Beyond AI: Responses to Hate Speech and Disinformation

Jessica Young[1], **Preetha Swamy**[1] **& David Danks**[2,3]

[1]-Heinz College of Public Policy; [2]-Department of Philosophy; [3]-Department of Psychology
Carnegie Mellon University
5000 Forbes Ave., Pittsburgh, PA 15213 USA
preethas@andrew.cmu.edu/jessicayoung@cmu.edu ddanks@cmu.edu

Abstract

Recent years have seen numerous and rampant instances of disinformation and hate speech across technology platforms such as Facebook, Twitter, Reddit, WhatsApp, and YouTube. Moreover, some of these instances have spilled from the digital to the physical world, resulting in significant harms to people, even including death. At the same time, there is little consensus, or understanding, about possible best (or merely better) responses. However, one consistent theme is that AI will somehow resolve this problem (though with disagreement about how and when). We contend that a major impediment to progress on this challenge has been a lack of understanding about the common key features across these instances of (physical) harm-causing hate speech and disinformation. We examined a series of case studies, such as recent events in Myanmar, India, and the United States, to identify common themes. In particular, they all involve: efforts to normalize violence; exploitation of existing fears; implicit or explicit support from government and key third parties; viral transmission via social media; and lack of source transparency. Many of these challenges fall outside of the scope of AI methods as primarily used, and so such methods will be relatively limited in their ability to help. However, policy actions informed by these case studies, behavioral science, and the current capabilities of AI could mitigate some of the harms done by hate speech and disinformation.

Introduction

Over the past few years, disinformation and hate speech spread online have led to rampant physical harm offline across geographies and technology platforms. As a society, we have historically held that members should be free from libel (Malice Standard) and from hate speech likely to lead to lawless action (Brandenburg Standard). With our time increasingly spent in the digital realm, online environments should uphold and protect those standards, as we do not give up our social contracts when we enter an online domain. Technology platforms have tried to support those standards by adopting new policy guidelines, leveraging emerging technologies, and onboarding staff trained to detect and remove harmful content. In particular, AI technologies have repeatedly been put forward as a solution to these problems, whether through detection, removal, or mitigation. In fact, the scope of the problem—the sheer volume of hate speech and disinformation in online environments—seemingly *requires* an automated or autonomous solution, as no human could ever hope to keep up.

However, such efforts—whether based in AI, policy, or humans—have fallen woefully short, with numerous tragic cases where people have been psychologically or physically harmed as a result of hate speech or disinformation. Harmful speech continues to be an elusive problem with impediments from timely detection to verification of falsity to political will. As time progresses without remedy, we see two problematic impacts quickly growing: instances of real harm mount offline, while online community trust wanes and becomes harder to maintain.

One persistent challenge in efforts to address online hate speech and disinformation in a systematic manner is a lack of understanding of the common themes of the relevant cases. Leading research to date has largely focused on detecting "fake news" and disinformation at large (Kucharski, 2016). Little research has explored disinformation linked to violence, and when it has, it has focused on such cases in geographic isolation (Marwick and Lewis, 2017). We thus build on prior research by analyzing violence-inducing cases on an international scale to identify their commonalities. These commonalities help to explain the difficulty of developing purely AI solutions. Drawing lessons from law, media, and behavioral science/psychology, we then consider alternative policies and practices that technology platforms should adopt to manage content that one can reasonably presume will lead to physical harms.

Case Studies & Commonalities

In this paper, we focus on instances of hate speech and disinformation that plausibly or foreseeably could result in physical harms. We recognize that these acts can also lead to psychological or social harms, but the nature and extent of that harm is significantly harder to detect, judge, and

measure. We thus focus on the easier (relatively speaking) case of physical harms, leaving non-physical harms to future work. Even with this restriction, there is potential ambiguity about our uses of the terms 'hate speech' and 'disinformation'. These terms have subtly different meanings in different contexts, but we can use relatively expansive notions for the present paper. More specifically, we define 'hate speech' as "… a communication that carries no meaning other than the expression of hatred for some group, especially in circumstances in which the communication is likely to provoke violence" (US Legal, 2018). Specifying a particular group is key, whether based on gender, religion, or some other factor. 'Disinformation' represents "false information deliberately and often covertly spread…to influence public opinion or obscure the truth" (Merriam-Webster, 2018).

Given these understandings, we attempted to identify key cases in which there was online hate speech or disinformation that could reasonably be foreseen to produce physical harms. We used two main criteria to narrow our scope. First, we considered only cases in which there was a clear causal pathway by which the speech led to physical harms. Second, we focused on cases in which technology companies might hope to create policies or use their systems to curb instances of hate speech and disinformation. Our principal concern here is the possibility of AI technologies mitigating or minimizing these harmful events.

High-level Descriptions of Case Studies

Pizzagate in the United States: On October 29, 2016, during the politically tense U.S. presidential election season, Facebook user Carmen Katz shared a post claiming presidential candidate Hillary Clinton was involved in a pedophilia ring. The story went viral, spreading to Twitter, Reddit and YouTube, where Alex Jones seeded disinformation that the ring operated out of a Washington, D.C. restaurant, Comet Ping Pong. Convinced of the veracity of this story, on December 4th, Edgar Welch, a father of two, entered the restaurant armed with an AR-15, a .38 handgun, and a folding knife, searching for signs of children in danger. After firing three shots, he was arrested (Robb, 2017).

Child snatchers in India: In India, disinformation took root during a time when abduction of children and child brides was rising (Sengupta and Bassi, 2018). Messages about individuals abducting children in villages spread across the country playing into an apprehension of outsiders. Fear-inciting messages containing disinformation spread virally via WhatsApp resulting in mob violence. In two months of 2018 alone, more than 20 innocent victims were lynched by mobs as large as 2,000 people (Busby, 2018). A total of 69 mob-violence cases in India have been related to rumors of kidnapping with 77% "attributed to fake news spread through social media," including 28% of cases where WhatsApp was the rumor source (Saldanha, et al., 2018).

Rohingya violence in Myanmar: Existing tensions in Myanmar between the ethnic minority Rohingya, and the majority Buddhists, were reflected in social media posts on Facebook as early as 2013 (Miles, 2018). As the conflict escalated from 2017-2018, it did so in concert with orchestrated disinformation campaigns and hateful posts about the Rohingya. This content was spread primarily on Facebook and Facebook Messenger, but also through Twitter (Miles, 2018). Disinformation spread about mosques stock-piling weapons to destroy Buddhist pagodas (Safi, 2018), Rohingya burning their own homes (Specia and Mozur, 2017), and a "fabricated jihad" planned for September 2017 (McKirdy, 2018). Some of the most popular posts received 3,400 reactions or were shared up to 9,500 times (Rajagopalan, Vo, and Soe, 2018). According to the U.N., the diffusion of such content on Facebook directly contributed to genocide in which 25,000 Rohingya were killed and an additional 700,000 displaced (Stecklow, 2018).

In all three cases, regardless of the country or society, hate speech and disinformation led to disastrous consequences. The scale of these cases ranges from no one being hurt to potentially 25,000 individuals dead. A significant thread running through all of these cases is fear. A group's fear can manifest through systemic conditions such as cultural tensions, attacks on vulnerable populations such as children, or instilling values that help those in power or hurt those disenfranchised.

Common Themes

We divide the common factors in these case studies into *active* events or conditions (e.g., actions) and *background* conditions (e.g., technology infrastructure).

Active: Normalization of violence

In both Myanmar and India, abusive content created a narrative under which violence appeared to be legitimized and normalized. In Myanmar, anti-Rohingya content "exploded on Facebook at the start of the Rohingya crisis" (Safi, 2018) including "racist political cartoons, falsified images, [and] staged news reports" (Huish and Balazo, 2018). As this disinformation went viral, it was spread in concert with hate speech (Specia and Mozur, 2017; Safi, 2018). This hate speech served to dehumanize the Rohingya and shape public perception so that the resulting violence would be celebrated (Huish and Balazo, 2018). Hateful posts inciting violence both before and after state-led violence (Rajagopalan, et al., 2018; Stecklow, 2018) helped justify the acts of physical harm as normal responses. Similarly, in India, doctored videos were used to justify violence as a normal response to heinous acts by others. Depictions of mob violence as a legitimate response to suspected kidnappings (Bassi and Sengupta, 2018) normalized the use of violence.

Active: Leveraging existing fears and societal divisions
In all three case studies, social media users posted manipulative content that played into existing community tensions, societal divisions, and fears. In Myanmar, the ethnic-minority Rohingya were portrayed as "aggressive outsiders" (Specia and Mozur, 2017) who were procreating to outnumber Buddhists. In India, narratives were framed in insider-outsider terms, whether village natives vs. outsiders, Muslims vs. Hindus, or hetero vs. transgender communities (Saldanha, et al., 2018). In the US, partisan distrust of political leaders was leveraged to drive fear and outrage over pedophilia (Robb, 2017). A focus on fear and division helped increase the sharing of hate speech and disinformation (Kraemer, Donsbach, Heidenreich, and Goutheir, 2014).

Active: Complicit or ineffectual powerful groups
In Myanmar, government and military social media accounts spread disinformation (Specia and Mozur, 2017), posted hateful content, and in some instances called for violence (Rajagopalan, et al., 2018). For example, "10% of the Arakan National Party's posts had Facebook-defined hate speech." In the US, foreign agent and domestic political operatives' accounts shared disinformation about Pizzagate, thereby signaling that political elites endorsed the content (Robb, 2017). Finally, messages in India were largely sent by the educated elite, who endorsed manipulative content to the detriment of those lacking media literacy (Bengali and Parth, 2018).

In all three cases, the relevant technology companies were ineffectual in stopping the hate speech and disinformation. Users flagged abusive content to limited avail. In Myanmar, Facebook took more than 48 hours to address flagged content, rather than the claimed average six-hour response (Rajagopalan, et al., 2018). Some posts violating community standards remained on Facebook for six years (Stecklow, 2018). In India, abusive content was largely shared on WhatsApp and could not be investigated or addressed due to encryption. WhatsApp was left to place advertisements on media literacy which unsurprisingly failed to curtail the violence (Saldanha, et al., 2018). Likewise in the US, the owner of Comet Ping Pong asked Facebook and Twitter to remove abusive posts to no avail, leading him to contact the FBI upon receiving death threats (LibGuides, 2018). In all three cases, the burden of addressing these problems fell instead to third parties, such as NGOs (Rajagopalan, et al., 2018) or the police (Bassi and Sengupta, 2018).

Background: Virality
In all three cases, platform features enabled content to go viral. That is, the platform provided a key background condition for the wide spread of the hate speech and disinformation. In Myanmar, anti-Rohingya posts were much more popular on Facebook than non-Rohingya ones, and so could spread rapidly. In October 2017, there were nearly 500 anti-Rohingya posts per hour (McKirdy, 2018). In India, the message forwarding feature of WhatsApp permitted chain-messages to reach virality. And in the US, the Pizzagate story spread rapidly on Twitter and YouTube. In fact, Google Trends research shows that between the story's debut on Facebook and its InfoWars coverage on YouTube, Google searches for "Hillary" and "pedophile" went from 0 to 100 (on a 0-100 scale) in only four days (Robb, 2017).

Moreover, an over-reliance on social media platforms for news contributed to the rapid spread of the content. In Myanmar, Facebook is the only online news source for the majority of the country (Safi, 2018). In India, many of WhatsApp's 200 million active users (Bengali and Parth, 2018) view the platform as its primary source of news information (Bassi and Sengupta, 2018). And even though a multitude of media sources exist in the US, 68% of Americans consume news through social media (Matsa and Shearer, 2018), thereby amplifying the reach of stories like Pizzagate.

Background: Lack of source transparency
Finally, all three of these case studies were marked by an inability for people to determine the original source of the (dis)information. Content sharing (whether on Facebook Messenger, WhatsApp, or Twitter) can obscure a post's origin. Indeed, with content forwarding on WhatsApp, a post's source is intentionally anonymous. In theory, one can still find the original source on other platforms, but that becomes exceedingly unlikely when messages are shared or forwarded thousands of times. Similarly, when (dis)information is spread by bots (as in Pizzagate), the original content source can quickly become untraceable when accounts are deactivated.

The Case for Tech Platform Responsibility

Although these events are widely agreed to be horrific, there is less agreement that technology companies bear responsibility. If a company or organization builds some infrastructure, then they are not thereby automatically responsible for all harms that involve that infrastructure. We contend, however, that the terms and conditions of each platform is appropriately similar to a social contract. The company determines the rules and protocols that each individual should uphold on a social media platform (D'Agostino, Gaus, and Thrasher, 2017). As part of these terms and conditions, the companies assert the right, and thereby the responsibility, to sanction individuals who fail to adhere to that social contract. Hence, when the company fails to exercise this role, they can be held responsible for resulting harms. This responsibility is especially apparent in Myanmar, as numerous unsuccessful attempts were made to flag and remove violating content (Rajagopalan, et al., 2018).

Moreover, some technology companies actively exacerbate the situation with the curation algorithms they use.

These algorithms shape what their users see, whether through active preference elicitation or more passive measures such as tracking the amount of time spent on a post or news article (Kim, 2018). This active curation can lead to a one-sided view of a situation or condition. And as individuals turn to social media as their main news source, disinformation and hate speech can be given disproportionate weight or trust. Even the term 'newsfeed' implies that the content is informative or newsworthy. Infrastructure companies are not necessarily culpable for harms performed with their systems. However, the active choices of technology companies mean that they cannot dodge responsibility in these cases.

Ineffective Responses to Date

We earlier noted the ineffectual responses of technology companies during the case study crises. Though progress on some fronts has been made since, company responses to date have still been insufficient for curbing the problem. In December 2017, Twitter adopted a new set of Hateful Conduct Policies. Meanwhile in April 2018, Facebook publicly disclosed its Community Standards Enforcement Guidelines, including rules for removing dehumanizing and inflammatory content. In addition, in response to an EU Commission report on disinformation, Google, Facebook and Twitter have committed to working with independent researchers to analyze the diffusion of disinformation, as well as develop new features like fact check mechanisms (Mantzarlis, 2018).

Facebook's primary strategy, used during our case studies, was to rely on reports from its users which it then manually assessed (Specia and Mozur, 2017). Unsurprisingly, this process was ineffective due to a multitude of issues. Facebook receives millions of reports every week (Stecklow, 2018), and so they have grown content moderation teams, pledging to double safety and security staff to 20,000 employees in 2018 (Frier, 2017). In Myanmar, this expansion included an increase from two to 100 Myanmar language experts. However, even that number is arguably inadequate to review posts coming from the country's 18 million active users. Activists and technologists have complained of clunky user interfaces that might not even display a country's local language. In fact, content that fully violates community standards, eliciting violence and even genocide, has remained online for months (Stecklow, 2018). Such slow response times have had grave consequences on the ground, to the point where governments (in India, Myanmar, and Sri Lanka) have had to shut down the internet during peak violence (Heanue, 2018; Safi, 2018).

Instead of removal, some platforms have sought to limit the spread of hate speech and disinformation. Both Google and Facebook have tweaked their algorithms to downgrade disinformation in newsfeeds (BBC, 2017; Ingram, 2017). Facebook and Twitter have begun de-activating the accounts of their most inflammatory offenders. Our case studies reveal challenges inherent in this after-the-fact approach. Blocked users migrate to other platforms or simply create new accounts. For example, when Myanmar's government banned Ashin Wirathu from public preaching, he migrated to Facebook and found a larger following for his hate speech (Specia and Mozur, 2017). Likewise, in the US, when Reddit banned a Pizzagate subreddit, users went to Twitter and sent approximately 145,000 tweets in one day (Robb, 2017).

In response to these challenges, platform leaders have touted the (potential) power of AI methods to better detect abusive content at scale (Rajagopalan, et al., 2018). However, current AI capabilities fall short in many key areas, including context recognition, code-switching, and use of novel slang. AI for inflammatory language detection largely failed in Myanmar due to present AI's limits in detecting context, which was particularly problematic since actors found ways to avoid detection by including messages in photos, memes and videos (Rajagopalan, et al., 2018).

Lessons from Diverse Fields

Given the failings of AI responses, we might hope to find insight from fields such as law, media, and behavioral sciences. They can help us understand why responses to date have not worked, and also what elements will be needed for successful solutions.

The Law and its Limits

In the US, the First Amendment protects individual speech from government censorship unless online hate speech enters the terrain of "incitement to imminent lawless action" (*Brandenburg v. Ohio*) or "true threats" (reasonably understood as such) about an individual or group (*Virginia v. Black*). While both of these conditions could be used to hold *individuals* accountable, that legal responsibility does not extend to technology platforms. To date, courts have ruled that social media companies have limited responsibilities as platforms rather than media publishers (*Fields v. Twitter*) in line with US legislation "asserting that platforms cannot be liable for content users post on their sites" (47 USC § 230).

Matters are different elsewhere in the world. Most notably, in Germany, the Network Enforcement Act requires large social media platforms to remove "illegal content" (including threats of violence) within 24 hours, or face fines up to 50 million euro (Miller, 2018). The German law places the burden on companies to determine which content violates the law, thereby giving them incentive to err on the side of removal. Other countries are quickly following suit, including Croatia, India, Indonesia, Malaysia, and Singapore (Funke, 2018). These local legal shifts may have global impact if companies decide to adopt "one size fits all" practices

(analogous to many responses to GDPR). Nonetheless, most social media platforms do not currently have significant legal incentive to change their practices.

Media Studies

As numerous adults consume social media as their prime news source, it is important to identify practices that can be utilized against hate speech and disinformation. To start, platforms could use algorithms to flag disinformation from a bot rather than a person (West, 2017). Content from a bot arguably does not have the same trustworthiness as from a person, so it should be flagged. In addition, third parties, particularly journalists, use sites like Knowhere to actively rate the level of disinformation from a news source so the user can decide whether to engage with that content (Houser, 2018).

Another mitigation effort aims to educate readers on how to identify hate speech or disinformation through social media literacy (West, 2017). For example, Twitter has a media literacy hashtag, primarily for children, to help draw attention to these educational campaigns (Twitter, 2018). Relatedly, the GDPR aims to promote terms and conditions that are easy to decipher by the user, thereby improving people's ability to know which content violates community standards (Baird, 2017).

Behavioral Science and Psychology

Cognitive biases played a key role in the rapid spread of disinformation and hate speech leading to physical harm. As a result of confirmation bias, people search for information confirming what they already believe, and cling to such beliefs with greater conviction when another party attempts to refute them. Moreover, it is much harder to discount the credibility of some (dis)information after one has internalized it. We see these challenges in all three case studies. In Myanmar, the public continued to believe disinformation even after the United Nations and western nations provided evidence of genocide (Beech and Saw, 2018). Meanwhile, in the Indian state of Tripura, a local activist who tried to debunk misinformation in local villages was killed by angry mobs (Bassi and Sengupta, 2018). In Pizzagate, a debunking article actually caused a *surge* in Twitter traffic supporting the conspiracy theory (Robb, 2017).

Hate speech can significantly increase the probability of violence by changing sentiment (Kramer, Guillory, and Hancock, 2014), fueling fear and anger, and scrambling right and wrong. It can even change perceptions of social norms. For example, anti-refugee Facebook content predicts a significant increase in anti-refugee violence, even when other factors are included (Müller and Schwarz, 2018).

Given people's cognitive biases, the most effective strategy (on psychological grounds) may well be simply to limit exposure in the first place, rather than attempting to change opinions afterwards. At the least, we should aim to remove such content before its sentiment becomes normalized, particularly given the known addictive properties of social media (Blackwell, et al., 2017).

Proposed Recommendations & Solutions

Although this is a challenging problem, we suggest that there are viable options worth exploring. In particular, the case studies, lessons learned, and ideas from other disciplines all point towards steps that can be taken. To mitigate hate speech and disinformation that plausibly cause real-world physical harms, we propose three key approaches. First, we can use existing AI methods to effectively reduce virality and increase transparency to contain the spread and reach of hate speech and disinformation. Second, we can try to reduce the impact of the harmful content that is still displayed, partly (but not entirely) using AI. Third, we can better engage the user community in addressing these challenges without any AI at all.

Detect and Limit with AI

AI can be a useful tool, but we propose that it be used quite differently than how it is largely employed. Using AI, technology platforms have frequently focused on automatic detection and removal of problematic speech. (Of course, human flagging is also quite prevalent.) However, as explained above, limitations of AI preclude (at the current time) fully autonomous solutions to hate speech and disinformation. In contrast, we contend that AI should be used to target the background condition of virality. If we can thwart the spread of the most dangerous content before irreversible damage is done, then we need not solve the hard problem of detection. We can use AI to identify posts approaching virality, and then proactively investigate these outliers with humans, thereby working to block the transmission of posts having the most potential to create harm at scale.[1] This strategy would not attempt to remove or censor all harmful posts, but rather allow platforms to focus on posts with the potential to cause the greatest harm. Of course, we should continue to pursue improved AIs to detect problematic content, but this strategy provides a good solution while we pursue the best.

Additionally, AI can be used to detect content that is contributed by bots, not users, as the former have frequently contributed to the viral spread of misinformation. As seen

[1] We acknowledge that Facebook has recently announced a shift toward this approach for handling misinformation at large (Thompson, 2018). We hold this should be a strategy employed by all major technology platforms facing this issue, with a priority focus on targeting the most (physically) harmful misinformation: disinformation and hate speech.

with the Pizzagate case, bots worldwide were working to get the Pizzagate hashtag trending and thereby provide legitimacy to the disinformation (Robb, 2017). Current AIs are easily capable of detecting posting rates and patterns of automated services, and so AI can help determine posts that are being amplified through bots.

Relatedly, we can improve the inputs to AI detection algorithms by incorporating commonly recurring features of our case studies. For example, those who share manipulative content to incite violence often use content that triggers primal reactions such as fear and anger, appeal to existing power and influence, and repeat hateful or false narratives to normalize violence. We should thus incorporate the following variables into our AI detection and spread-prediction algorithms: sentiment analysis; number of followers; history of problematic content; and other measures of authenticity.[2] As terms associated with hate speech evolve, they should continually be incorporated into algorithms, while enhanced lexicons (incorporating contextual variations of words) should continue to be developed with NGO partners worldwide. And of course, these development processes should include robust testing and monitoring to facilitate continual system improvement.

Signal Questionable Content

Given the difficulty of changing one's opinion once confirmation bias has set in, platforms could better leverage cues to signal the veracity of content to users. This could be done by signaling the likely validity of the source, as well as the validity of post content. First, leveraging best practices from the media, platforms should maintain and display a post's original author, preserving this source information through the sharing chain. AI could be used to mine posts and detect if the original account had since been deactivated, providing the opportunity for platforms to update information about the source. Using this type of signaling, platforms could further indicate if content was likely spread by a bot network rather than by a person.

Second, platforms could signal content (if likely or confirmed to be disinformation) as manipulative, much like traditional media, social media, and search algorithms all signal manipulative content with "Ad" labels. Factors that could be used in weighing content validity could include, among others, use of links (already flagged as abusive or not), source (if known to host problematic content or calls to violence), and account history (counts of prior removed content).

Improve Communication and Feedback Loops

In addition to better leveraging current AI capabilities and signaling, improvements in user design should also be made. As platforms are largely dependent on user reports of inappropriate content, measures to facilitate and expedite this process would leverage one of the greatest resources platforms have available: billions of active users. Taking one platform as an example, when attempting to report content on Facebook, a user is currently presented with 10 possible options for flagging content (including suicide, hate speech, and false information). However, there is no clear channel indicating where this information goes, or how it is managed after the content is flagged. Enhanced channels of communication that detail where and how reported information will be processed, as well as the status of processing, would create valuable transparency for community members who are also trying to minimize harm and violence. For processing reports, at a base level, platforms should disclose content removal speeds and live up to relevant community standards. Finally, an opportunity exists to reward users who report content that was verified and removed as a result. In this way, platforms could incentive beneficial actions while growing/maintaining members of the community who help ensure its quality and safety.

Conclusion

Given the lethal consequences that have resulted from online hate speech and disinformation, it is imperative that technology companies adopt better solutions that address the recurring themes underlying the most egregious cases. Recent events illustrate the continued problem of online hate speech and disinformation. For example, we observe that, the Tree of Life Synagogue attack in Pittsburgh involved four of the five common threads identified in this paper. Legal frameworks are not equipped to address this issue with speed, nor are they set up to proactively prevent harm. Rather, they only reactively address grievances. Furthermore, as the law varies by locality and new laws regarding hate speech and disinformation are still emerging, companies should continue to develop their own standards and solutions as the law evolves.

In the meantime, solutions need to be developed with urgency. When the most harmful posts cause damage, we have seen government shut-downs of entire internet infrastructure, showing communities around the world are grappling with a problem that tech platforms are uniquely equipped to address. While regulatory bodies can prohibit the harmful actions and speech inciting harm, only platforms can manipulate the algorithms that lend such content power and

[2] Of course, we recognize that social media companies might already be doing so, but their algorithms are not publicly disclosed.

credibility through its reach, exposure, and focal placement. The above solutions have the potential to help technology platforms target even more of the core elements that enable harmful content to thrive.

Acknowledgments

[removed for anonymous review]

References

Baird, S. 2017. GDPR Matchup: The Health Insurance Portability and Accountability Act. *iapp Privacy Tracker*, May 17, 2017.

Bassi, S. and Sengupta, J. 2018. WhatsApp Cracks Down on Fake Content after Child-kidnap Rumours Spark Killings across India. *CBC News*, July 8, 2018.

BBC. 2017. Google Search Changes Tackle Fake News and Hate Speech. *BBC News*, April 25, 2017.

Beech, H. and Saw, N. 2018. In Myanmar, a Facebook Blackout Brings More Anger Than a Genocide Charge. *The New York Times*, August 31, 2018.

Bengali, S., and Parth, M.N. 2018 Rumors of Child-kidnapping Gangs and Other WhatsApp Hoaxes Are Getting People Killed in India. *Los Angeles Times*, May 30, 2018.

Blackwell, D., Leaman, C., Tramposch, R., Osborne, C. and Liss, M. 2017. Extraversion, Neuroticism, Attachment Style and Fear of Missing Out as Predictors of Social Media Use and Addiction. *Personality and Individual Differences*, *116*: 69-72.

Brandenburg v. Ohio, 395 U.S. 444 (1968).

Busby, M. 2018. 2,000-strong Mob Lynch Engineer over Child-kidnapping Rumours in India. *The Independent*, July 16, 2018.

D'Agostino, F., Gaus, G., and Thrasher, J. 2017. Contemporary Approaches to the Social Contract. In *Stanford Encyclopedia of Philosophy*, ed. E. N. Zalta.

Feingold, R., et al., 2017. Fake News and Misinformation: The Roles of the Nation's Digital Newsstands, Facebook, Google, Twitter, and Reddit. October 1, 2018.

Fields v. Twitter, Inc., 2018 WL 626800 (9th Cir. Jan. 31, 2018)

Frier, S. 2017. Facebook Says It Will Double Safety and Security Staff to 20,000. *Bloomberg News*, October 31, 2017.

Funke, D. 2018. A Guide to Anti-misinformation Actions around the World. *Poynter*, October 31, 2018.

Heanue, S. 2018. WhatsApp Video of Fake Kidnapping Sparks Hysteria in India as Mob Beats Men to Death. *ABC News*, July 24, 2018.

Houser, K. 2018. A New AI "Journalist" Is Rewriting the News to Remove Bias. *Futurism*, April 6, 2018.

Huish, R. and Balazo, P. 2018. Unliked: How Facebook Is Playing a Part in the Rohingya Genocide. *The Conversation*, Oct. 24, 2018.

Human Rights Watch. 2018. *Germany: Flawed Social Media Law*. Human Rights Watch. February 15, 2018.

Ingram, D. 2017. Facebook Changes Algorithm to Curb 'Tiny Group' of Spammers. *Reuters*, June 30, 2017.

Kim, S. A. 2018. Social Media Algorithms: Why You See What You See. *Georgetown Law Technology Review*. Sept. 14, 2018.

Kraemer, T., Donsbach, J., Heidenreich, S., and Goutheir, M. HJ. 2014. The Good, the Bad, and the Ugly–How Emotions Affect Online Customer Engagement Behavior. Technical report.

Kramer, A.D., Guillory, J.E. and Hancock, J.T., 2014. Experimental Evidence of Massive-scale Emotional Contagion through Social Networks. *Proceedings of the National Academy of Sciences*, p. 201320040.

Kucharski, A. 2016. Study Epidemiology of Fake News. *Nature*, 540: 525.

LibGuides. 2018. Evaluating News: Case Study: Pizzagate. *LibGuides*, August 16. 2018.

Mantzarlis, A. 2018. Six Key Points from the EU Commissions New Report on Disinformation. *Poynter*, March 15, 2018.

Marwick, A. and Lewis, R. 2017. Media Manipulation and Disinformation Online. *Data & Society Research Institute*, May 15, 2017.

Matsa, K.E. and Shearer, E. 2018. News Use Across Social Media Platforms 2018. *Pew Research Center*, September 10, 2018.

McKirdy, E. 2018. When Facebook Becomes 'the Beast': Myanmar Activists Say Social Media Aids Genocide. *CNN*, April 7, 2018.

Merriam-Webster. 2018. *Disinformation*.

Miles, T. 2018. U.N. Investigators Cite Facebook Role in Myanmar Crisis. *Reuters*, March 12, 2018.

Miller, J. 2017. Germany Votes for 50m Euro Social Media Fines. *BBC News*. June 30, 2017.

Müller, K. and Schwarz, C. 2018. Fanning the Flames of Hate: Social Media and Hate Crime. *SSRN*, May 21, 2018.

Rajagopalan, M., Vo, L.T., and Soe, A.N. 2018. How Facebook Failed The Rohingya In Myanmar. *BuzzFeed News*, Aug. 27, 2018.

Robb, A. 2017. Pizzagate: Anatomy of a Fake News Scandal. *Rolling Stone*, November 16, 2017.

Safi, M. 2018. Revealed: Facebook Hate Speech Exploded in Myanmar during Rohingya Crisis. *The Guardian*, April 3, 2018.

Saldanha, A., Rajput, P., and Hazare, J. 2018. Child-Lifting Rumours: 33 Killed In 69 Mob Attacks Since Jan 2017. Before That Only 1 Attack In 2012. *IndiaSpend*, August 8, 2018.

Specia, M., and Mozur, P. 2017. A War of Words Puts Facebook at the Center of Myanmar's Rohingya Crisis. *New York Times*, October 27, 2017.

Stecklow, S. 2018. Why Facebook Is Losing the War on Hate Speech in Myanmar. *Reuters*, August 15, 2018.

Taub, A., and Fisher, M. 2018. Where Countries are Tinderboxes and Facebook is a Match. *The New York Times*, April 21, 2018.

Thompson, N. 2018. How Facebook Checks Facts and Polices Hate Speech. *Wired*, July 06, 2018.

Twitter. 2018. #medialiteracy Hashtag on Twitter. *Twitter*.

US Legal, Inc. 2018. Hate Speech Law and Legal Definition. *USLegal, Inc*.

Virginia v. Black, 538 U.S. 343 (2003)

West, D.M. 2017. How to Combat Fake News and Disinformation." *Brookings Institution*. December 18, 2017.