every Post and Photo.' *The Wall Street Journal*, October 22, 2017. https://www.proquest.com/newspapers/how-facebooks-master-algorithm-powers-social/docview/1953638742/se-2?accountid=12205.

Meta. "Hate Speech." Accessed May 22, 2021. https://www.facebook.com/communitystandards/hate_speech.

Meta. "Our Mission." Accessed March 15, 2022, https://about.facebook.com/company-info/.

Molla, Rani. "Mark Zuckerberg Said a Lot of Nothing in His Big Speech." *Vox*, October 17,

2019. https://www.vox.com/2019/10/17/20919505/mark-zuckerberg-georgetown-free-speech-facebook.

Morrison, Toni. "Nobel Lecture." Transcript of Speech delivered at the Nobel Prize in Literature, December 7, 1993. https://www.nobelprize.org/prizes/literature/1993/morrison/lecture/.

Müller, Karsten and Carlo Schwarz. "Fanning the Flames of Hate: Social Media and Hate Crime." *Social Science Research Network* (June 2020). https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3082972.

Newton, Casey. "Why You Can't Say 'Men are Trash' on Facebook." *The Verge*, October 3, 2019. https://www.theverge.com/interface/2019/10/3/20895119/facebook-men-are-trash-hate-speech-zuckerberg-leaked-audio.

Ohira, Hideki. "Eyeblink Activity in a Word-Naming Task as a Function of Semantic Priming and Cognitive Load." Perceptual and Motor Skills 82, no. 3 (June 1996): 835–42. https://doi.org/10.2466/pms.1996.82.3.835.

Pepitone, Julianne. "Facebook Trading Sets Record IPO Volume." *CNN Money,* May 18, 2012. https://money.cnn.com/2012/05/18/technology/facebook-ipo-trading/index.htm.

Rose-Stockwell, Tobias. "This is How Your Fear and Rage are Being Shared for Profit." *Quartz,* July 28, 2017. https://qz.com/1039910/how-facebooks-news-feed-algorithm-sells-our-fear-and-outrage-for-profit/.

Siapera, Eugenia and Paloma Viejo-Otero. "Governing Hate: Facebook and Digital Racism." *Television & New Media* 22, no. 2 (February 2021): 112—30. https://doi.org/10.1177/1527476420982232.

Simon, Herbert A., Karl W. Deutsch, Martin Shubik, and Emilio Q. Daddario. "Designing

Organizations for an Information-Rich World." *Computers, Communication, and the Public Interest* (1971): 37-72.

United Nations General Assembly. *Promotion and Protection of the Right to Freedom of Opinion and Expression,* A/74/486, October 9, 2019. https://www.ohchr.org/Documents/Issues/Opinion/A_74_486.pdf.

CHAPTER TWO:
CONSEQUENCES OF RACIALIZED HATE SPEECH ON FACEBOOK

Hatred, of course, has existed both before and beyond Facebook. Social media does not cause hate, but it can provide a platform for people to connect over their prejudice and fear, allowing it to metastasize with unprecedented intensity and rapidity. Nowhere does this phenomenon manifest more clearly than Myanmar, where the Rohingya ethnic minority have faced what the UN called a "textbook example of ethnic cleansing."[1] Though perpetrated offline, the genocide was organized and encouraged on Facebook. Myanmar may represent the most extreme example of social media exacerbating hate, but it is far from the only one. Facebook possesses almost 3 billion monthly users across the world and its proclivity to aggravate social tensions—particularly racial and ethnic ones—has raised concerns everywhere from India, to France, to Ethiopia, to the United States. Clearly, it is far too late to put this genie back in the bottle; Facebook is going nowhere. We must therefore ask some critical questions: what qualities inherent to social media allow for such virulent hate speech? What other environmental factors encourage this proliferation? How does online hate speech impact real-world conditions? Understanding the nature of the hatred behind the Rohingya genocide can illuminate how Facebook's technological innovations amplify fundamental human tendencies, facilitating hate speech that, if left unchecked, may burgeon into real-world violence.

1. "UN Human Rights Chief Points to 'Textbook Example of Ethnic Cleansing' in Myanmar." *UN News,* September 11, 2017, https://news.un.org/en/story/2017/09/564622-un-human-rights-chief-points-textbook-example-ethnic-cleansing-myanmar.

A BRIEF HISTORY OF MYANMAR

The region now known as Myanmar[2] possesses a rich history of Theravada Buddhism, adopted from South India in the fourth century. Religious isolation from its geographic neighbors led to expansionist policies that spanned multiple dynasties, and by the late 1500s the Burmese Empire stretched across modern-day Laos and Thailand. Its later forays into China and Assam caught the attention of Great Britain. A series of Anglo-Burmese wars played out through much of the 19th century, culminating in British annexation in 1885. Eager to use the territory as a stepping stone into China, Britain appended Myanmar as an Indian province. Substantial resistance to colonization persisted, particularly in the northern region, and British administrators resorted to burning and looting entire villages in order to suppress potential rebellion. They also employed classic divide-and-conquer tactics; certain groups received favor and privileged positions within the colonial regime, fomenting inter-ethnic conflict that prevented unified opposition against the British. Renewed calls for independence gained momentum in the 1920s, particularly on university campuses. One particularly charismatic law student, Aung San, worked with the imperial Japanese to overthrow British rule. When Japan invaded Burma in 1942, however, he switched allegiances, siding with Great Britain to drive out the Japanese instead. Aung San eventually reached an agreement with Britain, and Burma emerged as a newly independent state on January 4, 1948.

However, a 1962 coup deposed the democratically elected government, suspended the constitution, and established a dictatorship under the Tatmadaw, the official name of the Burmese military. The following decades witnessed intense economic hardship and suppression

2. The state's official name changed in 1989 from the Union of Burma to the Union of Myanmar. The United States government still refers to the country as Burma; however, this paper will follow the conventions of the rest of the world and use Myanmar when referring to the country after 1989.

of press freedom, prompting nation-wide demonstrations in the summer of 1988 that the military violently dispersed. Over 3000 people were massacred, and thousands more fled the country. These protests nevertheless planted seeds of resistance. Under the leadership of Aung San's daughter, Aung San Suu Kyi, various ethnic nationalist groups joined together in a new democratic movement.[3] Facing immense pressure, the reigning dictator stepped down, and a new military junta took power over the newly renamed Union of Myanmar.[4] Elections were held in 1990, yet despite an overwhelming victory for Aung San Suu Kyi's party, the Tatmadaw refused to surrender power and Suu Kyi herself remained under house arrest. For her leadership and dedication to non-violence in the struggle for democracy, she received the 1991 Nobel Peace Prize.[5]

Increased international scrutiny and a desire to attract foreign investment led to the adoption of a new constitution in 2008, and the junta itself dissolved entirely in 2011. After a transitional period, Aung San Suu Kyi ascended as the de facto leader of Myanmar, though the 2008 constitution prevents her from officially assuming the presidency. Free and fair elections in 2015 reaffirmed Burmese support for Suu Kyi, with her National League for Democracy party winning in a landslide. However, the Tatmadaw continues to exercise enormous control over the civilian government. One quarter of parliamentary seats are reserved for the military, giving it veto power over any constitutional amendment. Myanmar's 2020 elections firmly repudiated the military's proxy political party, and after its allegations of election fraud were disproven, the

---

3. Dinyar Godrej, "A Short History of Burma," *New Internationalist,* April 18, 2008, https://newint.org/features/2008/04/18/history.

4. Lindsay Maizland, "Myanmar's Troubled History: Coups, Military Rule, and Ethnic Conflict," *Council on Foreign Relations,* February 10, 2021, https://www.cfr.org/backgrounder/myanmar-history-coup-military-rule-ethnic-conflict-rohingya.

5. Godrej, "A Short History of Burma."

Tatmadaw staged another coup in February 2021. Aung San Suu Kyi and other elected officials were detained, and Myanmar remains under martial law.[6]

ETHNIC TENSIONS IN MYANMAR

Against this backdrop of colonization and military rule, ethnic conflict has permeated Myanmar's history. Immensely diverse, the state recognizes over 100 ethnic groups, with ethnic Burmans constituting almost 70% of the population. Bamars, as they are known, have long enjoyed privileged social and economic status as other minorities endure persistent discrimination and disenfranchisement.[7] Members of the Rohingya ethnicity in particular face extreme oppression; the UN Human Rights Council has referred to them as "the most persecuted minority in the world."[8] Inhabitants of the Arakan region—now the Rakhine state of northern Myanmar—since the eighth century, they possess their own distinct language and practice Sunni Islam. This sets them apart from the Hindu Bamars, who occupied Arakan in 1785. When Burma came under British rule, colonial administrators sought to combat Bamar nationalist sentiments by favoring Rohingya Muslims with bureaucratic positions. Favoritism policies like this increased ethnic tensions during the Japanese invasion of World War II; while the Bamar sided with Japan, Rohingya largely supported the British and even received military resources to help combat Japanese imperialism. In light of this, upon gaining independence the Burmese government refused citizenship to any Rohingya individual.[9]

---

6. Maizland, "Myanmar's Troubled History."

7. Ibid.

8. Office of the High Commissioner of Human Rights, "Human Rights Council Opens Special Session on the Situation of Human Rights of the Rohingya and Other Minorities in Rakhine State in Myanmar," December 5, 2017, https://www.ohchr.org/EN/NewsEvents/Pages/DisplayNews.aspx?NewsID=22491.

9. Rohingya Cultural Center, "History of the Rohingya," accessed October 16, 2021, https://rccchicago.org/history/.

Their statelessness was further codified in the 1982 Citizenship Law passed by the Tatmadaw, which identified 135 official Burmese ethnic groups but notably excluded the Rohingya. According to the junta's reasoning, any group not settled in Burma prior to colonization immigrated under British imperial rule, and therefore illegally occupy land that belongs to native Burmese.[10] This rhetoric indicates a broader appeal to Buddhist nationalist sentiment. According to the Rohingya Cultural Center, "In the same vein as the early nationalist movement under British occupation, [the Tatmadaw] fostered the belief that Burma is a land purely for the Burmese Buddhists, and used the 'us' and 'them' discriminatory rhetoric to unite the population under its military rule."[11] Attempting to 'purify' Burma, the Tatmadaw launched Operation Pyi Thaya, or "Clean and Beautiful Nation," in 1991. Amidst a widespread campaign of military violence, nearly one-quarter of a million Rohingya fled to Bangladesh, prompting the establishment of a special border security unit that further harassed people either fleeing or returning to the country.[12]

Long-simmering tensions exploded in 2012 when ethnic riots erupted between Rohingya Muslims and Rakhine Buddhists. Government-armed citizens murdered dozens of Rohingya and burned their villages, displacing over 140,000 people. In response to the conflict, the Tatmadaw imposed strict curfews and deployed military troops to Rakhine State, which only increased distrust and hostility. Continued persecution drove many Rohingya to attempt perilous sea journeys to Bangladesh, Thailand, and Malaysia. Countless refugees drowned or starved, and those who arrived at their destinations often faced imprisonment or became victims of human

---

10. Ibid.

11. Ibid.

12. United States Holocaust Memorial Museum, "Burma's Path to Genocide," accessed October 18, 2021, https://exhibitions.ushmm.org/burmas-path-to-genocide/timeline.

trafficking.[13] Even after the most acute violence abated, systematic cultural and political exclusion persisted. The early 2010s saw a proliferation of Buddhist nationalist movements, as well as organized attempts to further disenfranchise the Rohingya. A 2014 national census excluded them entirely, and in 2015 the government began issuing special verification cards that identified them as illegal immigrants from Bangladesh.[14]

The situation worsened in 2016. A small Rohingya militia assaulted a handful of police stations in Rakhine State, resulting in nine casualties. In response, the Tatmadaw commenced what they called a 'clearance operation:' systematic murder, rape, and looting. Over 86,000 Rohingya fled the country in what proved a harbinger of violence to come.[15] During August 2017, another Rohingya rebel group attacked several military outposts in Rakhine State and the military retaliated with brutal force, terrorizing villages, killing over 10,000 people, and displacing 740,000 more. The troops targeted men, women, and children indiscriminately, often trapping them in their homes and setting the buildings alight, burning them inside before disposing of their remains in mass graves. Those who fled faced rape, injury, and death.[16] Myanmar's 'clearance operations' have culminated in the worst refugee crisis since the Rwandan genocide, the creation of one of the world's largest stateless populations, and, in the words of the UN human rights chief, a "textbook example of ethnic cleansing."[17]

---

13. Rohingya Cultural Center, "History of the Rohingya."

14. United States Holocaust Museum, "Burma's Path to Genocide."

15. Ibid.

16. United Nations Human Rights Council, *Report of the Independent International Fact-finding Mission on Myanmar*, A/HRC/39/64, September 12, 2018, https://www.ohchr.org/Documents/HRBodies/HRCouncil/FFM-Myanmar/A_HRC_39_64.pdf.

17. "The Rohingya Refugee Crisis is the Worst in Decades," *The Economist,* September 21, 2017, https://www.economist.com/graphic-detail/2017/09/21/the-rohingya-refugee-crisis-is-the-worst-in-decades.

DESIGNING A GENOCIDE

Such violent acts did not occur spontaneously. Rather, a United Nations report found, "they were foreseeable and planned."[18] Ethnic tension percolated for years in a sociopolitical brew of legal discrimination and cultural persecution, but the Rohingya genocide was enabled, perpetuated, and exacerbated on Facebook. Masquerading as fan pages of pop stars, national heroes, and other celebrities, Myanmar military operatives launched a social media campaign to dehumanize and denigrate the Rohingya minority. "The campaign…included hundreds of military personnel who created troll accounts and news and celebrity pages on Facebook and then flooded them with incendiary comments and posts timed for peak viewership."[19] Reuters completed an extensive investigation into hate speech posted by the Myanmar military. One such post reads in Burmese, "Kill all the kalars [pejorative for Rohingya] that you see in Myanmar; none of them should be left alive." Another post from October 2016 proclaims that "If it's kalar, get rid of the whole race," and an entire Facebook page devotes itself to the declaration, "We will genocide all of the Muslims and feed them to the dogs."[20] This sort of rhetoric actually predates the ethnic cleansing campaign; in December 2013 one Facebook user implored their viewers, "We must fight [the Rohingya] the way Hitler did the Jews, damn kalars!" Many posts "call the Rohingya or other Muslims dogs, maggots and rapists, suggest they be fed to pigs, and urge they be shot or exterminated. The material also includes crudely pornographic anti-Muslim

18. United Nations Human Rights Council, "Report."

19. Paul Mozur, "A Genocide Incited on Facebook, With Posts From Myanmar's Military," *The New York Times*, October 15, 2018, https://www.nytimes.com/2018/10/15/technology/myanmar-facebook-genocide.html.

20. Steve Stecklow, "Why Facebook is Losing the War on Hate Speech in Myanmar," *Reuters,* August 15, 2018. https://www.reuters.com/investigates/special-report/myanmar-facebook-hate/.

images."[21]

A propaganda campaign of this scope and complexity did not emerge overnight; the Tatmadaw labored over its Facebook operations for years and devoted hundreds of personnel to the effort. In 2017, sham accounts administered by the military began spreading rumors to their respective Rohingya and Buddhist audiences about imminent attacks from the other side, sowing fear and distrust while increasing popular support for the Tatmadaw itself.[22] The posts were sensational, but seldom unbelievable. As one Burmese official trained in psychological warfare put it, "If one quarter of the content is true, that helps make the rest of it believable." Thus, the military manipulated existing prejudices and anxieties in order to wreak violence with unprecedented brutality and systematicity.

Anthropologist Alexander Hinton uses the Khmer Rouge in Cambodia to illustrate "the importance of viewing genocide as a process…that is generated by a variety of factors and has diverse outcomes."[23] According to this framework, societies go through a period of 'priming'—it may be enduring or acute, low-grade or high-intensity—that makes genocide more or less likely. In considering the Rohingya genocide, it is useful to examine the environmental primes taking place in the physical world, as well as the psychological primes playing out on Facebook. Of course, these macro- and micro-level factors did not operate separately; rather, they occurred in tandem, (re)producing one another in a symbiosis inconceivable before the advent of social media. Analyzing the interplay between them is crucial to understanding when, where, and how hate speech on platforms like Facebook can influence or exacerbate violence on the ground.

---

21. Ibid.

22. Mozur, "A Genocide Incited on Facebook."

23. Alexander Laban Hinton, *Why Did They Kill? Cambodia in the Shadow of Genocide,* (Berkeley and Los Angeles: University of California Press, 2005), 280.

The actual violence carried out against the Rohingya could not have occurred without a system of power that encouraged such harm. Philip Zimbardo, mastermind behind the infamous Stanford Prison Experiment, emphasizes the importance of macro-level analysis when understanding instances of hatred. In seeking to answer the question "How can good or ordinary people do extraordinarily evil things?" Zimbardo argues that we must "recognize the extent and limits of personal power, situational power, and systemic power."[24] Thus, examining the massacre of the Rohingya necessitates an analysis of the situational dynamics influencing individual behavior, as well as the larger systemic forces that foment the situation itself. Similarly, Chirot and McCauley name organization as one of the prerequisites for political mass murder. "The actual perpetrators [of genocide] function more efficiently and can overcome their moral reservations if they are well organized and led," suggesting the need for an institution to orchestrate and oversee large-scale massacres.[25]

For Myanmar, that institution is the military. With a standing force of over half a million, the Tatmadaw wields immense political, physical, and cultural force within the country. "From the moment they enter boot camp, Tatmadaw troops are taught that they are guardians of a country — and a religion — that will crumble without them."[26] This existential responsibility serves as a rationalization for extreme action. Zimbardo identifies this phenomenon, which he terms a 'justifying ideology' as a key tool in 'System Power's' institutional authorization of violence. "Ideology," he explains, "is a slogan or proposition that usually legitimizes whatever

---

24. Philip Zimbardo, *The Lucifer Effect: Understanding How Good People Turn Evil,* (New York: Random House, 2007), x.

25. Daniel Chirot and Clark McCauley, *Why Not Kill Them All? The Logic and Prevention of Mass Political Murder*, (Princeton: Princeton University Press, 2006), 57.

26. Hannah Beech, "Inside Myanmar's Army: 'They See Protesters as Criminals.," *The New York Times*, March 28, 2021, https://www.nytimes.com/2021/03/28/world/asia/myanmar-army-protests.html.

means are necessary to attain an ultimate goal…. Those in authority present the program as good and virtuous, as a highly valuable moral imperative."[27] The Tatmadaw constructed the Rohingya as an 'Other' that not only existed outside of Burmese Buddhism, but actively threatened its integrity. By elevating its members as defenders of their nation and faith, the military created a system that both permitted and rewarded the elimination of that 'threat.' Its particular brand of Buddhist nationalism goes far beyond simply justifying atrocities committed against the Rohingya; it validates them as necessary and even heroic.

Power hierarchies cultivated and enforced by military culture also create the necessary permission structure for mass murder. Both Chirot and McCauley and Zimbardo examine the role that obedience to authority plays in individuals' willingness to carry out violence against another human being. This can stem from the diffusion of responsibility as well as the fear of punishment that comes with power stratification. The former dynamic came to the forefront of the world's collective conscience after the Holocaust when scholars, political leaders, and laypeople alike began asking who deserves more blame: the Nazi soldier who killed an innocent Jewish person, or the Nazi leader who did not directly take a single life but ordered the deaths of thousands? An individual in the Tatmadaw would likely find it easier to commit otherwise unthinkable atrocities when directed by a superior to do so because the moral—and potentially legal—liability then falls upon the commanding officer. As Alexander Hinton notes in his analysis of the Cambodian genocide, "For most people, killing is easier when it is authorized by another person or institution."[28] Cultures that encourage the blind following of orders can therefore allow for large-scale perpetration of violence.

---

27. Zimbardo, *The Lucifer Effect*, 226.

28. Hinton, *Why Did They Kill?* 279.

In addition to the suspension of moral reservations, organizers of political mass murder such as the Rohingya genocide often employ fear and coercion to impel violence. This can take the form of individual threats, as well as ontological ones. Zimbardo notes that high exit costs can make ordinary people more likely to commit violence by intensifying the risks of disobedience.[29] Sure enough, an anonymous Tatmadaw military doctor confessed, "I want to quit, but I can't. If I do, they will send me to prison. If I run away, they will torture my family members."[30] Credible possibility of punishment can compel individuals to harm innocents even if they possess moral misgivings.

However, organizations that do not directly intimidate their members may still invoke threat as an impetus to violence. "Fear," conclude Chirot and McCauley, "is perhaps the key emotion for understanding genocide."[31] The justifying ideologies employed by System Power often fabricate an enemy, then appeal to fear of that enemy in order to encourage particular actions. National security, Zimbardo points out, is often marshalled as a justification for going to war even when no real threat exists.[32] This framework allows individuals to rationalize violence as acts of self-defense. Indeed, "the most powerful fear is fear of extinction, the fear that 'our' people, 'our' cause, 'our' culture, 'our' history may not survive. This fear will elicit the most violent and extreme reactions."[33] Members of the Tatmadaw imbibe a steady stream of Buddhist

29. Zimbardo, *The Lucifer Effect,* 274.

30. Beech, "Inside Myanmar's Army."

31. Chirot and McCauley, *Why Not Kill Them All?* 61.

32. Zimbardo, *The Lucifer Effect,* 274.

33. Chirot and McCauley, *Why Not Kill Them All?* 62.

nationalist propaganda that construe the Rohingya's very existence as a threat to their own.[34] Thus, murdering Rohingya men, raping Rohingya women, or burning Rohingya villages does not constitute a gross violation of innocent life, but a laudable defense of one's home and religion.

Though these ethnic tensions have existed in Myanmar for decades, if not centuries, they are increasingly aggravated by a larger trend towards retribalization. Humans' innate tendency to self-identify with a collective and equate their safety and wellbeing with one's own can instill a sense of belonging and camaraderie, but also enables inter-group conflict.[35] Erich Fromm names this phenomenon 'group narcissism,' and observes that it comprises "one of the most important sources of human aggression."[36] Perceived threats to the group simultaneously constitute threats to oneself. In Myanmar, where existence has been rendered a zero-sum game and cohabitation an impossibility, heterogeneity—the very presence of diverse peoples and customs—represents just such a threat. "Any group within the state's borders that does not accept its legitimacy on cultural grounds threatens its very integrity, the life of the state, and of the nation represented by that state. Even if a non-state cultural group wants to be loyal to the state it inhabits, suspicion that it is untrustworthy threatens the state and opens that group up to persecution."[37] This cycle of difference, distrust, and oppression has long characterized the Rohingya's status in Myanmar. De facto and de jure inequality between groups only increases the risk of conflict between them. Explaining the conditions that give rise to sadistic behaviors, Fromm emphasizes that "the power through which one group exploits and keeps down another tends to generate sadism in the

34. Beech, "Inside Myanmar's Military."

35. Chirot and McCauley, *Why Not Kill Them All?* 62.

36. Erich Fromm, *The Anatomy of Human Destructiveness*, (Greenwich: Fawcett Press, 1973), 231.

37. Chirot and McCauley, *Why Not Kill Them All?* 49.

controlling group."[38] Thus, Myanmar's extreme polarization leads to toxic group identification in which the endurance of the Bamar group is seen to depend on the Rohingya's elimination. Violence produced by this sort of retribalization is exacerbated by social stratification, which normalizes the ethnic majority's control, dominance, and exploitation of the minority.

In conceptualizing genocide as a dynamic and variable process, Hinton adopts a metaphor of heat to assess the relative likelihood of political mass murder. A 'cool' situation may not possess many or particularly extreme genocidal primes, while a 'hot' situation will feature a multitude of them.[39] Though the particular constellation of primes differs from situation to situation, he identifies political upheaval, structural divisions, identifiable target groups, and discriminatory political changes as common precursors to genocide. After the dissolution of the military regime in 2011, Myanmar underwent rapid sociopolitical transformation: free elections were held, Aung San Suu Kyi stepped up as the head of the civilian government, and independent media proliferated. Though much of the 'Western' world lauded these steps towards liberalization, they may have detrimentally impacted stability within the state. Retrospective analysis of Suu Kyi's rhetoric reveals long-standing anti-Islamic sentiments and a strong Buddhist nationalist bent. Despite her Nobel Peace Prize, she has never vocally intervened in the military's persecution of the Rohingya. Indeed, she expressed on record to the International Court of Justice her belief "that Rohingya is not an identity that should be recognized."[40] Aung San Suu Kyi's leadership may have actually lent legitimacy to her country's ethnic cleansing campaigns by casting it in the warm glow of democratization. However, research suggests that

38. Fromm, *The Anatomy,* 331.

39. Hinton, *Why Did They Kill?* 280.

40. Azeem Ibrahim, "Myanmar Has Blazed a Path to Democracy Without Rights," *Foreign Policy,* January 16, 2020, https://foreignpolicy.com/2020/01/16/myanmar-democracy-rohingya/.

"democratization has amplified polarization in the country, especially among different Buddhist groups, the government, and various ethnic groups, causing numerous factions to fend for their interests and leaving the Rohingya without a political ally."[41] The rapid expansion of political liberties has created space for the proliferation of Buddhist nationalist extremism. Though it may seem counterintuitive, sociopolitical transformation—even ostensibly 'good' transformation such as the expansion of democracy—often presages genocide. "Such events," writes Hinton, "upset the status quo, destabilize previous understandings and people's sense of well-being…intensify group division, force people to take sides, undermine social structures that promote cohesion and solidarity, and create a sense of threat and danger."[42] Anti-Rohingya sentiment certainly existed prior to Myanmar's governmental transition. The uncertainty, instability, and power vacuums that accompany political turmoil, however, exacerbated those tensions.

Codification of ethnic inequality under the Tatmadaw constitutes a genocidal prime as well. Myanmar possesses significant ethnolinguistic diversity, as well as a protracted history of ethnolinguistic hierarchization. "While all societies have a degree of pluralism, structural divisions crystallize more readily in situations in which the cleavages between groups cut across a number of domains, involve domination, exploitation, and inequality, are linked to a history of tension and conflict, or are reflected by political polarizations."[43] All of these dynamics are at play in Myanmar. Decades-long disenfranchisement campaigns gave legal clout to sociocultural distinctions between the Bamars and the Rohingya. Geographically isolated, religiously separate,

41. Jennifer Whitten-Woodring et al., "Poison If You Don't Know How to Use It: Facebook, Democracy, and Human Rights in Myanmar," *The International Journal of Press/Politics* 25, no. 3 (July 2020): 407–25. https://doi.org/10.1177/1940161220919666.

42. Hinton, *Why Did They Kill?* 282.

43. Ibid, 284.

linguistically unique, denied citizenship, and forced to bear special identification cards, the Rohingya were constructed as a stigmatized and identifiable group ripe for targeting.

GENOCIDAL PRIMES IN CYBERSPACE

The domestic institutional, social, and political factors discussed above created a 'hot' situation within Myanmar, but by themselves were not sufficient to propel the country to genocide. As Hinton discusses, genocide constitutes a "bricolage" of macrosocial, local, and personal elements. Ethnic cleansing operations like that of the Tatmadaw therefore involve not only particular environmental conditions, but also individual constructions of significance.[44] Much of the psychosocial meaning-making that fueled the Rohingya genocide took place on Facebook. An independent UN fact-finding mission into the situation in Myanmar found that "The role of social media [in the genocide] is significant. Facebook has been a useful instrument for those seeking to spread hate, in a context where, for most users, Facebook is the Internet. Although improved in recent months, the response of Facebook has been slow and ineffective."[45] The social media platform's reach and ubiquity made it the ideal means through which the Tatmadaw could spread inflammatory propaganda. Justifying ideologies constructed by the System Power in order to validate, necessitate, and reward extreme violence are propagated on Facebook, circulating with a speed and reach unmatched by any other form of media. Technological amplification of standard propagandic strategies therefore facilitated the spread of Buddhist nationalist pride, linked closely to anti-Rohingya rhetoric.

This unique cyberscape metastasizes prejudice and bigotry into a virulent hatred. In his psychological analysis of hatred, Willard Gaylin defines prejudice as "when the negative

---

44. Ibid, 287.

45. United Nations Human Rights Council. *"Report,"* 14.

attributes ascribed to a person by virtue of his or her being a member of a disdained or despised group are highlighted."[46] Essentializing an out-group—distilling an entire collective down to qualities possessed by a few individuals—enables moral disengagement that allows suffering and even justifies abuse. Anti-Rohingya Facebook posts display exactly this sort of prejudice. Many of them refer to all Muslims with terms of sexual deviance or transgression; the essentialization of the entire ethnic group as 'rapists' is particularly prominent. Drawing on Myanmar's history of colonization and British favoritism, Facebook posts frequently disparage the Rohingya as dangerous traitors as well.

Closely tied to prejudice disfavoring an out-group is bigotry favoring the in-group. Gaylin identifies this as a second precursor to hatred, one characterized by an intense "[partiality] to one's own group, religion, race, or politics and [intolerance] of those who differ."[47] In the case of Myanmar, this most frequently manifests in a strong sense of Buddhist nationalism. "On Facebook and offline, ultranationalists have framed Muslims as posing both a personal threat and a threat to the Buddhist majority nation. They have made claims about high Muslim birthrates, increasing Muslim economic influence, and Muslim plans to take over the country."[48] Though these claims do not reflect demographic realities, they nevertheless foster a stratified subjectivity in which 'superior' Buddhist Bamars face constant danger of impurification or even extinction from 'inferior' Muslim Rohingyas. Civilian government

---

46. Willard Gaylin, *Hatred: The Psychological Descent into Violence*, (New York: Public Affairs, 2003), 23.

47. Ibid, 26.

48. Christina Fink, "Dangerous Speech, Anti-Muslim Violence, and Facebook in Myanmar," *Journal of International Affairs* 73, no. 2 (Spring/Summer 2018).

officials have not vocally opposed this conceptualization for fear of alienating an influential voting bloc. The Tatmadaw, meanwhile, actively encourages ultranationalism because "military leadership benefits from Buddhist perceptions that it is defending the Buddhist-majority nation."[49] Vilification of the Rohingya as an existential threat to both Buddhism and Myanmar further entrenches the military. When substantial portions of the population view the Tatmadaw as critical to the defense of faith and country, they tend to turn a blind eye to its power abuses. The military, therefore, possesses a vested interest in stoking prejudice and bigotry.

Though hatred cannot be reduced to prejudice and bigotry, they are, as Gaylin writes, important "waystations" on the road to it. The transition to hate requires "a feeling of being threatened or humiliated," fueling fear and rage.[50] Facebook has provided the optimal platform for this transition. As part of its propaganda campaign, the Tatmadaw would post decontextualized images of dead bodies, then claimed they were the victims of Muslim-perpetrated massacres. Sham accounts also spread simultaneous rumors that Muslim groups were planning terrorist attacks on Buddhist communities, and vice versa.[51] This fear-mongering invoked widespread feelings of vulnerability and uncertainty. When anxiety and anger become linked to a particular person or group, hatred arises. Hatred in this sense does not refer merely to a strong negative emotion, but to a complex sentiment that requires an obsessive attachment to a subject. Gaylin defines it as "a sustained emotion of rage that occupies an individual through much of his life, allowing him to feel delight in observing or inflicting suffering on the hated

49. Ibid.

50. Gaylin, *Hatred*, 26.

51. Mozur, "A Genocide Incited on Facebook."

one. [Hatred] is always obsessive and almost always irrational."[52] The emphasis on longevity and relationality sets hatred apart from its more ephemeral and isolated cousin, anger. Nevertheless, anger plays a significant role in the fomentation of hate, which Gaylin sees as arising out of fear and rage directed towards a self-created enemy. Hate speech posted on Facebook—by both military and civilian accounts—effectively constructs the Rohingya as a target for those emotions.

Chirot and McCauley identify hatred as derivative of not only anger and anxiety, but also contempt and disgust. Targeted groups are often essentialized as dirty or profane, thereby endangering the 'purity' of the in-group and instilling a deep-seated fear that persists independent of any tangible threat. It is not a fear of physical harm, but of pollution. Fear of pollution involves the belief that an 'Other' is inherently dirty, disgusting, foul, and/or profane not by virtue of what they do, but simply because of who they are. Throughout history, countless instances of ethnic cleansing have stemmed from just such a fear; the forcible removal of a people "[mirrors] the same wish to cleanse the land of infidel pollution and danger."[53] The same principle appears at play in the Myanmar genocide. One emblematic Facebook posts reads, "These non-human kalar dogs, the Bengalis [Rohingya], are killing and destroying our land, our water and our ethnic people. We need to destroy their race."[54] This rhetoric explicitly links the existence of the Rohingya to the degradation of physical territory and cultural integrity. In this way, they are constructed as what Gaylin terms a 'territorial enemy.'[55] Ideological differences

52. Gaylin, *Hatred,* 34.

53. Chirot and McCauley, *Why Not Kill Them All?* 38.

54. Stecklow, "Why Facebook is Losing the War."

55. Gaylin, *Hatred,* 179.

between Buddhism and Islam do not necessarily underpin the hatred Bamars feel towards Rohingya; rather, a proximal group is targeted, and land used as a validation for hatred by fabricating the threat of an unjust deprivation of resources. The Rohingya have long constituted a territorial scapegoat, their very presence in the region ahistoricized and maligned as illegal. This rationalization also perpetuates the idea of Myanmar as a (Buddhist) 'homeland' that requires protection from outsiders. Thus, territory becomes a symbol that transcends its actual, physical importance. Preserving Myanmar and the Bamar ethnicity, in this formulation, therefore requires the elimination of the Rohingya entirely; the mere existence of the latter precludes that of the former.

Repeated invocations of Muslim men raping Buddhist women also conjure fears of pollution. In this case, the pollution is transferred from one party to another through sexual contact, symbolizing the implantation of contamination into the broader collective. The 'threatened' ethnic group is literally impregnated with impurity. Similarly, many inflammatory Facebook posts liken Rohingya to cockroaches, dogs, or pigs—animals commonly associated with filth and disease.[56] Such comparisons serve a dual purpose: they dehumanize the Rohingya, suspending the moral reservations that usually accompany inflicting harm upon another human, while also reinforcing the essentialization of the ethnic group as dirty and polluting.[57] Repeatedly associating the Rohingya with repulsiveness "justifies the violence against them because their disgusting characteristics threaten to pollute the environment and must be eliminated."[58] Anti-Rohingya hate speech on Facebook generates and feeds into fears that they will contaminate the

56. Stecklow, "Why Facebook is Losing the War."

57. Zimbardo, *The Lucifer Effect,* 223.

58. Chirot and McCauley, *Why Not Kill Them All?* 81.

Bamar 'essence,' thus rendering ethnic cleansing both permissible and obligatory as a form of group self-defense.

FACEBOOK AS A TOOL FOR HATE

However, Facebook is not merely a benign platform unwittingly hosting hateful language, or even a neutral tool harnessed for malevolent purposes. In fact, several of its defining characteristics exacerbate psychological tendencies recognized as key factors in the perpetration of genocide. The wealth of information available in modern media environments can drive individuals to take cognitive shortcuts due to an inability to allocate sufficient attention to all relevant inputs. Social media users may therefore resort to heuristics to simplify information-processing and expedite decision making. Though not inherently negative, such strategies can create and reinforce prejudices. Heuristic-based cognition in social media environments "encourages fringe-thinking through the confirmation of pre-existing biases."[59] Confronted with billions of data points jostling for finite attention, Facebook users often take mental shortcuts that can produce sweeping conclusions from minimal informational input. This leads to the perpetuation of stereotypes and prejudicial views of particular groups. Through this process, users "see millions of diverse individuals as a single object."[60] Universalizing judgements and character traits are assigned without care for nuance or even accuracy; *all* women are emotional, *all* Muslims are terrorists, *all* Rohingya are evil. An entire race, gender, nationality, or ethnicity is boiled down to a single perceived essence, categorized as entirely good or entirely bad. Such essentializing logic determines an individual's ontological guilt based not upon what they have

59. Peter J. Brown, Peter and Mario G. Beruvides, "The Heuristic-Based Framework for Attitude Certainty: How Technology and the Attention Economy Are Systematically Eroding Systematic Thinking," *The Psychologist-Manager Journal* 23, no. 2 (May 2020): 76–94. https://doi:10.1037/mgr0000107, 79.

60. Chirot and McCauley, *Why Not Kill Them All?* 82.

or have not done, but merely upon what group they belong to.

Often, essentialization occurs in tandem with the separate but parallel psychological process of dehumanization. "The misperception of certain others as subhuman, bad humans, inhuman, infrahuman, dispensable, or 'animals' is facilitated by means of labels, stereotypes, slogans, and propaganda images."[61] Facebook posts repeatedly referred to the Rohingya as 'kalars,' a derogatory, racialized slur used to highlight perceived phenotypical and ethnic differences between them and Bamars. This linguistic decision already establishes the Rohingya as an inferior 'other.' However, the dehumanization is further exacerbated by frequent essentializations of Rohingya as rapists, illegal immigrants, terrorists, and even animals such as pigs and dogs.[62] Hate speech like this is reminiscent of propaganda techniques during the Rwandan genocide, the perpetrators of which often called Tutsis 'cockroaches.' Indeed, dehumanization and essentialization have been repeatedly identified as key parts of the genocidal process.

> Each person is assessed not on the basis of his or her individual characteristics, but in terms of his or her membership in an abstract category that is essentialized, stigmatized, and targeted for elimination…. This ideological marking…further sets 'them' apart from the larger social community through devaluation. As less than fully human beings, these 'others' are depicted as legitimate targets of violence whose execution should not pose a moral dilemma.[63]

Deriding all Rohingya as terrorists or dogs crystallizes normally fluid and even minor differences between them and the Bamars, castigating them as an irreconcilable 'Other' fundamentally opposed to the ethnic majority.

---

61. Zimbardo, *The Lucifer Effect*, 223.

62. Stecklow, "Why Facebook is Losing the War."

63. Hinton, *Why Did They Kill?* 284.

Subjugating them to an infrahuman status creates a permission structure for normally unthinkable atrocities. It does not matter that one would never commit such acts upon another human being, because the Rohingya are not seen as human at all. The moral architecture that governs interactions between people no longer apply. To use Martin Buber's framework, relationships have devolved from 'I-Thou' to 'I-It.'[64] This fosters moral disengagement, allowing individuals to perpetrate extreme violence without causing an internal ethical crisis. Unburdened by self-censure, rape, murder, and massacre suddenly become reasonable courses of action. Suspending the humanity of an entire group of people can therefore heat up a situation already primed for genocide.

Though essentialization and dehumanization may constitute common hallmarks of genocide, these cognitive processes usually permeate society slowly and organically. With the introduction of communications technology, this can happen far more swiftly. Radio programs, for example, were effectively used during the Rwandan genocide to popularize dehumanizing rhetoric.[65] But while a radio program may reach a few thousand people who happen to tune in at a particular moment within a particular geographic area, a social media post persists and propagates without regard to space or time. Individuals in Myanmar could receive near-constant exposure to anti-Rohingya hate speech online, (re)creating dehumanizing prejudices over and over again.

Deindividuation fosters moral disengagement as well. In his analysis of the Stanford Prison Experiment, Zimbardo observes the pivotal role that physical disguises such as uniforms

---

64. Zimbardo, *The Lucifer Effect*, 223.

65. Rwandan Stories. "Hate Speech." Accessed November 15, 2021. http://www.rwandanstories.org/genocide/hate_radio.html.

or masks play in creating abusive environments. Such "disguises of one's usual appearance…promote anonymity and reduce personal accountability. When people feel anonymous in a situation, as if no one is aware of their true identity…they can be more easily induced to behave in antisocial ways."[66] Secure in the knowledge that identification of individual perpetrators is unlikely, deindividuated people may feel more comfortable committing acts of violence. Social media amplifies deindividuation to a degree seldom found in the physical world. Under cover of a screen name, with personal information limited or even fabricated, individuals recognize their own anonymity as one user in a community of billions. This generates a pervasive sense of impunity. Individuals will post or share opinions on Facebook that they may never express in real life, operating within a "general norm of permissiveness…that created a sense that [they] could do pretty much whatever they felt like doing because they were not personally accountable and could get away with anything because no one was watching."[67] Deindividuation on social media can therefore serve as a license for extremism and hate speech. Facebook played a critical role in the moral disengagement of an entire nation as it offered a high degree of anonymity that encouraged essentializing and dehumanizing rhetoric about the Rohingya, contributing to the activation of existing genocidal primes.

In fact, Facebook may have facilitated not only the spread of hate speech, but the collectivization of hatred itself. Gaylin identifies two related yet distinct groups bound by hate: communities of haters and cultures of hatred. "The community of haters," he writes, "is a group of disparate individuals who find one another and band together because of their shared

66. Zimbardo, *The Lucifer Effect*, 19.

67. Ibid, 368.

passion."[68] Facebook facilitates the development of such communities, creating a forum for people to bond over their common hatred and encouraging a process of identification in which individuals merge their own identities with that of the group as a whole. Technology makes this process easier than ever. Social media in particular allows for the transcendence of time and space, factors that may once have delimited when and where communities of haters could arise and what impact they could have. In allowing for the proliferation of anti-Rohingya sentiment, Facebook spawned an online community for individuals to share and (re)produce hate.

However, the social media platform seems to have expanded this phenomenon beyond an ad hoc group of virtual haters; it may have aided in the transformation of Myanmar into a culture of hatred. While a community of haters is artificially founded by its members based on their shared passion, a culture of hatred refers to "a natural community that breeds and encourages hatred."[69] Such cultures often possess a collective history and shared land, and leaders indoctrinate members into hatred towards a designated enemy. With its entrenched military, powerful Buddhist nationalist influence, and historic oppression of minority ethnic groups, Myanmar certainly seems to fit this description. Gaylin notes that when quotidian biases intersect with religious or nationalist credos, those biases can intensify and fixate on a particular scapegoated group.[70] This shared ideology and agreed-upon enemy can collectivize individual prejudice into communal hatred. Social media enables this process to occur on a vast scale and with a hitherto unimaginable speed, immune to the geographic and temporal limitations that bind other modes of communication. Though Myanmar already possessed the making of a culture of

68. Gaylin, *Hatred*, 218.

69. Ibid, 195.

70. Ibid, 244.

hatred, its deterioration was hastened by online hate speech. In short, Facebook did not cause the ethnic cleansing of the Rohingya, but it did enable the organizational, psychological, social, and cultural prerequisites of genocide.

FACEBOOK HATRED BEYOND MYANMAR

Worse, internal Facebook documents suggest that history is repeating itself. Unrest in Ethiopia between the government and residents of the northern Tigray region has escalated into civil war, with mass displacement, murder, and sexual violence. As in Myanmar, hate speech on social media has only aggravated the situation. "Facebook said it had observed a cluster of accounts affiliated with the [government-backed] militia group… using its platform to 'seed calls for violence,' promote armed conflict, recruit and fundraise."[71] Despite pledges of reform after the Rohingya genocide, the company has clearly failed to invest sufficient resources and money into protecting vulnerable minorities.

Halting the cycle of cyberhate and violence requires a close examination of the tragedies that have already unfolded, and what enabled them. In Myanmar, a unique array of sociopolitical factors already primed the country for genocide: the organizing institution of the military, a hierarchy permissive of atrocity, retribalization of ethnic identity, political unrest, rapid media liberalization, codification of difference and inequality. However, the mere presence of these conditions does not guarantee genocide as an inevitability. Facebook played a crucial role in perpetuating the psychological and ideological frameworks that transitioned Myanmar from political unrest to political mass murder. Posts on the platform played into existing prejudices,

71. Eliza Mackintosh, Eliza, "Facebook Knew it was Being Used to Incite Violence in Ethiopia. It Did Little to Stop the Spread, Documents Show," *CNN Business,* October 25, 2021, https://www.cnn.com/2021/10/25/business/ethiopia-violence-facebook-papers-cmd-intl/index.html.

rendering the Rohingya not simply a disenfranchised minority but the object of an obsessive and destructive hatred. This sentiment is intensified through rhetoric that constructs them as a corrosive danger to Buddhism and Myanmar itself. Confronted with this existential threat, the Bamar ethnic majority must exterminate the Rohingya in order to protect their own self-identity. Fears of pollution are reinforced by Facebook posts that dehumanize Rohingya individuals. By engaging in virtual essentializing discourse, individuals cultivate a moral disengagement that renders violence easier to rationalize, atrocities easier to commit, and the unthinkable easier to do. Though these cognitive processes exist in almost every instance of genocide, social media exacerbates them. The deindividuation that comes with an online avatar lessens users' sense of impunity, and Facebook's ability to transcend time and space enables the proliferation of moral disengagement and justifying ideologies on a hitherto unfathomable scale.

Patterns from Myanmar have emerged in Ethiopia, India, Afghanistan, Iraq, Sri Lanka, and even the United States.[72] Hate speech on Facebook amplifies tension, unrest, and prejudices that may otherwise not have precipitated genocide, serving as a trigger or activation switch for mass violence. If the global community hopes to prevent the Rohingya genocide from repeating itself in Tigray, in Kashmir, and in countless other places across the world, it must place pressure on Facebook to resource its market expansions and implement structural algorithmic reform. According to whistleblower Frances Haugen, "The raw version [of Facebook] roaming wild in most of the world doesn't have any of the things that make it kind of palatable in the United States, and I genuinely think there's a lot of lives on the line—that Myanmar and Ethiopia are

72. Amnesty International, "The Facebook Papers: What Do They Mean from a Human Rights Perspective?" November 4, 2021, https://www.amnesty.org/en/latest/campaigns/2021/11/the-facebook-papers-what-do-they-mean-from-a-human-rights-perspective/; Isabel Debre and Fares Akram, "Facebook's Language Gaps Weaken Screening of Hate, Terrorism," *Associated Press,* October 25, 2021, https://apnews.com/article/the-facebook-papers-language-moderation-problems-392cb2d065f81980713f37384d07e61f.

like the opening chapter."[73] If that is true, then the world cannot afford for the rest of the book to be written.

73. Mackintosh, "Facebook Knew."

CHAPTER TWO BIBLIOGRAPHY

Amnesty International. "The Facebook Papers: What Do They Mean from a Human Rights Perspective?" November 4, 2021, https://www.amnesty.org/en/latest/campaigns/2021/11/the-facebook-papers-what-do-they-mean-from-a-human-rights-perspective/.

Beech, Hannah. "Inside Myanmar's Army: 'They See Protesters as Criminals.'" *The New York Times*, March 28, 2021, https://www.nytimes.com/2021/03/28/world/asia/myanmar-army-protests.html.

Brown, Peter J., and Mario G. Beruvides. "The Heuristic-Based Framework for Attitude Certainty: How Technology and the Attention Economy Are Systematically Eroding Systematic Thinking." *The Psychologist-Manager Journal* 23, no. 2 (May 2020): 76–94. https://doi:10.1037/mgr0000107.

Chirot, Daniel and Clark McCauley. *Why Not Kill Them All? The Logic and Prevention of Mass Political Murder*. Princeton: Princeton University Press, 2006.

Debre, Isabel and Fares Akram. "Facebook's Language Gaps Weaken Screening of Hate, Terrorism." *Associated Press.* October 25, 2021, https://apnews.com/article/the-facebook-papers-language-moderation-problems-392cb2d065f81980713f37384d07e61f.

Fink, Christina. "Dangerous Speech, Anti-Muslim Violence, and Facebook in Myanmar." *Journal of International Affairs* 73, no. 2 (Spring/Summer 2018).

Fromm, Erich. *The Anatomy of Human Destructiveness*. Greenwich: Fawcett Press, 1973.

Gaylin, Willard. *Hatred: The Psychological Descent into Violence*. New York: Public Affairs, 2003.

Godrej, Dinyar. "A Short History of Burma." *New Internationalist,* April 18, 2008, https://newint.org/features/2008/04/18/history.

Hinton, Alexander Laban. *Why Did They Kill? Cambodia in the Shadow of Genocide.* Berkeley and Los Angeles: University of California Press, 2005.

Ibrahim, Azeem. "Myanmar Has Blazed a Path to Democracy Without Rights." *Foreign Policy,* January

16, 2020, https://foreignpolicy.com/2020/01/16/myanmar-democracy-rohingya/.

Mackintosh, Eliza. "Facebook Knew it was Being Used to Incite Violence in Ethiopia. It Did Little to Stop the Spread, Documents Show." *CNN Business.* October 25, 2021. https://www.cnn.com/2021/10/25/business/ethiopia-violence-facebook-papers-cmd-intl/index.html.

Maizland, Lindsay. "Myanmar's Troubled History: Coups, Military Rule, and Ethnic Conflict." *Council on Foreign Relations,* February 10, 2021, https://www.cfr.org/backgrounder/myanmar-history-coup-military-rule-ethnic-conflict-rohingya.

Mozur, Paul. "A Genocide Incited on Facebook, With Posts From Myanmar's Military." *The New York Times*, October 15, 2018, https://www.nytimes.com/2018/10/15/technology/myanmar-facebook-genocide.html.

Office of the High Commissioner of Human Rights. "Human Rights Council Opens Special Session on the Situation of Human Rights of the Rohingya and Other Minorities in Rakhine State in Myanmar." December 5, 2017, https://www.ohchr.org/EN/NewsEvents/Pages/DisplayNews.aspx?NewsID=22491.

Rwandan Stories. "Hate Speech." Accessed November 15, 2021. http://www.rwandanstories.org/genocide/hate_radio.html.

Rohingya Cultural Center. "History of the Rohingya." Accessed October 16, 2021. https://rccchicago.org/history/.

Stecklow, Steve. "Why Facebook is Losing the War on Hate Speech in Myanmar." *Reuters,* August 15, 2018. https://www.reuters.com/investigates/special-report/myanmar-facebook-hate/.

"The Rohingya Refugee Crisis is the Worst in Decades." *The Economist,* September 21, 2017, https://www.economist.com/graphic-detail/2017/09/21/the-rohingya-refugee-crisis-is-the-worst-

in-decades.

"UN Human Rights Chief Points to 'Textbook Example of Ethnic Cleansing' in Myanmar." *UN News,* September 11, 2017, https://news.un.org/en/story/2017/09/564622-un-human-rights-chief-points-textbook-example-ethnic-cleansing-myanmar.

United Nations General Assembly. *Promotion and Protection of the Right to Freedom of Opinion and Expression,* A/74/486, October 9, 2019, https://www.ohchr.org/Documents/Issues/Opinion/A_74_486.pdf.

United Nations Human Rights Council. *Report of the Independent International Fact-finding Mission on Myanmar*, A/HRC/39/64, September 12, 2018, https://www.ohchr.org/Documents/HRBodies/HRCouncil/FFM-Myanmar/A_HRC_39_64.pdf.

United States Holocaust Memorial Museum. "Burma's Path to Genocide." Accessed October 18, 2021. https://exhibitions.ushmm.org/burmas-path-to-genocide/timeline.

Whitten-Woodring, Jennifer, Kleinberg, Mona S., Thawnghmung, Ardeth, and Thitsar, Myat The. "Poison If You Don't Know How to Use It: Facebook, Democracy, and Human Rights in Myanmar." *The International Journal of Press/Politics* 25, no. 3 (July 2020): 407–25. https://doi.org/10.1177/1940161220919666.

Zimbardo, Philip. *The Lucifer Effect: Understanding How Good People Turn Evil.* New York: Random House, 2007.

CHAPTER THREE:
SOLUTIONS TO RACIALIZED HATE SPEECH ON FACEBOOK

Facebook's exacerbation of hate speech simultaneously represents a moral dilemma, a policy puzzle, as well as a business conundrum. Each of these aspects are inextricably intertwined. However, current efforts ranging from corporate responsibility to global governance fail to grapple with the often-contradictory interests of the company, the sovereign state, and the world at large. Examining the strengths and shortcomings of ongoing initiatives at each of these scales can illuminate the steps that must be taken to mitigate the harm of Facebook. No existing measure alone is adequate, nor will an adequate measure be easy. The issue of hate speech on Facebook falls at the fuzzy nexus of united regulation and unimpeded sovereignty, private free speech and public censorship. To safeguard one principle, one often must suppress its corollary. Corporations, nations, and the international community must therefore consider the sacrifices they are willing to make: is Facebook's vision of connecting the globe worth the cost of safety and wellbeing? Thwarting the recurrence of social media-fueled genocide will require unprecedented and innovative action coordinated from the United Nations to Menlo Park.

THE UNITED NATIONS: A CRISIS OF JURISDICTION

Since its conception, the United Nations has dedicated itself to the advancement of lofty, arguably unattainable ideals. Its founding charter outlines the organization's purpose in a series of starry-eyed articles: "to take collective effective measures for the…removal of threats to the peace;" "to strengthen universal peace;" to "[promote and encourage] respect for human rights and for fundamental freedoms for all;" and to "[harmonize] the actions of nations."[1] Two hundred states have agreed to these ambitions in theory, at least. However, the realization of

1. United Nations, *Charter of the United Nations,* 1945, 1 UNTS XVI, https://www.un.org/en/about-us/un-charter/full-text, art. 1.1-1.4.

universal peace and harmony has proven elusive, regularly running up against the harsh and contradictory conditions of reality: differential power, prejudice and racisms, human greed, governmental and corporate corruption. In fact, shortly after the delineation of the UN's admirable aims, the body's charter includes a clause acquiescing to the preeminence of national vice over global virtue: "Nothing contained in the present Charter shall authorize the United Nations to intervene in matters which are essentially within the domestic jurisdiction of any state."[2] Article 2.7 may constitute the most consequential sentence of the entire UN charter. It forbids intervention into domestic affairs, yet fails to specify exactly what falls under that category, empowering Member States to cry 'sovereignty' whenever they disagree with a UN edict. Although independence is an important and arguably inherent right of any state, the typical interpretation of Article 2.7 cripples any collective effort to actually achieve any of the objectives laid out in the first article. The United Nations archives overflow with documents that report findings or request action on pressing global issues but lack any enforcement power. For fear of violating state sovereignty, the UN must recommend, not require.

This stalemate between efficacy, authority, and mandate on the one hand and impotence, sovereignty, and entreaty on the other characterizes international efforts to address hate speech on social media. Fernand de Varennes, the UN special rapporteur on minority issues, has called digital hate speech "one of today's most acute challenges to human dignity and life."[3] He has advocated for Facebook to give greater consideration towards marginalized groups when assessing controversial content, and for the unification of corporate and international hate speech

2. Ibid, art. 2.7

3. "Hate Speech on Facebook Poses 'Acute Challenges to Human Dignity' – UN Expert," *UN News,* December 23, 2020, https://news.un.org/en/story/2020/12/1080832.

standards. Despite the rapporteur's astute and timely advice, the United Nations cannot impose legally obligatory directives to that end. Jurisdiction—already a perilous quagmire for the UN—grows even trickier when it concerns private enterprises like corporations. Because it is not a signatory of the UN charter, Facebook need not abide by UN treaties or mandates. The United Nations has nevertheless worked within its narrow sphere of influence to formulate policy addressing online hate speech from multiple angles. Lack teeth they might, but the UN's efforts to promote corporate social responsibility and protect the freedom of expression can serve as inspiration for more substantive action.

Upon first glance, it appears that the United Nations already possesses the requisite documents to justify the regulation of hate speech on Facebook and other social media platforms. The Universal Declaration of Human Rights—one of the premier components of the international law pantheon—unequivocally establishes the right to "life, liberty, and security of person."[4] Online rhetoric that incites violence and even genocide demonstrably violates this fundamental right. However, while documents such as this purportedly occupy a space of universality, the realm in which they hold the force of law remains far more restricted. Only 48 UN member states signed on to it, and as a mere 'declaration' it did not initially impose legal obligations. Most scholars concur that it has since achieved the status of customary international law, which does grant it binding authority.[5] Further, many of the rights within the Universal Declaration have been expanded and codified in subsequent treaties such as the International Covenant on Civil and Political Rights and the International Covenant on Economic, Social, and

4. United Nations, *Universal Declaration of Human Rights*, December 10, 1948, 217 A (III), https://www.un.org/en/about-us/universal-declaration-of-human-rights, art. 3.

5. Hurst Hannum, "The UDHR in National and International Law," *Health and Human Rights* 3, no. 2 (1998): 144–58, https://doi.org/10.2307/4065305.

Cultural Rights. These act with the force of law upon every nation in the world, preventing government action that would violate the terms of such covenants and imposing certain obligations upon states to protect the rights therein. Yet these duties and constraints apply only to sovereign states; corporations—such as Facebook—need not abide by them.

In fact, the status of private entities under international law has long remained murky. Though they have legal personhood under United States law as of the *Citizens United vs. FEC* decision in 2010, international courts do not "specifically distinguish between natural persons and juridical persons."[6] Such ambiguity renders the assignation of liability difficult in instances such as the Myanmar genocide, when corporations contribute to genocide or crimes against humanity without actually perpetrating them. More recent documents such as the UN's Guiding Principles on Business and Human Rights may elucidate the subject.

Published in 2011, it begins by calling for the implementation of its subsequent recommendations with regard to the vulnerability and systemic marginalization of particular groups. Against this backdrop, the Guiding Principles establish the responsibility of states to protect against human rights abuses within their jurisdiction, including by third parties like business enterprises. Individual states and the international community must therefore devise legal frameworks to ensure that businesses respect human rights. The businesses themselves, however, do not bear the same responsibilities as states. Corporations should respect—though they need not proactively protect—human rights and avoid infringing upon them, addressing violations should they arise. These obligations exist "over and above compliance with national laws and regulations," indicating the responsibility of businesses to respect human rights

6. José E. Alvarez, "Are Corporations 'Subjects' of International Law?" *Santa Clara Journal of International Law* 9, no. 1 (2011): 1—36, https://digitalcommons.law.scu.edu/scujil/vol9/iss1/1/, 9.

transcends the responsibility of states to protect them.[7] Therefore, even if an individual nation violates the rights of its citizens, companies doing business within that nation do not receive a free pass to do the same. In instances of state-sanctioned infringement, "the responsibility to respect human rights requires that business enterprises…seek to prevent or mitigate adverse human rights impacts that are directly linked to their operations, products or services by their business relationships, even if they have not contributed to those impacts."[8] By such logic, Facebook bears clear responsibility for preventing and mitigating the impact of hate speech on its site. This rebuts the company's claim that it merely serves as a neutral platform. Instead, Facebook must proactively prevent the proliferation of hate speech and, if such efforts fail, take substantive action to remedy the situation. Of course, any claim that Facebook 'must' do something should be scrutinized. The very first page of the document explicitly states that "nothing in these Guiding Principles should be read as creating new international law obligations."[9] In short, none of the bold claims of corporate responsibility or liability can be enforced, rendering them little more than pretty words and pleasant ideas.

The challenging question of Facebook's obligations to uphold human rights is complicated not only by its status as a private corporation, but also by its ethereality. How does international law operate in cyberspace? Can online behavior be juridically demonstrated to perpetrate offline human rights abuses? Who possesses jurisdiction over the Internet, and how can we demarcate those boundaries? These questions grew far more urgent after Facebook's role

---

7. Office of the High Commissioner of Human Rights, *Guiding Principles on Business and Human Rights*, 2011, HR/PUB/11/04, https://www.ohchr.org/sites/default/files/documents/publications/guidingprinciplesbusinesshr_en.pdf, 13.

8. Ibid, 14.

9. Ibid, 1.

in the Rohingya genocide became clear. In its 2019 report regarding the application of human rights law to online hate speech, the UN Special Rapporteur on Promotion and Protection of the Right to Freedom of Opinion and Expression called out this atrocity specifically. "The consequences of ungoverned online hate can be tragic," the report notes, "as illuminated by Facebook's failure to address incitement against the Rohingya Muslim community in Myanmar."[10] With the Rohingya genocide as its backdrop, it lays out recommendations for both states and companies, urging all parties to protect human rights online with the same rigor as they do offline.

Despite this resolute opening, the report quickly grows more tepid and nebulous. The Special Rapporteur makes a broad call for corporations to consider human rights when addressing hate speech on their platforms, but fails to elaborate on what such a consideration should look like. Similarly, recommendations for increased transparency regarding allowable versus removable content neglect to provide examples, targets, or evaluative measures. More specific injunctions fall short as well. A proposal that companies draw upon human rights law to designate protected identities appears sound, until one realizes that no human rights law explicitly or consistently establishes such categories. Even if such frameworks existed, this report faces the same issue as the Guiding Principles on Business and Human Rights; it lacks potency. Its recommendations for corporations are just that—recommendations. The Special Rapporteur can advise that companies take its findings under consideration, but cannot require action or inflict consequences.

10. Office of the High Commissioner of Human Rights, *Report on Online Hate Speech*, October 9, 2019, A/74/486, https://documents-dds-ny.un.org/doc/UNDOC/GEN/N19/308/13/PDF/N1930813.pdf?OpenElement, 16.

One ongoing project may solve the enforceability issues faced by other United Nations measures. Convened annually since 2014, the Intergovernmental Working Group on Transnational Corporations and Other Business Enterprises with Respect to Human Rights represents the international community's attempt to devise a legally binding instrument that would regulate the activities of transnational corporations with respect to human rights. The treaty's current draft addresses victim rights and protections, prevention of human rights atrocities, access to remedy, issues of legal liability, and jurisdiction.[11] If the working group ever completes its task, the final product could possess the clout of international law. However, that is a big 'if.' Given that it has taken seven years to yield a draft document, finalization and ratification seems a distant goal.

The working group's testudinal pace exemplifies a major drawback of transnational policy coordination. Unifying two hundred-odd states into a coalition with a cohesive approach towards any issue—let alone one as novel and complex as digital hate speech—represents a major diplomatic and administrative challenge. Global governance systems are bureaucratic behemoths, and even a successful attempt to design a legally binding regulatory instrument may take so long as to render the final product anachronistic. These impracticalities inhibit substantive and actionable contributions. Ultimately, the United Nations may be best positioned to articulate values and visions, and leave the actual mitigation of Facebook hate speech to other players.

---

11. Office of the High Commissioner of Human Rights, *Legally Binding Instrument to Regulate, in International Human Rights Law, the Activities of Transnational Corporations and Other Business Enterprises*, August 17, 2021, A/HRC/49/65/Add. 1, https://www.ohchr.org/sites/default/files/Documents/HRBodies/HRCouncil/WGTransCorp/Session6/LBI3rdDRAFT.pdf.

SOVEREIGN STATES: STRONG BUT SEPARATE ENFORCEMENT

Because the United Nations so often finds itself hamstrung by accusations of jurisdictional overreach, sovereign entities may prove more effective regulators. Within the limits of some very loose international agreements, states have absolute control over the relative liberty corporations possess within their borders. Some states—such as China—severely restrict that liberty. Though at least a partial market economy, the state retains extensive influence even over ostensibly private enterprises; one law, for example, requires any citizen, organization, or business to "support and cooperate in national intelligence work."[12] The near-omnipotence of the Chinese government gives it an easy path out of the moral and political quagmire of Facebook: it simply avoids the quagmire entirely. Dubbed "The Great Firewall," China operates an extensive censorship system that blocks most social media platforms owned by Western companies—including Facebook.[13] Despite its ubiquity throughout the rest of the world, Facebook cannot be accessed in China without a VPN. China's government instituted the ban in 2009 after unrest in the heavily oppressed Xinjiang region in order to sever connections between residents and the outside world. Eliminating Facebook in a country does, of course, eliminate Facebook hate speech in that country. However, trading virtual vitriol for centralized censorship does not constitute a desirable solution, and few states would be willing to risk the public backlash that would surely follow a unilateral prohibition of the platform.

On the opposite side of the spectrum lies the United States. Fervently capitalist, beholden to corporate interests, and seemingly allergic to federal economic regulation, the US has

---

12. Richard McGregor, "How the State Runs Business in China," *The Guardian,* July 25, 2019, https://www.theguardian.com/world/2019/jul/25/china-business-xi-jinping-communist-party-state-private-enterprise-huawei.

13. Eloise Barry, "These Are the Countries Where Twitter, Facebook and TikTok Are Banned," *Time*, January 18, 2022, https://time.com/6139988/countries-where-twitter-facebook-tiktok-banned/.

historically dealt with hate speech on Facebook by turning a blind eye to it. The country's founding commitment to free expression renders any limitation on speech—even hateful speech—a political minefield. More recently, some legislators and attorneys general have targeted tech giants, including Facebook, for violating antitrust laws. A bipartisan coalition in the House of Representatives has proposed bills that would place a higher burden of proof on companies to demonstrate that planned mergers are sufficiently competitive.[14] Such a policy would make it much more challenging for corporations to acquire and neutralize potential rivals, as Facebook did with Instagram. Similarly, in December 2020 the Federal Trade Commission sued the company, alleging that Facebook maintains its veritable monopoly over social networking through anticompetitive practices.[15]

Though these measures indicate unprecedented legislative bellicosity towards Big Tech, they do not directly address hate speech. Reducing Facebook's market power could allow for the emergence of new social media platforms with more stringent content standards, moving users away from Facebook and curtailing the scope and intensity of its impact. However, given that over half of the world's Internet users regularly access Facebook, this outcome is far from certain. Both the Chinese and American approaches to Facebook fail to substantively address hate speech on Facebook. In the European Union, however, proposed legislation offers innovative and wide-ranging policies that could not only require social media companies to more

14. Cecilia Kang, "Lawmakers, Taking Aim at Big Tech, Push Sweeping Overhaul of Antitrust," *The New York Times*, June 11, 2021, https://www.nytimes.com/2021/06/11/technology/big-tech-antitrust-bills.html#:~:text=WASHINGTON%20%E2%80%94%20House%20lawmakers%20on%20Friday,to%20monopoly%20laws%20in%20decades.

15. Federal Trade Commission, "FTC Sues Facebook for Illegal Monopolization," news release, December 9, 2020, https://www.ftc.gov/news-events/news/press-releases/2020/12/ftc-sues-facebook-illegal-monopolization.

assiduously remove hateful content from their platforms, but proactively prevent such content in the first place.

Approved by the European Parliament in January 2022, the Digital Services Act (DSA) imposes a system of rules, regulations, and penalties meant to make the Internet a safer place for users.[16] It requires social media platforms to publicly disclose how many content moderators they employ and what languages they speak, a direct response to Facebook's failures in Myanmar. Platforms would also need to establish accessible and timely mechanisms for complaints and redress, allowing users to contest content moderation decisions. However, the DSA recognizes that corporations' first loyalty is to their own bottom line, not the rights and security of their users. Therefore, the legislation would subject companies to external risk audits and create stronger public oversight of the platforms. Unlike the United Nations, which has published multiple reports suggesting similar measures, the European Union possesses the ability to punish noncompliance. Corporations in violation of the Digital Services Act could face fines of up to 10% of their global revenue.[17]

Passed in tandem with the DSA, the Digital Markets Act (DMA) addresses the unjust market systems that have allowed Facebook to accrue power virtually unchecked. The DMA places new obligations on industry 'gatekeepers,' or firms that hold significant market power. It would require companies such as Facebook to cease favoring their own products and prevent other noncompetitive practices that quash the success of new, potentially more ethical social

16. European Parliament. "Digital Services Act: Regulating Platforms for a Safer Online Space for Users." News release, January 20, 2022, https://www.europarl.europa.eu/news/en/press-room/20220114IPR21017/digital-services-act-regulating-platforms-for-a-safer-online-space-for-users.

17. Ibid.

media platforms.[18] Failure to uphold these standards and comply with a fair, open digital marketplace could provoke additional fines of up to 10% of global revenue. Enforcement of both the DSA and DMA would rest with the European Commission, not individual European states.[19] This technicality is meant to bypass nations with lax regulations such as Ireland, where most Big Tech companies house their European headquarters. If approved by EU Member States, the Digital Services Act and Digital Markets Act would represent the world's most comprehensive action against online hate speech and its offline harm.

Both of these measures build on the 2016 Code of Conduct on Countering Illegal Hate Speech Online. Actualizing the dual-track schemas put forth by the United Nations, the Code of Conduct creates responsibilities for both nations and corporations. It commits IT companies to supporting EU efforts to "respond to the challenge of ensuring that online platforms do not offer opportunities for illegal online hate speech to spread virally."[20] Platforms must respond "expeditiously" to reports of hate speech—ideally within 24 hours—and maintain a high standard of transparency regarding impermissible content and notification procedures.[21] Unlike many other national technological policies, the Code of Conduct also obligates Internet companies to support educational programs that build critical thinking skills, inoculating the

18. European Commission, "The Digital Markets Act: Ensuring Fair and Open Digital Markets," accessed December 1, 2021, https://ec.europa.eu/info/strategy/priorities-2019-2024/europe-fit-digital-age/digital-markets-act-ensuring-fair-and-open-digital-markets_en.

19. Ibid.

20. European Commission. "European Union Code of Conduct on Countering Illegal Hate Speech Online." June 30, 2016, https://ec.europa.eu/info/policies/justice-and-fundamental-rights/combatting-discrimination/racism-and-xenophobia/eu-code-conduct-countering-illegal-hate-speech-online_en, 1.

21. Ibid, 1.

populace against disinformation and heuristic-induced prejudices. These measures are designed to prevent the type of online behavior that led to the Rohingya genocide in Myanmar.

The Code of Conduct on Countering Illegal Hate Speech Online is also extrapolated from a 2008 EU framework criminalizing certain forms of racist and xenophobic expression. European Union member states must punish speech that is "publicly inciting to violence or hatred directed against a group of persons or a member of such group defined by reference to race, colour, religion, descent or national or ethnic origin," and/or "publicly condoning, denying, or grossly trivialising" crimes of genocide, crimes against humanity, and war crimes.[22] This wording already lends the Code of Conduct more clout than similar limitations on free speech in the United States; it forbids not only incitement to violence, but also incitement to hatred. Issuing this broader injunction relieves the hefty burden of proving a direct connection between digital language and physical violence.

The initial 2008 framework also contains an article holding legal persons liable for punishable speech. Constituting another departure from United States law, this raises the possibility of Facebook facing culpability for hate speech posted on its platform. In the US, infamous Section 230 of the Communications Decency Act states that "no provider or user of an interactive computer service shall be treated as the publisher or speaker of any information provided by another information content provider."[23] This small provision possesses massive consequences. It essentially removes liability from any social media company for the content hosted on their platforms, even content that is libelous, inciteful, or otherwise judicially

---

22. European Council, *Council Framework Decision on Combating Certain Forms and Expressions of Racism and Xenophobia by Means of Criminal Law,* November 28, 2008, 2008/913/JHA, https://eur-lex.europa.eu/legal-content/en/TXT/?uri=CELEX%3A32008F0913.

23. Congress.gov, "S.314 - 104th Congress (1995-1996): Communications Decency Act of 1995," February 1, 1995, https://www.congress.gov/bill/104th-congress/senate-bill/314/, § 230.

condemnable. Passed in 1996, the United States Congress intended Section 230 to provide the legal wiggle room viewed as necessary for the growth of the burgeoning tech industry.[24] More than 25 years later, however, it is evident that corporations no longer need governmental stimulus. Where once Internet companies required protection from the public, now the public needs protection from the Internet companies. The European Union's Code of Conduct on Countering Illegal Hate Speech addresses this development, recognizing that some degree of regulation—a small sacrifice of personal liberty—may be necessary to guarantee a greater degree of security—a realization of collective liberty. However, despite its innovative approach, the Code of Conduct possesses significant geographical limitations. It is, after all, the 'European Union's' code, and therefore does not apply to the vast majority of the world's population that dwells outside of the EU. The Digital Markets Act and Digital Services Act face similar constraints. The advantage of enforceability that sovereign solutions afford comes at the cost of universality, leaving the issue of online hate speech only spottily addressed.

CORPORATE RESPONSIBILITY: NECESSITY OR IMPOSSIBILITY?

Reining in the harm caused by Facebook may elude governing bodies. Multilateral institutions such as the United Nations can standardize regulations, coalescing the international community into a united front against digital hate speech and its real-world damage. However, such measures often lack the force of law, rendering them all but ineffective. Policies initiated by sovereign entities such as the European Union face the inverse problem; they can impose enforceable rules and consequences, but only within their limited jurisdictions. If each nation were to impose distinct laws governing hate speech on social media platforms, cyberspace could

24. Electronic Frontier Foundation, "CDA 230," Accessed February 14, 2022, https://www.eff.org/issues/cda230.

be carved into dozens of disparate regulatory zones. Such an outcome would be not only inconvenient, but almost inconceivable. After all, the Internet in general and social media in particular derive their utility from their unsurpassed ability to connect individuals across the globe. Discordant and uncoordinated policing of Facebook seems both impractical and unpopular.

Further, both global and national governmental systems are riddled with red tape. Drafting even limited or impotent legislation would likely take so long that it would be obsolete by the time of ratification, forcing the process to begin anew. Technology progresses with a rapidity utterly foreign to the halls of bureaucracy. Facebook's algorithm—and the way users engage with it—changes frequently, producing phenomena and potential problems that lawmakers may not anticipate. Cumbersome legislative leviathans lack the speed and agility to keep pace with Silicon Valley. Indeed, rapid advancement and a high tolerance for risk has characterized Facebook ever since the company's infancy when Zuckerberg famously exhorted his employees to "move fast and break things."[25] With plenty of things now broken, perhaps only Facebook itself can move fast enough to repair them.

Once again, Myanmar provides a grim illustration for the rest of the world. Places as diverse as India and France, Ethiopia and the United States possess high degrees of racial/ethnic tension, exacerbated by social media and their own unique constellations of genocidal primes, as discussed in the previous chapter. Given the ubiquity of both conflict and Facebook, what steps can be taken to decouple the platform from its aggravation of global hatred? Any such progress will first require Facebook to devote its resources more equitably around the world.

---

25. "Mark Zuckerberg's Letter to Investors: 'The Hacker Way,'" *Wired*, February 1, 2012, https://www.wired.com/2012/02/zuck-letter/.

Investigations in the aftermath of the Rohingya genocide revealed a gross negligence on the part of the company as it aggressively pursued market expansion in Myanmar without allocating adequate attention, funding, or staffing to its operations there. During the rapid deregulation of media after the dissolution of the military junta, telecommunications access exploded. In 2012, Internet penetration hovered around just 1%; by 2018, it reached 30%.[26] Simultaneously, the price of SIM cards plunged by 99%, leading millions of Burmese citizens to purchase smartphones—smartphones from which they could access Facebook. With its synchronized messaging system, news content, and entertainment platform, Facebook use in Myanmar skyrocketed. In order to capitalize on the app's popularity, "Myanmar's mobile phone operators began offering a sweet deal: use Facebook without paying any data charges."[27] The coincidence of media liberalization, technologization, and economic incentivization produced an exponential increase in the prevalence of Facebook. According to a Reuters investigation, Myanmar possessed 1.2 million Facebook users in 2014. Just four years later, it had 18 million.[28] Unlike in the United States where the platform serves primarily to connect people, many people in Myanmar utilize it as their primary news source, and even equate it with the Internet itself.[29]

However, the expansion of Burmese Facebook users was not accompanied by an expansion of Burmese Facebook infrastructure. The company had no permanent staff members in the country, and in 2015 it employed just four Burmese-speaking content moderators to

---

26. Jennifer Whitten-Woodring et al., "Poison If You Don't Know How to Use It: Facebook, Democracy, and Human Rights in Myanmar," *The International Journal of Press/Politics* 25, no. 3 (July 2020): 407–25. https://doi.org/10.1177/1940161220919666.

27. Steve Stecklow, "Why Facebook is Losing the War on Hate Speech in Myanmar," *Reuters,* August 15, 2018. https://www.reuters.com/investigates/special-report/myanmar-facebook-hate/.

28. Ibid.

29. Whitten-Woodring et al., "Poison."

review the posts of Myanmar's 7.3 million users.[30] Spread that thinly, the moderators had no hope of examining every reported violation, and countless instances of anti-Rohingya content fell through the cracks. Hate speech also benefited from a scarcity of technological resources. Facebook's algorithm for identifying and reporting problematic posts operates primarily in 'global' languages such as English, Spanish, and Mandarin, creating a language barrier that exposes linguistic minorities to higher levels of online hatred.[31] Further, the company's translation technology is largely incompatible with Burmese script, producing dangerously erroneous results. One post discovered by Reuters read in Burmese, "Kill all the kalars that you see in Myanmar; none of them should be left alive;" the English translation rendered it as, "I shouldn't have a rainbow in Myanmar."[32] The effects of such technological failures like this are worsened by other sociopolitical dynamics within the country. "Facebook gained influence at a time when government publications seemed to condone extreme speech and when trust in foreign media was declining. Therefore, Facebook became popular when conditions in Myanmar were ripe for online extreme speech to occur and for disinformation to remain unchallenged."[33] Despite the presence of domestic upheaval, ethnic tension, and other genocidal primes, the company did not curb the pace of its expansion in Myanmar, nor did it adjust its content moderation strategies. Facebook's aggressive business practices and refusal to sufficiently resource expanding markets contributed to one of the worst humanitarian crises of the 21st century.

---

30. Stecklow, "Why Facebook is Losing the War."

31. Billy Perrigo, "Facebook Says It's Removing More Hate Speech Than Ever Before. But There's a Catch," *Time,* November 27, 2019, https://time.com/5739688/facebook-hate-speech-languages/.

32. Stecklow, "Why Facebook is Losing the War."

33. Whitten-Woodring et al., "Poison."

These shortcomings are not limited to Myanmar. Rather, they constitute a pattern within Facebook's global operations, demonstrating a clear prioritization of its Euro-American users. A recent study of Facebook content in India exposed rampant disinformation and hate speech, particularly targeted towards Muslims. India currently constitutes Facebook's largest market, with over 340 million regular users. However, despite the nation's importance to the company's bottom line, Facebook still chooses to allocate its ample resources elsewhere. According to internal documents, it devotes 87% of its anti-misinformation budget to the United States, despite Americans representing only 10% of Facebook users.[34] This blatant disregard for the majority of the global population perpetuates the hatred that already exists on the platform and precludes the mitigation of resulting violence. Preventing another genocide like that in Myanmar will require Facebook to invest time, funding, technology, and personnel outside the United States.

Part of this negligence also manifests in Facebook's 'race-blind' approach to content moderation. By blinding itself entirely to historic and contemporary realities of marginalization, the company fails to adequately protect vulnerable groups from violence seeded on its platform. The effects of this decontextualization can be starkly seen in Myanmar. Privileging dominant Buddhist nationalist extremists to the same degree as the disenfranchised Rohingya minority permitted the proliferation and reinforcement of hateful rhetoric that—while perhaps not in strict violation of the Community Standards—nevertheless engendered harm. Facebook has made repeated assurances that it takes 'local nuances' such as racial slurs into consideration during the evaluation process. However, content moderators have confessed that "the rules were

34. Frenkel, Sheera and Davey Alba. "In India, Facebook Grapples with an Amplified Version of Its Problems." *The New York Times*, November 9, 2021. https://www.nytimes.com/2021/10/23/technology/facebook-india-misinformation.html?referringSource=articleShare.

inconsistent; sometimes they could make exceptions and sometimes they couldn't."[35] In many instances of cultural or linguistic ambiguity, users received the benefit of the doubt and questionable content remained on the platform.[36] If Facebook hopes to impede the repetition of history, it must reimagine its approach to hate speech. It must recognize that cyberspace does not exist in a vacuum, and therefore ground its Community Standards in local contexts of power, oppression, and privilege.

Additionally, Facebook should also consider reforming the underlying design that not merely enables but actively encourages the spread of hate speech on its platform. The piece of technology that prophesied Facebook's innovation and enshrined its dominance—the News Feed algorithm—preys upon the natural human proclivity towards fear-, disgust-, or rage-inducing content. This proclivity, known as negativity bias, "accounts for our tendency to remember episodes of threat and fear more strongly than periods of calm and peaceful relations with other groups."[37] It may have begun as an adaptive strategy, but in modern times the bias leads to individuals devoting more time, attention, and memory to negative events. Chirot and McCauley note that this can engender intergroup conflict; group members focus more on the perceived dangers and uncertainties posed by the 'Other,' creating an abstract yet pervasive sense of threat that may incite violence.[38] Social media algorithms only exacerbate these effects. Meant to maximize user engagement, the News Feed serves customized content to each user based on what that individual has demonstrated an interest in or preference for. However, because human

35. Stecklow, "Why Facebook is Losing the War."

36. Ibid.

37. Daniel Chirot and Clark McCauley, *Why Not Kill Them All? The Logic and Prevention of Mass Political Murder*, (Princeton: Princeton University Press, 2006), 64.

38. Ibid, 64-65.

beings possess a psychological disposition towards negative information, the News Feed disproportionately 'dishes up' extremist content. Subsequent adjustments to the algorithm—such as the incorporation of Meaningful Social Interaction (MSI)—only intensified this tendency, saturating people's Facebook feeds with posts more likely to stoke existing racial or ethnic tensions.[39]

The MSI revision is not irreversible. In fact, when confronted with its dangers, Facebook executives decided to scale back the News Feed's MSI component, but only for particularly controversial topics in particularly volatile places—including Myanmar. Despite suggestions from internal research teams to extrapolate that decision to the rest of the platform, the company refused to curtail MSI across the board, fearing that doing so would compromise the integrity of what made Facebook, Facebook.[40] However, this retroactive, compartmentalized concession remains insufficient. Confining algorithmic reform to states that have already experienced social media-influenced ethnic conflict fails to forestall the exacerbation of genocidal primes in other locations. Diluting MSI and implementing other systemic changes to social media platforms would by no means eradicate the possibility of mass rape, murder, or ethnic cleansing. Yet in light of the role of Facebook in enabling the Rohingya genocide, taking preventative action to defang the aspects of its algorithm that embolden extremism and inspire violence constitutes a crucial step in ensuring such atrocities do not repeat themselves.

---

39. Christopher Mims, "How Facebook's Master Algorithm Powers the Social Network," *The Wall Street Journal*, October 22, 2017, https://www.proquest.com/newspapers/how-facebooks-master-algorithm-powers-social/docview/1953638742/se-2?accountid=12205.

40. Jeff Horowitz, interview with Ryan Knutson and Keach Hagey, *The Facebook Files, Part 4: The Outrage Algorithm,* podcast audio, September 18, 2021, https://www.wsj.com/podcasts/the-journal/the-facebook-files-part-4-the-outrage-algorithm/e619fbb7-43b0-485b-877f-18a98ffa773f.

In response to the release of the Facebook Files—a series of damning revelations about the platform exposed by an internal whistleblower—Facebook has emphasized its commitment to safety and security. Spokespeople have touted the billions of dollars and thousands of new personnel that the company has devoted to the assurance of its users' well-being. After all, as Mark Zuckerberg said, permitting the proliferation of harmful content is "deeply illogical;" what advertiser would pay to post ads on a platform that sparks hatred and violence?[41] Though Zuckerberg's argument appears rational, it stands in direct opposition to reality. Since 2017, the year the Rohingya genocide peaked, Facebook's annual net income has grown by more than $13 billion and its userbase has expanded to encompass 3 billion people—more than 1/3 of the world's population.[42] This success comes despite the company's failure to address the structural issues that encourage hate speech and enable genocide.

Even seemingly small changes to Facebook's underlying technology can yield potentially consequential impacts. Internal company studies have found that slightly reducing platform speeds gives users a split second longer to think about what they're posting, encouraging them to exercise greater prudence. One surprising and controversial way to do this, the research team discovered, is to remove the reshare button. They found that doing so led to "huge gains off the bat," but Facebook would never condone such a radical step, even one proven to significantly mitigate harm on its platform.[43] However, less extreme measures can also produce similar results. Other potential solutions include limiting how many invitations a group can send or

41. Mike Isaac, "Facebook Wrestles with the Features It Used to Define Social Networking," *The New York Times*, October 9, 2021, https://www.nytimes.com/2021/10/25/technology/facebook-like-share-buttons.html?referringSource=articleShare.

42. Statista, "Facebook's Revenue and Net Income from 2007 to 2020," accessed November 6, 2021, https://www.statista.com/statistics/277229/facebooks-annual-revenue-and-net-income/.

43. Horowitz, *The Outrage Algorithm*.

comments an individual can post in a day. These schemes are known within Facebook as "break glass measures"—ideas the company knows about, suspects could work, and could deploy at any time, but refuses to enact.[44]

This represents the greatest shortcoming of harm mitigation efforts that depend upon corporate responsibility; Facebook has repeatedly demonstrated its unwillingness to take inconsequential hits to its revenue even when doing so would increase the safety and wellbeing of its users. Its pattern of behavior suggests a pervasive devaluation of human rights. A change in policy and practice, therefore, may require a change in the corporate culture, which in turn may require a change in the highest echelons of Facebook's leadership. Removing Mark Zuckerberg and installing a new CEO could reprioritize people over profit.

Separating the inventor from his invention, however, may prove difficult. Former Facebook chief security officer Alex Stamos notes that the company's top leadership has remained essentially the same since it went public in 2012, and the calcification of turnover "[creates] kind of this bubble where Zuckerberg gets to be detached."[45] As a Harvard dropout with a brilliant technological mind and limited life experience, Zuckerberg surrounded himself with veterans of marketing, management, and lobbying. Many of these veterans continue to comprise Facebook's elite inner circle. They tend to shield Zuckerberg from the harsh realities of his platform, a dynamic which does not seem to perturb Zuckerberg himself. "He's okay being in this bubble of people who are telling him, you know, not necessarily what he wants to hear, but

---

44. Ibid.

45. Jon Favreau, interview with Alex Stamos, *Offline: Alex Stamos on Leaving Facebook and Zuckerberg's Reign,* podcast audio, January 9, 2022, https://crooked.com/podcast/alex-stamos-on-leaving-facebook-and-zuckerbergs-reign/.

they are formatting things in the way he wants them to be," Stamos contends.[46] Shaking up Facebook's executive roster may overhaul its culture more broadly, reprioritizing human rights even if it means slight losses to their growth rate.

Further, a change in leadership could serve as a means of redesigning the corporate structure. "There's a couple of…fundamental organizational flaws at Facebook," Stamos observes, "that I think are real problems."[47] Human rights and information security concerns fall under the purview of the communications and policy teams, which have historically erred on the side of less transparency. Stamos recommends that Facebook divorce its integrity and security departments from comms and policy in order to preserve an independent human rights approach. If unmired from the swamp of public relations and obsessive growth, such an approach could actually prioritize the safety and wellbeing of the platform's users.

Of course, there is no guarantee that installing a new CEO would tangibly reform Facebook's culture or its policy. Zuckerberg's successor could prove even less invested in preserving human rights and may worsen rather than ameliorate violations on the platform. Even if they possessed a commitment to expanding the company's focus beyond growth and revenue production, a top-down transformation may not prove effective. The corporate culture may be too deeply ingrained for a mere change in leadership to alter it. At the very least, an executive shake-up would require either the acquiescence of Zuckerberg himself or the exhortation of Facebook's board of directors. Both seem unlikely. Though enhanced privacy requirements from smartphone companies—not to mention the rocky debut of Meta as Facebook's futuristic parent company—have sent the company's stock into a tailspin, investors are snapping up shares,

46. Ibid.

47. Ibid.

betting on significant long-term growth.[48] With that kind of market optimism, a replacement of the CEO seems unlikely. Unless and until Facebook faces enduring legal or financial consequences for its actions, internal corporate responsibility measures will remain weak and performative.

48. Ryan Vlastelica, "Meta's Stock-Market Wipeout Is Unmatched in the Megacap Era," *Bloomberg,* February 18, 2022, https://www.bloomberg.com/news/articles/2022-02-18/meta-s-collapse-is-unmatched-in-the-era-of-big-tech-tech-watch; Fitri Wulandari, "Meta Platforms Stock Forecast: Will Metaverse Drive FB Higher?" *Capital.com,* March 18, 2022, https://capital.com/meta-platforms-fb-stock-forecast.

CHAPTER THREE BIBLIOGRAPHY

Alvarez, José E. "Are Corporations 'Subjects' of International Law?" *Santa Clara Journal of International Law* 9, no. 1 (2011): 1—36. https://digitalcommons.law.scu.edu/scujil/vol9/iss1/1/.

Amnesty International. "The Facebook Papers: What Do They Mean from a Human Rights Perspective?" November 4, 2021, https://www.amnesty.org/en/latest/campaigns/2021/11/the-facebook-papers-what-do-they-mean-from-a-human-rights-perspective/.

Barry, Eloise. "These Are the Countries Where Twitter, Facebook and TikTok Are Banned." *Time*, January 18, 2022, https://time.com/6139988/countries-where-twitter-facebook-tiktok-banned/.

Chirot, Daniel and Clark McCauley. *Why Not Kill Them All? The Logic and Prevention of Mass Political Murder*. Princeton: Princeton University Press, 2006.

Congress.gov. "S.314 – 104th Congress (1995-1996): Communications Decency Act of 1995." February 1, 1995. https://www.congress.gov/bill/104th-congress/senate-bill/314/.

Debre, Isabel and Fares Akram. "Facebook's Language Gaps Weaken Screening of Hate, Terrorism." *Associated Press,* October 25, 2021, https://apnews.com/article/the-facebook-papers-language-moderation-problems-392cb2d065f81980713f37384d07e61f.

Electronic Frontier Foundation. "CDA 230." Accessed February 14, 2022, https://www.eff.org/issues/cda230.

European Commission. "European Union Code of Conduct on Countering Illegal Hate Speech Online." June 30, 2016, https://ec.europa.eu/info/policies/justice-and-fundamental-

rights/combatting-discrimination/racism-and-xenophobia/eu-code-conduct-countering-illegal-hate-speech-online_en.

European Commission. "The Digital Markets Act: Ensuring Fair and Open Digital Markets." Accessed December 1, 2021, https://ec.europa.eu/info/strategy/priorities-2019-2024/europe-fit-digital-age/digital-markets-act-ensuring-fair-and-open-digital-markets_en.

European Council. *Council Framework Decision on Combating Certain Forms and Expressions of Racism and Xenophobia by Means of Criminal Law,* November 28, 2008, 2008/913/JHA, https://eur-lex.europa.eu/legal-content/en/TXT/?uri=CELEX%3A32008F0913.

European Parliament. "Digital Services Act: Regulating Platforms for a Safer Online Space for Users." News release, January 20, 2022, https://www.europarl.europa.eu/news/en/press-room/20220114IPR21017/digital-services-act-regulating-platforms-for-a-safer-online-space-for-users.

Facebook. "Hate Speech." Accessed May 22, 2021, https://www.facebook.com/communitystandards/hate_speech.

Favreau, Jon. Interview with Alex Stamos. *Offline: Alex Stamos on Leaving Facebook and Zuckerberg's Reign.* Podcast audio. January 9, 2022. https://crooked.com/podcast/alex-stamos-on-leaving-facebook-and-zuckerbergs-reign/.

Federal Trade Commission. "FTC Sues Facebook for Illegal Monopolization." News release, December 9, 2020, https://www.ftc.gov/news-events/news/press-releases/2020/12/ftc-sues-facebook-illegal-monopolization.

Frenkel, Sheera and Davey Alba. “In India, Facebook Grapples with an Amplified Version of Its Problems.” *The New York Times*, November 9, 2021. https://www.nytimes.com/2021/10/23/technology/facebook-india-misinformation.html?referringSource=articleShare.

Hannum, Hurst. “The UDHR in National and International Law.” *Health and Human Rights* 3, no. 2 (1998): 144–58. https://doi.org/10.2307/4065305.

“Hate Speech on Facebook Poses ‘Acute Challenges to Human Dignity’ – UN Expert.” *UN News,* December 23, 2020, https://news.un.org/en/story/2020/12/1080832.

Horowitz, Jeff. Interview with Ryan Knutson and Keach Hagey. *The Facebook Files, Part 4: The Outrage Algorithm.* Podcast audio. September 18, 2021. https://www.wsj.com/podcasts/the-journal/the-facebook-files-part-4-the-outrage-algorithm/e619fbb7-43b0-485b-877f-18a98ffa773f.

Isaac, Mike. “Facebook Wrestles with the Features It Used to Define Social Networking.” *The New York Times*. October 9, 2021. https://www.nytimes.com/2021/10/25/technology/facebook-like-share-buttons.html?referringSource=articleShare.

Kang, Cecilia. “Lawmakers, Taking Aim at Big Tech, Push Sweeping Overhaul of Antitrust.” *The New York Times*, June 11, 2021, https://www.nytimes.com/2021/06/11/technology/big-tech-antitrust-bills.html#:~:text=WASHINGTON%20%E2%80%94%20House%20lawmakers%20on%20Friday,to%20monopoly%20laws%20in%20decades.

“Mark Zuckerberg’s Letter to Investors: ‘The Hacker Way.’” *Wired*, February 1, 2012, https://www.wired.com/2012/02/zuck-letter/.

McGregor, Richard. "How the State Runs Business in China." *The Guardian,* July 25, 2019, https://www.theguardian.com/world/2019/jul/25/china-business-xi-jinping-communist-party-state-private-enterprise-huawei.

Mims, Christopher. "How Facebook's Master Algorithm Powers the Social Network; The Algorithm Behind Facebook's News Feed, a 'Modular Layered Cake, Extracts Meaning from Every Post and Photo.' *The Wall Street Journal*, October 22, 2017.

Office of the High Commissioner of Human Rights. *Guiding Principles on Business and Human Rights*, 2011, HR/PUB/11/04, https://www.ohchr.org/sites/default/files/documents/publications/guidingprinciplesbusinesshr_en.pdf.

Office of the High Commissioner of Human Rights. *Legally Binding Instrument to Regulate, in International Human Rights Law, the Activities of Transnational Corporations and Other Business Enterprises*, August 17, 2021, A/HRC/49/65/Add. 1, https://www.ohchr.org/sites/default/files/Documents/HRBodies/HRCouncil/WGTransCorp/Session6/LBI3rdDRAFT.pdf.

Office of the High Commissioner of Human Rights. *Report on Online Hate Speech*, October 9, 2019, A/74/486, https://documents-dds-ny.un.org/doc/UNDOC/GEN/N19/308/13/PDF/N1930813.pdf?OpenElement.

Perrigo, Billy. "Facebook Says It's Removing More Hate Speech Than Ever Before. But There's a Catch." *Time,* November 27, 2019. https://time.com/5739688/facebook-hate-speech-languages/.

Statista. "Facebook's Revenue and Net Income from 2007 to 2020." Accessed November 6, 2021. https://www.statista.com/statistics/277229/facebooks-annual-revenue-and-net-income/.

Stecklow, Steve. "Why Facebook is Losing the War on Hate Speech in Myanmar." *Reuters,* August 15, 2018. https://www.reuters.com/investigates/special-report/myanmar-facebook-hate/.

United Nations. *Charter of the United Nations,* October 24, 1945, 1 UNTS XVI, https://www.un.org/en/about-us/un-charter/full-text.

United Nations. *Universal Declaration of Human Rights*, December 10, 1948, 217 A (III), https://www.un.org/en/about-us/universal-declaration-of-human-rights.

Vlastelica, Ryan. "Meta's Stock-Market Wipeout Is Unmatched in the Megacap Era." *Bloomberg,* February 18, 2022, https://www.bloomberg.com/news/articles/2022-02-18/meta-s-collapse-is-unmatched-in-the-era-of-big-tech-tech-watch.

Whitten-Woodring, Jennifer, Kleinberg, Mona S., Thawnghmung, Ardeth, and Thitsar, Myat The. "Poison If You Don't Know How to Use It: Facebook, Democracy, and Human Rights in Myanmar." *The International Journal of Press/Politics* 25, no. 3 (July 2020): 407–25. https://doi.org/10.1177/1940161220919666.

Wulandari, Fitri. "Meta Platforms Stock Forecast: Will Metaverse Drive FB Higher?" *Capital.com,* March 18, 2022, https://capital.com/meta-platforms-fb-stock-forecast.

CONCLUSION

Call it ingenuity, call it hubris, call it human nature—whichever your preferred explanation, the world has achieved a breathtaking stage of technological advancement. Innovation has always intended to ease and improve our lives, but its purpose now seems almost soteriological; we need invention to soothe our imperfect world, we need progress to save us from ourselves. It was not so long ago that Facebook felt that way. I remember when friends and family began making accounts and we realized that it enabled us to connect with each other despite miles between us. On a dauntingly expansive planet, social media makes our loved ones feel close. Eventually, Facebook came to represent hope for a liberated world order as activists took to cyberspace to dissolve calcified hierarchies, empower historically marginalized groups, and build a transnational vision of a just and equitable society. It seemed the technology would once again offer a means of salvation.

Yet in recent years, it has become clear that while Facebook may serve as a tool for humanity to fight its inner demons, it also actively exploits those demons for profit. Zuckerberg and other company executives introduced a product that preyed upon human tendencies towards negativity, divisiveness, and radicalism. Worse, they did so in volatile sociopolitical environments with little knowledge nor care for local contexts of power and oppression. Facebook posits that it simply provides a neutral platform for individuals to freely express themselves, but such claims disingenuously mask the way that the platform's algorithm intentionally amplifies sensational, even inflammatory content. It does not create social tensions out of nothing, but it can exacerbate those that already exist. This process was tragically exemplified by the Burmese military's Facebook hate speech campaign and the resulting Rohingya genocide. Despite the well-founded connection between the company's technology,

policies, and resource prioritization and the unspeakable violence in Myanmar, there has been little concerted global, national, or corporate effort to address the issue of hate speech on social media.

Over the two-year journey of writing this thesis, I have reflected often upon my personal relationship with Facebook. I still do not have an account on that platform nor any of its affiliates, in large part due to moral reservations that only intensified as I delved into my research. But my stubborn refusal to join my 3.7 billion fellow human beings on social media does not force Facebook to improve its dismal human rights record; indeed, it has often distanced me from the movements fighting to make the world—both virtual and physical—a safer place for those most harmed by digital hate speech. Even if I succumb neither to convenience nor friendly coercion and go the remainder of my life without a Facebook account, social media is now inextricably intertwined with the future of human civilization. We felt this viscerally on October 4, 2021. On that day, the world ground to a halt, half the global population reeling from an unforeseeable shockwave. Communications quieted and commerce stuttered as Facebook and its associated platforms—WhatsApp, Instagram, and Messenger—disappeared for five hours.[1] In this brief time, the company lost $50 billion in market value and caused incalculable damage to small businesses that depend on social media for sales and communities that depend on social media as the Internet.[2] The outage sparked conversation about the world's potentially dangerous

---

1. Mike Isaac and Sheera Frenkel, "Gone in Minutes, Out for Hours: Outage Shakes Facebook," *The New York Times*, October 8, 2021, https://www.nytimes.com/2021/10/04/technology/facebook-down.html.

2. Mark Sweney, "Facebook Outage Highlights Global Over-reliance on its Services," *The Guardian,* October 5, 2021, https://www.theguardian.com/technology/2021/oct/05/facebook-outage-highlights-global-over-reliance-on-its-services#:~:text=The%20fallout%20of%20Facebook's%20unprecedented,m%20of%20the%20advertising%20dollars.

social, commercial, and psychological reliance on Facebook. However, it also demonstrated beyond doubt that social media in general—and Facebook in particular—is here to stay.

Its permanence does not give it a free pass to violate human rights. Quite the opposite: it makes it all the more critical to maximize Facebook's potential to help while minimizing its tendency to harm. Doing so will require updating both legal and ethical frameworks to align with a deeply technologized and interconnected world. Zuckerberg and his fellow executives conceived of the social media network in a milieu of unregulated free speech and capitalism that fostered breakneck innovations and the absolute valuation of individual rights. These qualities are seen as fundamental features of Facebook. In many ways, they enabled the platform's meteoric rise to social success and economic dominance. However, Facebook's single-minded quest for connection actually sows division, encouraging people to bond over hatred and prejudice, and ultimately amplifying discord and difference.

After its highly publicized rebranding in 2021, Facebook—now known as 'Meta,'—incorporated language pledging itself to values of equity, justice, and service. It vows to "give people a voice," to "serve everyone, build connection and community, and keep people safe" as it carries out its mission to "bring the world closer together."[3] In order to accomplish this, the company has adopted the slogan: "move fast with stable infrastructure."[4] Of course, this somewhat uninspiring motto harkens back to the early years of Facebook, when Zuckerberg famously urged his employees to "move fast and break things"—break the status quo, break

---

3. "Company Info," *Meta,* accessed March 19, 2022, https://about.facebook.com/company-info/.

4. Isabel Asher Hamilton, "Mark Zuckerberg's New Values for Meta Show He Still Hasn't Truly Let Go of 'Move Fast and Break Things,'" *Business Insider,* February 16, 2022, https://www.businessinsider.com/meta-mark-zuckerberg-new-values-move-fast-and-break-things-2022-2.

expectations, break the bounds of what we are told is possible. Ironically, putting things back together again may require returning to that old mantra. To repair the damage caused by hate speech and other human rights violations on Facebook, the world must once more break things—break up monopolies, break with legal precedent, break racialized systems of power. Though it seems paradoxical, the decentralization and democratization that social media offers would empower such a movement; holding platforms accountable for their misdeeds may prove impossible without the platforms themselves.

In fact, some have already tried that strategy. Over 1000 companies and activist groups united in July of 2020 under the hashtag #StopHateforProfit and boycotted placing their ads on Facebook. The campaign intended to "pressure Facebook into taking more stringent steps to stop the spread of hate speech and misinformation on its platform," drawing major corporate names including Coca-Cola, Verizon, and HP.[5] Their unprecedented stance grabbed international headlines. However, rampant speculation that the boycott would significantly damage Facebook's earnings and force it to strengthen its human rights policy did not bear out. Most companies returned to the platform within a month, and an analysis following the campaign found that lost brand revenue represented less than 1% of third quarter growth.[6] Despite a concerted effort by some of the world's most influential businesses, no substantive policy change emerged; participants in the boycott were unable or unwilling to truly divorce themselves from social media advertisements, and the public has largely forgotten about #StopHateForProfit.

---

5. Megan Graham, "Zuckerberg was Right: Ad Boycotts Won't Hurt Facebook That Much," *CNBC,* August 4, 2020, https://www.cnbc.com/2020/08/04/some-major-companies-will-keep-pausing-facebook-ads-as-boycott-ends.html.

6. Ibid.

Clearly, Facebook has grown too powerful to be financially coerced into better behavior. The solution may rest upon some of the very qualities that enable the platform's human rights violations, namely its unmatched ability to share sensational content across physical and social barriers. For years, activists have turned to social media networks to raise awareness for civil rights campaigns, foment resistance against authoritarian regimes, and demand justice for historically oppressed peoples. Despite my conscious disinvolvement in Facebook, I am familiar with these tactics. In the brief time that I have been writing this thesis, I have witnessed countless transformative protest movements originate, expand, and organize on social media. Native activists in the United States—particularly my home in Minnesota—have used Facebook and its kin to bring the world's attention to missing and murdered Indigenous women. Their Facebook page has almost 150,000 followers, a huge step in combating the apathy that allows such atrocities to persist.[7] The decentralized nature of social media facilitates coordination between the national nonprofit and aligned grassroots movements, building a diverse coalition that does not depend on a single individual or hierarchy to ensure its continued existence. Instead, the movement grows from countless roots and nodes. Missing and Murdered Indigenous Women's social media organizing encourages the sharing of strategies, networks, and radical imaginings.

This principle is exemplified by the Hong Kong protests of 2019—2020. After China passed an aggressive extradition bill undermining the 'one country, two systems' arrangement that preserved Hong Kong's nominal sovereignty, thousands of people took to the streets. For months, protesters staged audacious actions, garnering international support. Drawing inspiration from the Arab Spring, the protest "exhibited certain features of modern, decentralized,

7. Missing and Murdered Indigenous Women USA (@mmiwusa). Facebook page, accessed April 4, 2022. https://www.facebook.com/mmiwusa/.

'networked' social movements: the refusal of leadership, the decentralization of protest activities, and the instrumental reliance on social media-based communication."[8] Just as Missing and Murdered Indigenous Women have done, the Hong Kong activists rejected hierarchical organization in favor of "participatory horizontality" that leveraged spontaneous and decentralized decision-making to maximize action impact while minimizing risks to privacy and security.[9] Movements were coordinated, tactics negotiated, and messages communicated via social media. In a battle for freedom against a regime notorious for its omniscient surveillance and harsh crackdowns, the ability of platforms like Facebook to foster both anonymity and unity is critical. It seems unlikely that the Hong Kong protests would have proven so effective without the use of digital social networks.

Social media not only enabled the sharing of tactics and rhetoric within the activist networks in Hong Kong, but to other movements across the world. As I write this from the Twin Cities, my community continues to see regular uprisings against racist policing, driven by the murders of Philando Castile, Jamar Clark, George Floyd, Daunte Wright, and so many others. Many Black Lives Matter organizers adopt strategies directly from their counterparts in Hong Kong. Scholars classify this phenomenon as "mimetic piggybacking," observing that shared videos and images serve as informal 'how-to' guides for protesting unjust or authoritarian systems.[10] Though social justice activists have long drawn inspiration from their predecessors, "the social internet has sped up a long history of direct and indirect dialogue between protest

8. Silvia Frosina, "Digital Revolution: How Social Media Shaped the 2019 Hong Kong Protests," *Italian Institute for International Political Studies,* June 9, 2021, https://www.ispionline.it/en/pubblicazione/digital-revolution-how-social-media-shaped-2019-hong-kong-protests-30756.

9. Ibid.

10. Tracy Ma, Natalie Shutler, Jonah E. Bromwich, and Shane O'Neill, "Why Protest Tactics Spread Like Memes," *The New York Times,* July 31, 2020, https://www.nytimes.com/2020/07/31/style/viral-protest-videos.html.

movements around the world."[11] I have seen countless examples of this with my own eyes: protesters covering tear gas canisters with traffic cones, batting away flash-bang grenades with tennis rackets, and organizing defensive walls of umbrellas to ward off projectiles. Thanks to social media, such tactics have been used from Hong Kong to Minneapolis, Catalonia to Kurdistan.

Despite the intentional ways that Facebook's algorithm and executives have encouraged the proliferation of racialized hate speech—and, by extension, racialized violence—the platform nevertheless serves as a powerful tool for demanding justice. And many of the same qualities that enable unspeakable atrocities also empower inspiring resistance. The company's uncompromising commitment to free speech permits harm to marginalized groups, yet prevents government censorship of popular opinion. Anonymity afforded by digital avatars licenses impunity both for those seeking to oppress others, and those seeking to liberate them. Transcendence of geographic and temporal boundaries creates space for simmering social tensions to boil over, and for burgeoning social movements to spread. Holding Facebook accountable for the physical repercussions of its digital content is of paramount importance, and ironically Facebook itself may represent the best tool to do so. In order to fight for transformative change both online and off, we must indeed move fast and break things.

11. Ibid.

CONCLUSION BIBLIOGRAPHY

"Company Info." *Meta,* accessed March 19, 2022. https://about.facebook.com/company-info/.

Graham, Megan. "Zuckerberg was Right: Ad Boycotts Won't Hurt Facebook That Much." *CNBC,* August 4, 2020. https://www.cnbc.com/2020/08/04/some-major-companies-will-keep-pausing-facebook-ads-as-boycott-ends.html.

Frosina, Silvia. "Digital Revolution: How Social Media Shaped the 2019 Hong Kong Protests." *Italian Institute for International Political Studies,* June 9, 2021. https://www.ispionline.it/en/pubblicazione/digital-revolution-how-social-media-shaped-2019-hong-kong-protests-30756.

Hamilton, Isabel Asher. "Mark Zuckerberg's New Values for Meta Show He Still Hasn't Truly Let Go of 'Move Fast and Break Things.'" *Business Insider,* February 16, 2022. https://www.businessinsider.com/meta-mark-zuckerberg-new-values-move-fast-and-break-things-2022-2.

Isaac, Mike and Sheera Frenkel. "Gone in Minutes, Out for Hours: Outage Shakes Facebook." *The New York Times*, October 8, 2021. https://www.nytimes.com/2021/10/04/technology/facebook-down.html.

Ma, Tracy, Natalie Shutler, Jonah E. Bromwich, and Shane O'Neill. "Why Protest Tactics Spread Like Memes." *The New York Times,* July 31, 2020. https://www.nytimes.com/2020/07/31/style/viral-protest-videos.html.

Missing and Murdered Indigenous Women USA (@mmiwusa). Facebook page, accessed April 4, 2022. https://www.facebook.com/mmiwusa/.

Sweney, Mark. "Facebook Outage Highlights Global Over-reliance on its Services." *The Guardian,* October 5, 2021.

https://www.theguardian.com/technology/2021/oct/05/facebook-outage-highlights-global-over-reliance-on-its-services#:~:text=The%20fallout%20of%20Facebook's%20unprecedented,m%20of%20the%20advertising%20dollars.