4

Online Hate Speech

Alexandra A. Siegel

INTRODUCTION

Once relegated to the dark corners of the Internet, online hate speech has become increasingly visible on mainstream social media platforms. From targeted anti-Semitic attacks on Jewish journalists to reports of social media's role in mobilizing ethnic violence in Myanmar and Sri Lanka, the offline consequences of online hate speech appear increasingly dire. Fearing that this harmful rhetoric is inciting violence and driving extremism, governments worldwide are passing regulation and pressuring social media companies to implement policies to stop the spread of online hate speech (Gagliardone et al. 2016).

However, these calls for action have rarely been motivated by comprehensive empirical evidence. Moreover, despite increased attention to online hate speech in the scientific literature, surprisingly little is known about the prevalence, causes, or consequences of different forms of harmful language across diverse platforms. Furthermore, researchers have only recently begun to examine the efficacy of approaches to countering online hate, and our understanding of the collateral costs of these interventions is especially limited.

This chapter examines the state of the literature – including scientific research, legal scholarship, and policy reports – on online hate speech. In particular, it explores ongoing debates and limitations in current approaches to defining and detecting online hate speech; provides an overview of what social media data and surveys can tell us about the producers, targets, and overall prevalence of harmful language; reviews empirical evidence of the offline consequences of online hate speech; and offers quantitative insights into what interventions might be most effective in combating harmful rhetoric online.

DEFINING ONLINE HATE SPEECH

There is no single agreed on definition of hate speech – online or offline – and the topic has been hotly debated by academics, legal experts, and policymakers alike. Most commonly, hate speech is understood to be bias-motivated, hostile, and malicious language targeted at a person or group because of their actual or perceived innate characteristics (Cohen-Almagor 2011; Faris et al. 2016). However, as Sellars (2016) argues, "for all of the extensive literature about the causes, harms, and responses to hate speech, few scholars have endeavored to systematically define the term."

A wide variety of content might or might not fit a definition of hate speech, depending on the context (Parekh et al. 2012; Sellars 2016). For example, while slurs and insults are easily identifiable, language containing epithets may not necessarily be considered hate speech by the speaker or target recipient (Delgado 1982). Conversely, more subtle language attacking an out-group, which can be harder for casual observers to identify, may have particularly damaging effects on individuals and group relations (Parekh et al. 2012). This is especially true in the online sphere, where speech is rapidly evolving and can be highly specialized (Gagliardone et al. 2015). The use of code words as stand-ins for racial slurs is also common in online communities, further complicating the definition of hate speech (Duarte et al. 2018). For example, among members of the alt-right, journalists have documented the use of the term "googles" to refer to the n-word; "skypes" as an anti-Semitic slur; "yahoos" as a derogatory term for Hispanics; and "skittles" as an anti-Muslim term (Sonnad 2016). Alt-right communities have also used steganography, such as triple brackets, to identify and harass Jews online (Fleishman and Smith 2016). In this way, when defining hate speech – and online hate speech in particular – the well-known "I know it when I see it" classification famously applied to obscene content clearly falls short.

As a result, existing definitions of hate speech can be extremely broad or fairly narrow. At one end of the spectrum are definitions that capture a wide variety of speech that is directed against a specified or easily identifiable individual or group based on arbitrary or normatively irrelevant features (Parekh et al. 2012). At the other end are definitions that require intended harm. The narrowest definitions imply that hate speech must be "dangerous speech" – language that is directly linked to the incitement of mass violence or physical harm against an out-group (Benesch 2013). This tension reflects the difficulty of developing a definition that adequately addresses the range of phenomena that might be considered hate speech, without losing valuable distinctions. Online hate speech can involve disparate instigators, targets, motives, and tactics. Sometimes perpetrators know those they attack, whereas others may galvanize anonymous online followers to target particular individuals. Speech that incites violence is distinct from speech that is "merely" offensive, and the use of harmful language by a single attacker is quite different from coordinated hate campaigns carried out by a digital mob (Sellars 2016). Recent work seeks to develop more comprehensive definitions and coding schemes for identifying hate speech that provide context and account for differences in severity and intent (Gagliardone et al. 2016;

Waseem and Hovy 2016; Kennedy et al. 2018; Olteanu et al. 2018). Yet despite these advances, there is still no consensus in the scientific literature on how to define online hate speech.

Legal definitions of hate speech are similarly murky. Governments are increasingly defining hate speech in their criminal codes in an attempt to directly regulate harmful rhetoric both on- and offline (Haraszti 2012). As with academic definitions, these range from the relatively broad, such as Canada's characterization of hate speech as language that "willfully promotes hatred against any identifiable group," to more narrow definitions, like the European Union's framework, which defines hate speech as: "Public incitement to violence or hatred directed against a group of persons or a member of such group defined on the basis of race, [color], descent, religion or belief, or national or ethnic origin" and "publicly condoning, denying or grossly trivializing crimes of genocide, crimes against humanity, and war crimes [as defined in EU law], when the conduct is carried out in a manner likely to incite violence or hatred against such group or a member of such group" (Sellars 2016). In the Uniited Kingdom, it is a criminal offense to incite racial or religious hatred, and variations on this legislation – while unconstitutional in the United States – exist in the majority of developed democracies, including Australia, Denmark, France, Germany, India, South Africa, Sweden, and New Zealand (Howard 2019), and in authoritarian contexts, particularly in the Arab World where laws banning online hate speech are often lumped together with laws countering extremism (Chetty and Alathur 2018). Yet despite the existence of laws explicitly banning hate speech, how these laws should be enforced in practice, particularly in the digital age, is a subject of ongoing debate.

More recently, online platforms themselves have developed definitions of hate speech for the purpose of moderating user-generated content. For example, YouTube's Community Guidelines "hateful content" section states "we don't support content that promotes or condones violence against individuals or groups based on race or ethnic origin, religion, disability, gender, age, nationality, veteran status, or sexual orientation/gender identity, or whose primary purpose is inciting hatred on the basis of these core characteristics" (YouTube 2018). Similarly, Twitter's terms of service state that the company prohibits "hateful conduct" including "promot[ing] violence against or directly attack[ing] or threaten[ing] other people on the basis of race, ethnicity, national origin, sexual orientation, gender, gender identity, religious affiliation, age, disability or disease." The company also emphasizes that it does not allow accounts whose "primary purpose is inciting harm towards others on the basis of these categories" (Twitter 2018). Facebook's definition of hate speech does not contain the incitement to violence language employed by Twitter and YouTube, instead identifying hate speech as "content that directly attacks people based on their race; ethnicity; national origin; religious affiliation; sexual orientation; sex, gender, or gender identity; or serious disabilities or diseases" (Facebook 2018).

Together, this absence of clear and consistent definitions of hate speech in academic research, legal scholarship, and among actors attempting to govern online spaces has meant that despite extensive research, and well-documented policy interventions, our knowledge of the causes, consequences, and effective means of combating online hate speech remains somewhat clouded by definitional ambiguity.

DETECTING ONLINE HATE SPEECH

Just as there is no clear consensus on the definition of hate speech, there is no consensus with regard to the most effective way to detect it across diverse platforms. The majority of automated approaches to identifying hate speech begin with a binary classification task in which researchers are concerned with coding a document as "hate speech or not," though multiclass approaches have also been used (Davidson et al. 2017).

Automated hate speech detection tends to rely on natural language processing or text mining strategies (Fortuna and Nunes 2018). The simplest of these approaches are dictionary-based methods, which involve developing a list of words that are searched and counted in a text. Dictionary-based approaches generally use content words – including insults and slurs – to identify hate speech (Dinakar et al. 2011; Dadvar et al. 2012; Liu and Forss 2015; Isbister et al. 2018). These methods can also involve normalizing or taking the total number of words in each text into consideration (Dadvar et al. 2012). Recognizing that online hate speech may obscure offensive words using accidental or intentional misspellings, some researchers have used distance metrics, such as the minimum number of edits necessary to transform one term into another, to augment their dictionary-based methods (Warner and Hirschberg 2012). Furthermore, given that code words may be used to avoid detection of hateful terms, other researchers have included known anti–out-group code words in their dictionaries (Magu et al. 2017).

Beyond pure dictionary-based methods, most state-of-the-art hate speech detection techniques involve supervised text classification tasks. These approaches, such as using Naive Bayes classifiers, linear support vector machines (SVM), decision trees, or random forest models, often rely on "bag-of-words" and "n-gram" techniques. In the bag-of-words method, a corpus is created based on words that appear in a training dataset, instead of a predefined dictionary. The frequencies of words appearing in text, which has been manually annotated as "hate speech or not," are then used as features to train a classifier (Greevy and Smeaton 2004; Kwok and Wang 2013; Burnap and Williams 2016). To avoid misclassification, if words are used in different contexts or spelled incorrectly, some researchers use n-grams, a similar approach to bag-of-words, which combines sequential words into bigrams, trigrams, or lists of length "n" (Burnap and Williams 2016; Waseem and Hovy 2016; Badjatiya et al. 2017; Davidson et al. 2017). More recent work has leveraged these approaches to improve the accuracy of dictionary-based methods – removing false positives by identifying which tweets containing slurs should indeed be classified as hate speech (Siegel et al. 2020). Rule-based approaches and theme-based grammatical patterns, which incorporate sentence structure, have also been used (Fortuna and Nunes 2018).

Other researchers have identified hate speech using topic modeling, aiming to identify posts belonging to a defined topic such as race or religion (Agarwal and Sureka 2017). Still others have incorporated sentiment into their analysis, with the assumption that hate speech is likely to be negative in tone (Liu and Forss 2014; Gitari et al. 2015; Davidson et al. 2017; Del Vigna et al. 2017). Word embedding or vector representations of text techniques including doc2vec, paragraph2vec, and FastText have also been used (Djuric et al. 2015; Schmidt and Wiegand 2017; Siegel et al. 2020), and deep learning techniques employing neural networks have become more common for both text classification and sentiment analysis related to detecting hate speech (Yuan et al. 2016; Zhang et al. 2018, Al-Makhadmeh and Tolba 2020).

Recognizing that these techniques may not be well-suited to identifying subtle or indirect forms of online hate, researchers have also employed more theoretically motivated approaches. For example, Burnap and Williams (2016) and ElSherief, Kulkarni et al. (2018) incorporate the concept of othering or "us vs. them" language into their measure of hate speech. They find that hate speech often uses third-person pronouns, including expressions like "send them all home." Other studies have incorporated declarations of in-group superiority – in addition to attacks directed at out-groups – into their measures (Warner and Hirschberg 2012). Another approach involves accounting for common anti–out-group stereotypes. For example, anti-Hispanic speech might make reference to border crossing, or anti-Semitic language might refer to banking, money, or the media (Alorainy et al. 2018). Additional work has distinguished between hate speech directed at a group (generalized hate speech) and hate speech directed at individuals (directed hate speech) to capture important nuances in the targets of online hate speech (ElSherief, Kulkarni et al. 2018). Beyond relying on textual features, researchers have also incorporated user characteristics, including network features and friend/follower counts to improve the accuracy of hate speech detection (Unsvåg and Gambäck 2018).

Another more recent set of approaches leverages large pre-classified datasets from online platforms to detect online hate speech. These include the bag-of-communities technique (Chandrasekharan, Samory et al. 2017), which computes the similarity of a post to the language used in nine other known hateful communities from 4chan, Reddit, Voat, and MetaFilter. Similar techniques have been employed by Saleem et al. (2017) and Siegel et al. (2020), using data from well-known hateful subreddits to classify hate speech on Twitter. An advantage of these methods is that they are not hindered by low intercoder reliability that can be found in training datasets or by the fact that rapidly evolving speech patterns online can make it difficult to use the same training data over time (Waseem 2016).

Despite these major advances in the automatic detection of online hate speech, existing methods largely have not been tested across multiple platforms or diverse types of hate speech. Owing to ease of data collection, most existing studies have relied on Twitter data. While other works have incorporated data from Reddit, YouTube, Facebook, Whisper, Tumblr, Myspace, Gab, the comment sections of websites, and blogs, these are relatively rare (Fortuna and Nunes 2018; Mathew, Dutt et al. 2019). Additionally, the vast majority of studies examine English-language content, though some researchers have developed methods to detect hate speech in other languages. These include empirical examinations of hate speech in Amharic (Mossie and Wang 2018), Arabic (Siegel 2015; De Smedt et al. 2018; Siegel et al. 2018, Albadi et al. 2019, Chowdhury et al. 2019), Dutch (Van Hee et al. 2015), German (Ross et al. 2017), Hindi (Santosh and Aravind 2019), Indonesian (Aulia and Budi 2019), Italian (Lingiardi et al. 2019), Korean (Kang et al. 2018), Polish (Czapla et al. 2019), Romanian (Meza 2016), and Spanish (Basile et al. 2019). Crowd-sourced multilingual dictionaries of online hate speech including Hatebase, the Racial Slur Database, and HateTrack have also been developed, demonstrating promising avenues for future work (ElSherief, Kulkarni et al. 2018, Siapera et al. 2018). Yet approaches to automated hate speech detection that are designed to scale across multiple languages are quite difficult to develop, and more work is needed in this area.

Additionally, the majority of studies of online hate speech seek to detect all types of hate speech at once, or "general hate speech" (Fortuna and Nunes 2018). However, other works have examined specific types of harmful language, including jihadist hate speech (De Smedt et al. 2018), sectarian hate speech (Siegel 2015; Siegel et al. 2018), anti-Muslim hate speech (Olteanu et al. 2018), anti-black hate speech (Kwok and Wang 2013), misogynistic hate speech (Citron 2011), and anti-immigrant hate speech (Ross et al. 2017). Recent work has also explored differences in types of hate speech, comparing hate speech targeting diverse out-groups and distinguishing between more and less severe types of hate speech (Beauchamp et al. 2018; Saha et al. 2019; Siegel et al. 2019).

PRODUCERS OF ONLINE HATE SPEECH

While extensive research has explored organized hate groups' use of online hate speech, less is known about the actors in informal communities dedicated to producing harmful content, or the accounts that produce hate speech on mainstream platforms. Moreover, no empirical work has systematically examined how these actors interact within and across platforms.

Organized hate groups established an online presence shortly after the invention of the Internet (Bowman-Grieve 2009) and have proliferated over time. More than a decade of primarily qualitative research has demonstrated that organized hate groups use the Internet to disseminate hate speech on their official websites (Adams and Roscigno 2005; Chau and Xu 2007; Douglas 2007; Flores-Yeffal et al. 2011; Castle 2012; Parenti 2013). This includes the use of interactive forums (Holtz and Wagner 2009) such as chat boards and video games (Selepak 2010). Hate groups use these channels both to broaden their reach and to target specific audiences. For example, the explicitly racist video games that originate on far-right extremist websites are designed to appeal to ardent supporters and potential members alike, especially youth audiences (Selepak 2010). Along these lines, hate groups have used the Internet to recruit new members and reinforce group identity (Chau and Xu 2007; Parenti 2013; Weaver 2013). Online platforms are also especially well-suited to tailoring messages to specific groups or individuals (Castle 2012). By providing efficient ways to reach new audiences and disseminate hateful language, the Internet enables hate groups to be well represented in the digital realm, fostering a sense of community among their members, and attracting the attention of journalists and everyday citizens alike (Bowman-Grieve 2009; McNamee et al. 2010).

In addition to the official websites of organized hate groups, the number of sites dedicated to producing hateful content operated by informal groups and individuals has also increased over time (Potok 2015). These include explicitly racist, misogynistic, or otherwise discriminatory pages, channels, or communities on mainstream social networking platforms like Facebook, Twitter, and YouTube, as well as forums on Reddit 4chan, and 8chan, listserves, internet chat communities, discussion forums, and blogs designed to disseminate hateful rhetoric (Douglas 2007; Marwick 2017). These range from fake Facebook profiles designed to incite violence against minorities (Farkas and Neumayer 2017) to infamous (now banned) Reddit forums like /CoonTown and /fatpeoplehate (Chandrasekharan, Pavalanathan et al. 2017). Well-known white nationalists and hateful accounts have also operated openly on mainstream social media platforms. For example, Richard Spencer, who organized the "Unite the Right" alt-right Charlottesville rally, has more than 75,000 followers and was verified by Twitter up until November 2017, when he was stripped of his verified status. Twitter accounts such as @SageGang and @WhiteGenocide frequently tweet violent racist and anti-Semitic language (Daniels 2017).

However, such concentrations of hateful content are sometimes banned and removed from particular platforms. As a result, these communities often disappear and resurface in new forms. For example, in 2011, 4chan's founder deleted the news board (/n/) due to racist comments and created /pol/ as a replacement forum for political discussion.

4chan's /pol/ board quickly became a home for particularly hateful speech – even by 4chan standards (Hine et al. 2016). Similarly, banned subreddits like Coontown have moved to Voat, a platform with no regulations with regard to hate speech (Chandrasekharan, Pavalanathan et al. 2017). While survey data and ethnographic work suggests that users of 4chan and Redditt are overwhelmingly young, white, and male (Daniels 2017; Costello and Hawdon 2018), because of the anonymous nature of these sites we do not know very much about the users that produce the most hate speech. In particular, we do not know the degree to which their rhetoric represents their actual beliefs or is simply trolling or attention-seeking behavior, which is quite common in these communities (Phillips 2015).

Outside of these official and unofficial pages and forums dedicated to hateful content, hate speech is also prevalent in general online discussions across a variety of popular platforms, including Facebook, YouTube, Myspace, Tumblr, Whisper, and Yik Yak (Black et al. 2016; Fortuna and Nunes 2018). While little is known about the specific individuals that produce hate speech on these mainstream platforms, recent work has begun to measure and characterize their behavior. Examining the trajectory of producers of hate speech over time, Beauchamp et al. (2018) find that producers of misogynistic and racist hate speech on Twitter tend to start out expressing "softer," more indirect hateful language and later graduate to producing more virulent hate. The authors posit that this may be due to gradually decreasing levels of social stigma as these users find themselves in increasingly extreme social networks. ElSherief, Nilizadeh et al. (2018) find on Twitter that accounts that instigate hate speech tend to be new, very active, and express lower emotional awareness and higher anger and immoderation in the content of their tweets, compared to other Twitter users who did not produce such content. Similarly, using a manually annotated dataset of about 5,000 "hateful users," Ribeiro et al. (2018) find that hateful users tweet more frequently, follow more people each day, and their accounts are more short-lived and recent. They also find that, although hateful users tend to have fewer followers, they are very densely connected in retweet networks. Hateful users are seventy-one times more likely to retweet other hateful users and suspended users are eleven times more likely to retweet other suspended users, compared to non-hateful users. Comparing users that produce hate speech to those that do not on Gab, Mathew, Dutt et al. (2019) also find that hateful users are densely connected to one another. As a result, they argue, content generated by hateful users tends to spread faster, farther, and reach a much wider audience as compared to the content generated by users that do not produce hate speech.

Such behavior may contribute to the overall visibility of hate speech on mainstream online platforms. For example, on Twitter, although tweets containing hate speech have lower numbers of replies and likes than non-hateful tweets, they contain a similar number of retweets (Klubicka and Fernandez 2018). The highly networked structure of hateful Twitter users also dovetails with qualitative evidence suggesting that people are mobilized on explicitly hateful subreddits or communities like the /pol/ board on 4chan to engage in coordinated racist or sexist attacks on Twitter (Daniels 2017).

Studying the network structure of users who produce online hate speech, Magdy et al. (2016) find that they can predict the likelihood that Twitter users tweet anti-Muslim messages after the 2015 Paris attacks with high levels of precision and accuracy based on their Twitter networks, even if they have never mentioned Muslims or Islam in their previous tweets. Twitter users who follow conservative media outlets, Republican primary candidates, evangelical Christian preachers, and accounts discussing foreign policy issues were significantly more likely to tweet anti-Muslim content following the Paris attacks than those that did not.

In one of the few existing surveys of social media users exploring the use of hate speech, Costello and Hawdon (2018) find that people who spend more time on Reddit and Tumblr report disseminating more hate speech online. Moreover, individuals who are close to an online community, or spend more time in communities where hate speech is common, are more inclined to produce hate material. Counter to their expectations, they find that spending more time online in general, however, is not associated with the production of hate and there is no association between the use of first-person shooter games and producing hate material online.

As with pages and forums explicitly dedicated to online hate speech, individual producers of online hate speech have increasingly been banned from Twitter and other mainstream platforms. While many of these users simply create new accounts after they have been suspended, others have moved to more specialized platforms where they can produce hate more freely. For example, in August 2016, the social network Gab was created as an alternative to Twitter. The platform stated that it was dedicated to "people and free speech first," courting users banned or suspended from other social networks (Marwick and Lewis 2017). Zannettou et al. (2018) found that Gab is mainly used for the dissemination and discussion of news and world events, and that it predominantly attracts alt-right users, conspiracy theorists, and trolls. The authors find that hate speech is much more common on Gab than Twitter but less common on Gab than on 4chan's /pol/ board. Similarly, Lima et al. (2018) found that Gab generally hosts banned users from other social networks, many of whom were banned due to their use of hate speech and extremist content.

TARGETS OF ONLINE HATE SPEECH

One of the few areas of consensus in defining hate speech, which separates it from other forms of harmful speech, is that hate speech targets groups or individuals as they relate to a group (Sellars 2016). A small body of literature has explicitly analyzed the targets of online hate speech. Studying the targets of online hate speech on Whisper (an anonymous online platform) and Twitter using a sentence-structure–based algorithm, Silva et al. (2016) find that targeted individuals on both platforms are primarily attacked on the basis of their ethnicity, physical characteristics, sexual orientation, class, or gender. Survey research suggests that victims of online hate speech tend to engage in high levels of online activity (Hawdon et al. 2014), have less online anonymity, and engage in more online antagonism (Costello, Rukus, and Hawdon 2018). Examining the targets of hate speech on Twitter, ElSherief, Nilizadeh et al. (2018) find that those targeted by hate speech were 60 percent more likely to be verified than the accounts of instigators and 40 percent more likely to be verified than general users, respectively. This suggests that more visible Twitter users (with more followers, retweets, and lists) are more likely to become targets of hate.

Along these lines, recent qualitative research suggests that journalists, politicians, artists, bloggers, and other public figures have been disproportionately targeted by hate speech (Isbister et al. 2018). For example, when the all-female reboot of *Ghostbusters* was released in July 2016, white supremacist Milo Yiannopoulos instigated a Twitter storm following the publication of his negative movie review on Breitbart. White supremacists began to bombard African American actress Leslie Jones's timeline with sexist and racist slurs and hateful memes, including rape and death threats. When the abuse escalated as Yiannopoulos began directly tweeting at Jones and egging on his followers, Jones left Twitter. After public pressure convinced the company to intervene, Yiannopoulos was banned from Twitter and Jones returned (Isaac 2016). Similarly, as the author Mikki Kendall describes, "I was going to leave Twitter at one point. It just wasn't usable for me. I would log on and have 2,500 negative comments. One guy who seemed to have an inexhaustible energy would Photoshop my image on top of lynching pictures and tell me I should be 'raped by dogs,' that kind of thing." Kendall was also doxxed – her address was made public online – and she received a picture of her and her family in a photo that "looked like it had been sighted through a rifle" (Isaac 2016).

In June 2016, several highly visible Jewish journalists began to report a barrage of online hate that involved steganography – triple parentheses placed around their names like (((this))) (Fleishman and Smith 2016). As a result, the Anti-Defamation League (ADL) added the triple parentheses to their database of hateful symbols. This "digital equivalent of a yellow star" was intended to identify Jews as targets for harassment online (Gross 2017). For example, Jonathan Weisman of the *New York Times* left Twitter after being subjected to anti-Semitic harassment beginning with a Twitter account known as @CyberTrump, which escalated to a barrage of hateful Twitter activity, voicemails, and emails containing slurs and violent imagery (Gross 2017).

As these examples suggest, online hate speech may be most visible in coordinated attacks detecting this behavior (Mariconti et al. 2018). Such attacks draw a great deal of attention both online and through traditional media outlets, making these strategic targets useful for both extremists and trolls seeking to reach a broader audience and elevate their messages. Such coordinated harassment campaigns allow groups of anonymous individuals to work together to bombard particular users with harmful content again and again (Chess and Shaw 2015; Chatzakou et al. 2017). One manifestation of this behavior is known as raiding, when ad hoc digital mobs organize and orchestrate attacks aimed to disrupt other platforms and undermine users who advocate issues and policies with which they disagree (Hine et al. 2016; Kumar et al. 2018; Mariconti et al. 2018). However, while raiding receives a great deal of media attention, we have little understanding of how common or pervasive these attacks are or on what platforms they most commonly occur.

PREVALENCE OF ONLINE HATE SPEECH

While a great deal of research has been devoted to defining and detecting online hate speech, we know surprisingly little about the popularity of online hate speech on either mainstream or fringe platforms, or how the volume of hate speech shifts in response to events on the ground. Social media platforms have increased the visibility of hate speech, prompting journalists and academics alike to assert that hate speech is on the rise. As a result, there is a tendency to characterize entire mainstream social media platforms as bastions of online hate, without using empirical evidence to evaluate how pervasive the phenomenon truly is. For example, after becoming the target of a hateful online attack, *Atlantic* editor Jeffrey Goldberg called Twitter "a cesspool for anti-Semites, homophobes, and racists" (Lizza 2016). While any online hate speech is of course problematic, suggesting that a platform used by more than a quarter of Americans and millions more around the globe is dominated by such speech is misleading and potentially problematic – particularly in countries where civil and political liberties are already under threat and social media provides a valuable outlet for opposition voices (Gagliardone et al. 2016).

With regard to empirical evidence, a small handful of studies have begun to systematically evaluate the prevalence of hate speech on online platforms, though more work is needed. Analyzing the popularity of hate speech in more than 750 million political tweets and in 400 million tweets sent by a random sample of American Twitter users between June 2015 and June 2017, Siegel et al. (2020) find that, even on the most prolific days, only a fraction of a percentage of tweets in the American Twittersphere contain hate speech. Similarly, studying the popularity of hate speech on Ethiopian Facebook pages, Gagliardone et al. (2016) find that only 0.4 percent of statements in their representative sample were classified as hate speech, and 0.3 percent of tweets were classified as dangerous speech, which directly or indirectly calls for violence against a particular group.

While these studies suggest that online hate speech is a relatively rare phenomenon, cross-national survey research suggests that large numbers of individuals have nonetheless been incidentally exposed to online hate speech.

In a cross-national survey of internet users between the ages of fifteen and thirty, 53 percent of American respondents report being exposed to hate material online, while 48 percent of Finns, 39 percent of Brits, and 31 percent of Germans report exposure. Using online social networks frequently and visiting "dangerous sites" are two of the strongest predictors of such exposure (Hawdon et al. 2017). Perhaps explaining the discrepancy between empirical findings that hate speech is quite rare on mainstream platforms and high rates of self-reported exposure, Kaakinen et al. (2018) find that, while hateful content is rarely produced, it is more visible than other forms of content. Hate speech is also more common in particular online demographic communities than others. For example, Saha et al. (2019) find that hate speech is more prevalent in subreddits associated with particular colleges and universities than popular subreddits that were not associated with colleges or universities.

In addition to exploring the prevalence of online hate speech, recent work has investigated how offline events may drive upticks in the popularity of such rhetoric. One avenue of research explores the impact of violent offline events on various types of hate speech. For example, studying the causal effect of terror attacks in Western countries on the use of hateful language on Reddit and Twitter, Olteanu et al. (2018) find that episodes of extremist violence lead to an increase in online hate speech, particularly messages directly advocating violence, on both platforms. The authors argue that this provides evidence that theoretical arguments regarding the feedback loop between offline violence and online hate speech are – unfortunately – well-founded. This finding supports other research suggesting hate speech and hate crimes tend to increase after "trigger" events, which can be local, national, or international, and often drive negative sentiments toward groups associated with suspected perpetrators of violence (Awan and Zempi 2015).

Similarly, seeking to assess the impact of diverse episodes of sectarian violence on the popularity of anti-Shia hate speech in the Saudi Twittersphere, Siegel et al. (2018) find that both violent events abroad and domestic terror attacks on Shia mosques produce significant upticks in the popularity of anti-Shia language in the Saudi Twittersphere. Providing further insight into the mechanisms by which offline violent events lead to increases in the use of online hate speech, the authors demonstrate that, while clerics and other elite actors both instigate and spread derogatory rhetoric in the aftermath of foreign episodes of sectarian violence – producing the largest upticks in anti-Shia language – they are less likely to do so following domestic mosque bombings.

Exploring the effect of political – rather than violent – events on the popularity of online hate speech, Siegel et al. (2020) find, contrary to the popular journalistic narrative, that online hate speech did not increase either over the course of Donald Trump's 2016 campaign or in the aftermath of his unexpected election. Using a dataset of more than 1 billion tweets, their results are robust whether they detect hate speech using a machine-learning–augmented dictionary-based approach or a community-based detection algorithm comparing the similarity of daily Twitter data to the content produced on hateful subreddits over time. Instead, hate speech was "bursty" – spiking in the aftermath of particular events and re-equilibrating shortly afterward. Similarly, Faris et al. (2016) demonstrate spikes in online harmful speech are often linked to political events, whereas Saleem et al. (2017) find that hate speech rose in the aftermath of events that triggered strong emotional responses like the Baltimore protests and the US Supreme Court decision on same-sex marriage.

Together, these studies demonstrate the importance of examining both the prevalence and the dynamics of online hate speech systematically over time and using large representative samples. More work is needed to better understand how different types of online hate speech gain traction in diverse global contexts and how their relative popularity shifts on both mainstream and specialized social media platforms over time.

OFFLINE CONSEQUENCES OF ONLINE HATE SPEECH

Systematically measuring the impact of online hate speech is challenging (Sellars 2016), but a diverse body of research suggests that online hate speech has serious offline consequences both for individuals and for groups. Surveys of internet users indicate that exposure to online hate speech may cause fear (Hinduja and Patchin 2007), particularly in historically marginalized or disadvantaged populations. Other work suggests that such exposure may push people to withdraw from public debate both on- and offline, therefore harming free speech and civic engagement (Henson et al. 2013). Indeed, observational data indicate that exposure to hate speech may have many of the same consequences as being targeted by hate crimes, including psychological trauma and communal fear (Gerstenfeld 2017). Along these lines, human rights groups have argued that failure to monitor and counter hate speech online can reinforce the subordination of targeted minorities, making them vulnerable to attacks, while making majority populations more indifferent to such hatred (Izsak 2015). That being said, recent work demonstrates that interpretations of hate speech – what is considered hateful content as well as ratings of the intensity of content – differ widely by country (Salminen et al. 2019), and men and political conservatives tend to find hate material less disturbing than women, political moderates, and liberals (Costello et al. 2019).

On the individual level, qualitative research suggests that Muslims living in the West who are targeted by online hate speech fear that online threats may materialize offline (Awan and Zempi 2015). Furthermore, surveys of adolescent internet users have found that large numbers of African American respondents have experienced individual or personal discrimination online, and such exposure is associated with depression and anxiety, controlling for measures of offline discrimination (Tynes et al. 2008). In studying the differential effects of exposure to online hate speech, Tynes and Markoe (2010) find from a survey experiment conducted on college-age internet users that African American participants were most bothered by racist content (images) on social networking sites, whereas European Americans – especially those who held "color-blind" attitudes – were more likely to be "not bothered" by those images. Similarly, individuals exposed to hate speech on university-affiliated subreddits exhibited higher levels of stress than those who were not (Saha et al. 2019). Survey data suggest that youth who have been exposed to online hate speech have weaker attachment to family and report higher levels of unhappiness, though this relationship is not necessarily causal (Hawdon et al. 2014). Exposure to hate speech online is also associated with an avoidance of political talk over time (Barnidge et al. 2019). At the group level, online hate speech has fueled intergroup tensions in a variety of contexts, sometimes leading to violent clashes and undermining social cohesion (Izsak 2015). For example, Facebook has come under fire for its role in mobilizing anti-Muslim mob violence in Sri Lanka and for inciting violence against the Rohingya people in Myanmar (Vindu, Kumar, and Frenkel 2018). Elucidating the mechanisms by which exposure to hate speech drives intergroup tension, survey data and experimental evidence from Poland suggest that frequent and repetitive exposure to hate speech leads to desensitization to hateful content, lower evaluations of populations targeted by hate speech, and greater distancing – resulting in higher levels of anti–out-group prejudice (Soral et al. 2018).

A diverse body of literature suggests that hate speech may foster an environment in which bias-motivated violence is encouraged either subtly or explicitly (Herek et al. 1992; Greenawalt 1996; Calvert 1997; Tsesis 2002; Matsuda 2018). Intergroup conflict is more likely to occur and spread when individuals and groups have the opportunity to publicly express shared grievances and coordinate collective action (Weidmann 2009; Cederman et al. 2010). Digital technology is thought to reduce barriers to collective action among members of the same ethnic or religious group by improving access to information about one another's preferences. This is thought to increase the likelihood of intergroup conflict and accelerate its spread across borders (Pierskalla and Hollenbach 2013; Bailard 2015; Weidmann 2015).

Moreover, while hate speech is just one of many factors that interact to mobilize ethnic conflict, it plays a powerful role in intensifying feelings of mass hate (Vollhardt et al. 2007; Gagliardone et al. 2014). This may be particularly true in the online sphere, where the anonymity of online communication can drive people to express more hateful opinions than they might otherwise (Cohen-Almagor 2017). As individuals come to believe that "normal" rules of social conduct do not apply (Citron 2014; Delgado and Stefancic 2014), intergroup tensions are exacerbated. Along these lines, online hate speech places a physical distance between speaker and audience, emboldening individuals to express themselves without repercussions (Citron 2014). Perhaps more importantly, online social networks create the opportunity for individuals to engage with like-minded others that might otherwise never connect or be aware of one another's existence (Posner 2001). Recognizing the importance of online hate speech as an early warning sign of ethnic violence, databases of multilingual hate speech are increasingly used by governments, policymakers, and NGOs to detect and predict political instability, violence, and even genocide (Gagliardone et al. 2014; Tuckwood 2014; Gitari et al. 2015).

Many have argued that there is a direct connection between online hate and hate crimes, and perpetrators of offline violence often cite the role online communities have played in driving them to action (Citron 2014; Cohen-Almagor 2017; Gerstenfeld 2017). For example, on June 17, 2015, twenty-one-year-old Dylann Roof entered the Emanuel African Methodist Episcopal Church and murdered nine people. In his manifesto, Roof wrote that he drew his first racist inspiration from the Council of Conservative Citizens (CCC) website (Cohen-Almagor 2018). Similarly, the perpetrator of the 2019 Pittsburgh synagogue attack was allegedly radicalized on Gab, and the perpetrator of the 2019 New Zealand mosque shootings was reportedly radicalized on online platforms and sought to broadcast his attack on YouTube.

While it is very difficult to causally examine the link between online hate speech and hate crimes, recent empirical work has attempted to do so. This work builds off of a larger literature exploring how the use of hate speech through traditional media platforms can be used to trigger violent outbursts or ethnic hatred. This includes work exploring the effect of hate radio on levels of violence during the Rwandan genocide (Yanagizawa-Drott 2014), research on how radio propaganda incited anti-Semitic violence in Nazi Germany, (Adena et al. 2015), and a study of how nationalist Serbian radio was used to incite violence in Croatia in the 1990s (DellaVigna et al. 2014).

Examining the effects of online hate, Chan et al. (2015) find that broadband availability increases racial hate crimes in areas with higher levels of segregation and a higher proportion of racially charged Google search terms. Their work suggests that online access is increasing the incidence of racial hate crimes executed by lone wolf perpetrators. Similarly, Stephens-Davidowitz (2017) finds that the search rate on Google for anti-Muslim words and phrases, including violent terms like "kill all Muslims," can be used to predict the incidence of anti-Muslim hate crimes over time. Other studies show an association between hateful speech on Twitter and hate crimes in the US context, but the causal links are not well identified (Williams et al. 2019; Chyzh et al. 2019).

In one of the only existing studies that explicitly examines the causal link between online hate and offline violence, Muller and Schwarz (2017) exploit exogenous variation in major internet and Facebook outages to show that anti-refugee hate crimes increase disproportionately in areas with higher Facebook usage during periods of high anti-refugee sentiment online. They find that this effect is particularly pronounced for violent incidents against refugees, including arson and assault. Similarly, in a second paper, Muller and Schwarz

(2019) exploit variation in early adoption of Twitter to show that higher Twitter usage is associated with an increase in anti-Muslim hate crimes since the start of Trump's campaign. Their results provide preliminary evidence that social media can act as a propagation mechanism between online hate speech and offline violent crime. Together, this work suggests that online hate speech may have powerful real-world consequences, ranging from negative psychological effects at the individual level to violent attacks offline.

COMBATING ONLINE HATE SPEECH

Rising concern regarding these real-world effects of online hate speech have prompted researchers, policymakers, and online platforms to develop strategies to combat online hate speech. These approaches have generally taken two forms: content moderation and counter-speech.

One strategy to combat online hate speech has been to moderate content, which involves banning accounts or communities that violate platforms' terms of service or stated rules (Kiesler et al. 2012). On May 31, 2016, the European Commission in conjunction with Facebook, Twitter, YouTube, and Microsoft issued a voluntary Code of Conduct on Countering Illegal Hate Speech Online that required the removal of any hate speech, as defined by the European Union (EU). This was spurred by fears over a rise in intolerant speech against refugees as well as worries that hate speech fuels terror attacks (Aswad 2016). Additionally, beginning in December 2017, facing pressure in the aftermath of the deadly August 2017 "Unite the Right" march in Charlottesville, Virginia, Twitter announced a new policy to ban accounts that affiliate with groups "that use or promote violence against civilians to further their causes" (Twitter 2017). The platform began by suspending several accounts with large followings involved in white nationalism or in organizing the Charlottesville march. In this period, Twitter also suspended a far-right British activist who had been retweeted by President Trump, as well as several other accounts affiliated with her ultranationalist group (Nedig 2017). The company announced that their ban on violent threats would also be extended to include any content that glorifies violence (Twitter 2017). Similarly, in April 2018, Facebook announced its twenty-five-page set of rules dictating what types of content are permitted on Facebook (2018). The section on hate speech states, "We do not allow hate speech on Facebook because it creates an environment of intimidation and exclusion and in some cases may promote real-world violence." The goal of banning hate speech from more mainstream online platforms is to reduce the likelihood that everyday internet users are incidentally exposed to online hate speech.

However, little is known about how these bans are actually implemented in practice or how effective they have been in reducing online hate speech on these platforms or exposure to such speech more broadly. Moreover, the use of automatic hate speech detection has come under fire in the media as the limits of these methods have been highlighted by embarrassing mistakes – like when Facebook's proprietary filters flagged an excerpt from the Declaration of Independence as hate speech (Lapin 2018). While a February 2019 review by the European Commission suggests that social media platforms including Facebook and Google were successfully removing 75 percent of posts flagged by users that violate EU standards within 24 hours, we do not know what portion of hate speech is flagged or how this may be biased against or in favor of certain types of political speech (Laub 2019).

Empirical work on the effectiveness of banning hateful content yields mixed results. Studying the effect of banning the /fatpeoplehate and /CoonTown subreddits on Reddit in 2015, Chandrasekharan, Pavalanathan et al. (2017) find the ban was successful. Analyzing more than 100 million Reddit posts and comments, the authors found that many accounts discontinued using the site after the ban, and those that stayed decreased their hate speech usage by at least 80 percent. Although many of these users migrated to other subreddits, the new subreddits did not experience an increase in hate speech usage, suggesting the ban was successful in limiting online hate speech on Reddit. Also on Reddit, Saleem and Ruths (2019) find that banning a large hateful subreddit (r/fatpeoplehate) prompted users of this subreddit to stop posting on Reddit. Similarly, other work suggests that banning accounts on Twitter disrupts extremist social networks, as users who are frequently banned suffer major drops in follower counts when they rejoin a particular platform (Berger and Perez 2016).

That being said, although bans may have decreased the overall volume of hate speech on Redditt, and disrupted extremist activity on Twitter, such activity may have simply migrated to other platforms. In response to the 2015 bans, Newell et al. (2016) find that disgruntled users sought out alternative platforms such as Voat, Snapzu, and Empeopled. Users who migrate to these fringe platforms often keep their usernames and attempt to recreate their banned communities in a new, less regulated domain (Chandrasekharan, Pavalanathan et al. 2017). In addition to moving hate speech from one platform to another, other work suggests that producers of harmful content simply become more creative about how to continue to use hate speech on their preferred platforms. For example, seeking to avoid content moderation, as previously described, members of online communities often use code words to circumvent detection (Chancellor et al. 2016; Sonnad 2016).

Additionally, attempts to ban user accounts may sometimes be counterproductive, galvanizing support from those who are sympathetic to hateful communities. When well-known users come under fire, people who hold similar beliefs may be motivated to rally to their defense and/or to express views that are opposed by powerful companies or organizations. For example, empirical studies of extremist behavior online examining pro-ISIS accounts suggest that online extremists view the blocking of their accounts as a badge of honor, and individuals who have been blocked or banned are often able to reactivate their accounts under new names (Vidino and Hughes 2015; Berger and Perez 2016). Moreover, banning users often prompts them to move to more specialized platforms, such as Gab or Voat, which may further radicalize individuals who produce online hate. Indeed, banning hateful users removes them from diverse settings where they may come into contact with moderate or opposing voices, elevating their grievances and feelings of persecution and pushing them into hateful echo chambers where extremism and calls for offline violence are normalized and encouraged (Marwick and Lewis 2017; Lima et al. 2018; Zannettou et al. 2018; Jackson 2019). While this is a compelling theoretical argument against banning users from mainstream platforms, more empirical work is needed to track the extent to which banned users migrate to more extreme platforms, as well as whether they indeed become further radicalized on these platforms (Jackson 2019).

In this way, existing empirical work on the effectiveness of content moderation suggests that, while it may reduce hate speech on particular platforms, as disgruntled users migrate to other corners of the Internet, it is unclear whether such efforts reduce hate speech overall. Moreover, thorny legal, ethical, and technical questions persist with regard to the benefits of banning hate speech on global social media platforms, particularly outside of Western democracies. For example, a recent ProPublica investigation found that Facebook's rules are not transparent and inconsistently applied by tens of thousands of global contractors charged with content moderation. In many countries and disputed territories, such as the Palestinian territories, Kashmir, and Crimea, activists and journalists have been censored for harmful speech as Facebook has responded to government concerns and worked to insulate itself from legal liability. The report concluded that Facebook's hate speech content moderation standards "tend to favor elites and governments over grassroots activists and racial minorities." Along these lines, governments may declare opposition speech to be hateful or extremist in order to manipulate content moderation to silence their critics (Laub 2019). Moreover, automated hate speech detection methods have not been well adapted to local contexts, and very few content moderators are employed that speak local languages – including those that are used to target at-risk minority groups who are often targeted by hate speech. In a famous example, in 2015, despite rising ethnic violence and rampant reports of hate speech on Facebook and other social media platforms targeting Muslims in Myanmar, Facebook allegedly just employed two Burmese-speaking content moderators (Stecklow 2018).

Recognizing that censoring hate speech may come into conflict with legal protections of free speech or may be manipulated by governments to target critics, international agencies such as UNESCO have generally maintained that "the free flow of information should always be the norm." As a result, they often argue that counter-speech is usually preferable to the suppression of hate speech (Gagliardone et al. 2015). Counter-speech is a direct response to hate speech intended to influence discourse and behavior (Benesch 2014a, 2014b). Counter-speech campaigns have long been used to combat the public expression of hate speech and discrimination through traditional media channels. Examples of this in the US context include the use of anti-KKK billboards in the Deep South (Richards and Calvert 2000), and the dissemination of information about US hate groups by the Southern Poverty Law Center (McMillin 2014). Interventions designed to prevent the incitement of violence have also been deployed, including the use of soap operas to counter intergroup tensions in Rwanda and the use of television comedy in Kenya to discourage the use of hate speech (Staub et al. 2003; Paluck 2009; Kogen 2013). Experimental evaluations of these interventions have found that they may make participants better able to recognize and resist incitement to anti–out-group hatred.

More recent work has explored the use of counter-speech in the online sphere. For example, fearing violence in the lead-up to the 2013 Kenyan elections, international NGOs, celebrities, and local businesses helped to fund "peace propaganda" campaigns to deter the spread of online hate speech – and offline violence – in Kenya. For example, one company offered cash and cell phone time to Kenyans who sent peace messages to each other online, including photos, poems, and stories (Benesch 2014a). Demonstrating that counter-speech occurs organically on online platforms, in the aftermath of the 2015 Paris attacks, Magdy et al. (2016) estimate that the vast majority of tweets posted following the attacks were defending Muslims, while anti-Muslim hate tweets represented a small fraction of content in the Twittersphere. Similarly, examining online hate speech in Nigerian political discussions, Bartlett et al. (2015) find that extreme content is often met with disagreement, derision, and counter-messages.

A nascent strand of literature experimentally evaluates what types of counter-speech messages are most effective in reducing online hate speech. Munger (2017) shows that counter-speech using automated bots can reduce instances of racist speech if instigators are sanctioned by a high-status in-group member – in this case, a white male with a large number of Twitter followers. Similarly, Siegel and Badaan (2020) deployed a sockpuppet account to counter sectarian hate speech in the Arab Twittersphere. They find that simply receiving a sanctioning message reduces the use of hate speech, particularly for users in networks where hate speech is relatively uncommon. Moreover, they show that messages priming a common Muslim religious identity containing endorsements from elite actors are particularly effective in decreasing users' posttreatment level of hate speech. Additional research is needed to further evaluate what types of counter-speech from what sources are most effective in reducing online hate in diverse contexts. Recognizing the potential of counter-speech bots, Leetaru (2017) proposed deploying AI bots en masse to fight online hate speech, though the feasibility and consequences of such an intervention are not well understood. Simulating how much counter-speech might be necessary to "drown out" hate speech on Facebook, Schieb and Preuss (2016) find that counter-speech can have a considerable impact on reducing the visibility of online hate speech, especially when producers of hate speech are in the minority of a particular community. In one of the only studies that explicitly detects naturally occurring counter-speech on social media, (Mathew et al. 2018; Mathew, Saha et al. 2019) find that counter-speech comments receive much more likes and engagement than other comments and may prompt produces of hate speech to apologize or change their behavior. More empirical work is needed, however, to see how this dynamic plays out more systematically on real-world social media platforms over time.

Explicitly comparing censorship or content monitoring to counter-speech interventions, Alvarez-Benjumea and Winter (2018) test whether decreasing social acceptability of hostile comments in an online forum decreases the use of hate speech. They first designed an online forum and invited participants to join and engage in conversation on current social topics. They then experimentally manipulated the comments participants observed before posting their own comments. They included a censoring treatment in which participants observed no hate comments and a counter-speech treatment in which hate speech comments were uncensored but were presented alongside posts highlighting the fact that hate speech was not considered acceptable on the platform. Comparing the level of hostility of the comments and instances of hate across the treatment conditions, they find that the censoring treatment was the most effective in reducing hostile comments. However, the authors note that the fact that they do not observe a statistically significant effect of the counter-speech treatment may be due to their small sample sizes and inability to monitor repeated interactions over time in their experimental setup. Together, this growing body of literature on the effects of censoring and counter-speech on online hate speech provides some optimism, particularly regarding the impact of content moderation on reducing hate speech on mainstream platforms and the ability of counter-speech campaigns to decrease the reach, visibility, and harm of online hate speech. However, we know very little about the potential collateral damage of these interventions. Future work should not only provide larger scale empirical tests of these types of interventions in diverse contexts but seek to evaluate the longer-term effects of these approaches.

CONCLUSIONS AND STEPS FOR FUTURE RESEARCH

As online hate speech has become increasingly visible on social media platforms, it has emerged at the center of academic, legal, and policy agendas. Despite increased attention to online hate speech, as this chapter demonstrates, the debate over how to define online hate speech is far from settled. Partly as a consequence of these definitional challenges, and partly as a result of the highly context-specific and evolving nature of online hate speech, detecting hateful content systematically is an extremely difficult task.

While state-of-the-art techniques employing machine learning, neural networks and incorporating contextual features have improved our ability to measure and monitor online hate speech, most existing empirical work is fairly fragmented – often detecting a single type of hate speech on one platform at one moment of time. Moreover, because of ease of data collection, the vast majority of studies have been conducted using English-language Twitter data and therefore do not necessarily tell us very much about other platforms or cultural contexts. Adding further complications, definitions of hate speech and approaches to detecting it are highly politicized, particularly in authoritarian contexts and conflict settings. Though some research has explored multiple types of hate speech, used several datasets, conducted research on multiple platforms, or examined trends in hate speech over time, these studies are the exception rather than the rule (Fortuna 2017). Drawing on the rich literature of hate speech detection techniques in computer science and social science, future work should attempt more systematic comparative analysis to improve our ability to detect online hate speech in its diverse forms.

Though less developed than the literature on defining and measuring online hate speech, recent work has explored both the producers of online hate speech and their targets. A large body of literature has evaluated how hate groups strategically use the Internet to lure recruits and foster a sense of community among disparate members, using primarily small-scale qualitative analysis of data from hate groups' official websites (Selepak 2010). Other work has conducted large-scale observational studies of the users that produce hate speech on mainstream social media platforms like Twitter and Reddit, including their demographic characteristics and network structures. These users tend to be young, male, very active on social media, and members of tightly networked communities in which producers of hate speech frequently retweet and like each other's posts (Costello and Hawdon 2018; Ribeiro et al. 2018).

With regard to the targets of hate speech, researchers have used both big data empirical analyses and surveys of the users targeted online to demonstrate that targets of hate speech are often prominent social media users with large followings (ElSherief, Nilizadeh et al. 2018). Additionally, qualitative and quantitative work demonstrates that one targeting strategy on mainstream social media platforms is for well-organized groups of users to launch coordinated hate attacks or "raids" on bloggers, celebrities, journalists, or other prominent actors (Mariconti et al. 2018). This may be one reason why online hate speech has received so much attention in the mainstream media, despite empirical evidence suggesting that hate speech is actually quite rare on mainstream social media platforms in aggregate.

Indeed, quantitative work evaluating the prevalence of online hate speech suggests that it may represent only a fraction of a percentage point of overall posts on sites like Facebook and Twitter (Gagliardone et al. 2016; Siegel et al. 2020). Moreover, studies exploring the dynamics of online hate speech over time on Twitter suggest that it is quite bursty – it increases in response to emotional or violent events and then tends to quickly re-equilibrate (Awan and Zempi 2015; Olteanu et al. 2018; Siegel et al. 2020).

Although hate speech may be rare, it can still have severe offline consequences. Survey data suggest that online hate speech negatively impacts the psychological well-being of individuals who are exposed to it and can have detrimental consequences for intergroup relations at the societal level (Tynes et al. 2008). A growing body of empirical evidence also suggests that online hate speech can incite people to violence and that it may be playing a particularly devastating role in fueling attacks on Muslim immigrants and refugees. Recent work exploring the causal effect of online hate speech on offline attitudes and behaviors (Chan et al. 2015; Muller and Schwarz 2017; Muller and Schwarz 2019) should be replicated, expanded, and adapted to enable us to better understand these dynamics in other contexts and over longer periods of time.

Scientific studies have also assessed what strategies might be most effective to combat online hate speech. Empirical evidence suggests that banning hateful communities on Reddit, for example, reduced the volume of hate speech on the platform overall (Chandrasekharan, Pavalanathan et al. 2017). However, other work indicates that users who are banned from discussing particular topics on mainstream platforms simply move elsewhere to continue their hateful discourse (Newell et al. 2016). Additionally, content and account bans could have galvanizing effects for certain extremist actors who view the sanction as a badge of honor (Vidino and Hughes 2015). More optimistically, experimental research using counter-speech to combat online hate speech suggests that receiving sanctioning messages from other Twitter users – particularly fellow in-group members, high-status individuals, or trusted elite actors – discourages users from tweeting hateful content (Munger 2017; Siegel and Badaan 2020). Moreover, large-scale empirical studies suggest that counter-speech is quite common in the online sphere, and the same events that trigger upticks in online hate speech often trigger much larger surges in counter-speech (Magdy et al. 2016; Olteanu et al. 2018). Future work should continue to explore what kinds of counter-speech might be most effective in diverse cultural contexts and on different platforms, as well as how counter-speech can be encouraged among everyday social media users. Given the dangerous offline consequences of online hate speech in diverse global contexts, academics and policymakers should continue to build on this existing literature to improve hate speech detection, gain a more comprehensive understanding of how hate speech arises and spreads, develop further understanding of hate speech's offline consequences, and build better tools to effectively combat it.

REFERENCES

Albadi, N., Kurdi, M., & Mishra, S. (2019). Investigating the effect of combining GRU neural networks with handcrafted features for religious hatred detection on Arabic Twitter space. *Social Network Analysis and Mining*, *9*(41), 1–19.

Al-Makhadmeh, Z., & Tolba, A. (2020). Automatic hate speech detection using killer natural language processing optimizing ensemble deep learning approach. *Computing*, *102*, 501–522. https://doi.org/10.1007/s00607-019-00745-0

Adams, J., & Roscigno, V. J. (2005). White supremacists, oppositional culture and the World Wide Web. *Social Forces*, *84*(2), 759–778.

Adena, M., Enikolopov, R., Petrova, M., Santarosa, V., & Zhuravskaya, E. (2015). Radio and the rise of the Nazis in prewar Germany. *The Quarterly Journal of Economics*, *130*(4), 1885–1939.

Agarwal, S., & Sureka, A. (2017). Characterizing linguistic attributes for automatic classification of intent based racist/radicalized posts on Tumblr microblogging website. arXiv.org. https://arxiv.org/abs/1701.04931

Alorainy, W., Burnap, P., Liu, H., & Williams, M. (2018). Cyber hate classification: "Othering" language and paragraph embedding. arXiv.org. https://arxiv.org/pdf/1801.07495.pdf

Alvarez-Benjumea, A., & Winter, F. (2018). Normative change and culture of hate: An experiment in online environments. *European Sociological Review*, *34*(3), 223–237.

Aswad, E. (2016). The role of US technology companies as enforcers of Europe's new Internet hate speech ban. *HRLR Online*, *1*(1), 1–14.

Aulia, N., & Budi, I. (2019). Hate speech detection on Indonesian long text documents using machine learning approach. In *Proceedings of the 2019 5th International Conference on Computing and Artificial Intelligence* (pp. 164–169). New York: ACM.

Awan, I., & Zempi, I. (2015). We fear for our lives: Offline and online experiences of anti-Muslim hostility. *Tell MAMA*, October. www.tellmamauk.org/wp-content/uploads/resources/We%20Fear%20For%20Our%20Lives.pdf

Badjatiya, P., Gupta, S., Gupta, M., & Varma, V. (2017). Deep learning for hate speech detection in tweets. In *Proceedings of the 26th International Conference on World Wide Web Companion* (pp. 759–760). Geneva: International World Wide Web Conferences Steering Committee.

Bailard, C. S. (2015). Ethnic conflict goes mobile: Mobile technology's effect on the opportunities and motivations for violent collective action. *Journal of Peace Research*, *52*(3), 323–337.

Barnidge, M., Kim, B., Sherrill, L. A., Luknar, Ž., & Zhang, J. (2019). Perceived exposure to and avoidance of hate speech in various communication settings. *Telematics and Informatics*, 44, 101263.

Bartlett, J., Krasodomski-Jones, A., Daniel, N., Fisher, A., & Jesperson, S. (2015). *Social Media for Election Communication and Monitoring in Nigeria*. Demos report, London.

Basile, V., Bosco, C., Fersini, E. et al. (2019). Semeval-2019 task 5: Multilingual detection of hate speech against immigrants and women in twitter. In J. May, E. Shutova, A. Herbelot, X. Zhu, M. Apidianaki, & S. M. Mohammad (Eds.), *Proceedings of the 13th International Workshop on Semantic Evaluation* (pp. 54–63). Minneapolis: Association for Computer Linguistics.

Beauchamp, N., Panaitiu, I., & Piston, S. (2018). Trajectories of hate: Mapping individual racism and misogyny on Twitter. Unpublished working paper.

Benesch, S. (2013). Dangerous speech: A proposal to prevent group violence. Voices That Poison: Dangerous Speech Project proposal paper. February 23. https://dangerousspeech.org/wp-content/uploads/2018/01/Dangerous-Speech-Guidelines-2013.pdf

(2014a). *Countering Dangerous Speech to Prevent Mass Violence During Kenya's 2013 Elections*. United States Institute of Peace final report, February 28. https://ihub.co.ke/ihubresearch/jb_BeneschCFPReportPeacebuildingInKenyapdf2014-3-25-07-08-41.pdf

(2014b). Defining and diminishing hate speech. In P. Grant (Ed.), *Freedom from Hate: State of the World's Minorities and Indigenous Peoples 2014* (pp. 18–25). London: Minority Rights Group International. https://minorityrights.org/wp-content/uploads/old-site-downloads/mrg-state-of-the-worlds-minorities-2014.pdf

Berger, J. M., & Perez, H. (2016). The Islamic State's diminishing returns on Twitter: How suspensions are limiting the social networks of English-speaking ISIS supporters. Occasional paper. Program on Extremism at George Washington University.

Black, E. W., Mezzina, K., & Thompson, L. A. (2016). Anonymous social media: Understanding the content and context of Yik Yak. *Computers in Human Behavior*, *57*(C), 17–22.

Bowman-Grieve, L. (2009). Exploring Stormfront: A virtual community of the radical right. *Studies in Conflict and Terrorism*, *32*(11), 989–1007.

Burnap, P., & Williams, M. L. (2016). Us and them: Identifying cyber hate on Twitter across multiple protected characteristics. *EPJ Data Science*, *5*(1), 1–15.

Calvert, C. (1997). Hate speech and its harms: A communication theory perspective. *Journal of Communication*, 47(1), 4–19.

Castle, T. (2012). Morrigan rising: Exploring female-targeted propaganda on hate group websites. *European Journal of Cultural Studies*, *15*(6), 679–694.

Cederman, L.-E., Wimmer, A., & Min, B. (2010). Why do ethnic groups rebel? New data and analysis. *World Politics*, *62*(1), 87–119.

Chan, J., Ghose, A., & Seamans, R. (2015). The Internet and racial hate crime: Offline spillovers from online access. *MIS Quarterly*, *14*(2), 381–403.

Chancellor, S., Pater, J. A., Clear, T., Gilbert, E., & De Choudhury, M. (2016). # thyghgapp: Instagram content moderation and lexical variation in pro-eating disorder communities. In *Proceedings of the 19th ACM Conference on Computer-Supported Cooperative Work and Social Computing* (pp. 1201–1213). New York: ACM.

Chandrasekharan, E., Samory, M., Srinivasan, A., & Gilbert, E. (2017a). The bag of communities: Identifying abusive behavior online with preexisting internet data. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems* (pp. 3175–3187). New York: ACM.

Chandrasekharan, E., Pavalanathan, U., Srinivasan, A., Glynn, A., Eisenstein, J., & Gilbert, E. (2017b). You can't stay here: The efficacy of Reddit's 2015 ban examined through hate speech. In *Proceedings of the ACM on Human-Computer Interaction*, Vol. 1 (CSCW) (pp. 1–22). New York: Association for Computing Machinery.

Chatzakou, D., Kourtellis, N., Blackburn, J., De Cristofaro, E., Stringhini, G., & Vakali, A. (2017). Mean birds: Detecting aggression and bullying on Twitter. In *Proceedings of the 2017 ACM on Web Science Conference* (pp. 13–22).

Chau, M., & Xu, J. (2007). Mining communities and their relationships in blogs: A study of online hate groups. *International Journal of Human-Computer Studies*, *65*(1), 57–70.

Chess, S., & Shaw, A. (2015). A conspiracy of fishes, or, how we learned to stop worrying about# GamerGate and embrace hegemonic masculinity. *Journal of Broadcasting and Electronic Media*, *59*(1), 208–220.

Chetty, N., & Alathur, S. (2018). Hate speech review in the context of online social networks. *Aggression and Violent Behavior*, *40*, 108–118.

Chowdhury, A. G., Didolkar, A., Sawhney, R., & Shah, R. (2019). ARHNet: Leveraging community interaction for detection of religious hate speech in Arabic. In F. Alva-Manchego, E. Choi, & D. Khashabi (Eds.), *Proceedings of the 57th Conference of the Association for Computational Linguistics: Student Research Workshop* (pp. 273–280). Florence: Association for Computational Linguistics.

Chyzh, O., Nieman, M. D., & Webb, C. (2019). The effects of dog-whistle politics on political violence. *Political Science Publications*, 1–10. https://lib.dr.iastate.edu/pols_pubs/59/

Citron, D. K. (2011). Misogynistic cyber hate speech. Written Testimony and Statement of Danielle Keates Citron, Professor of Law, Boston University School of Law hearing on “Fostering a Healthier Internet to Protect Consumers” before the House Committee on Energy and Commerce, October 16, 2019, Washington, DC.

(2014). *Hate Crimes in Cyberspace*. Cambridge, MA: Harvard University Press.

Cohen-Almagor, R. (2011). Fighting hate and bigotry on the Internet. *Policy and Internet*, *3*(3), 1–26.

(2017). Why confronting the Internet’s dark side? *Philosophia*, *45*(3), 919–929.

(2018). When a ritual murder occurred at Purim: The harm in hate speech. *El Profesional de la Informacion*, *27*(3), 671–681.

Costello, M., & Hawdon, J. (2018). Who are the online extremists among us? Sociodemographic characteristics, social networking, and online experiences of those who produce online hate materials. *Violence and Gender*, *5*(1), 55–60.

Costello, M., Hawdon, J., Bernatzky, C., & Mendes, K. (2019). Social group identity and perceptions of online hate. *Sociological Inquiry*, *89*(3), 427–452.

Costello, M., Rukus, J., & Hawdon, J. (2018). We don’t like your type around here: Regional and residential differences in exposure to online hate material targeting sexuality. *Deviant Behavior*, *40*(3), 1–17.

Czapla, P., Gugger, S., Howard, J., & Kardas, M. (2019). Universal language model fine-tuning for Polish hate speech detection. Paper presented at the Proceedings of the PolEval 2019 Workshop: 149. May 31, Warsaw, Poland.

Dadvar, M., de Jong, F. M. G., Ordelman, R., & Trieschnigg, D. (2012). Improved cyberbullying detection using gender information. In *Proceedings of the Twelfth DutchBelgian Information Retrieval Workshop (DIR 2012)* (pp. 23–25). Ghent: University of Ghent.

Daniels, J. (2017). Twitter and white supremacy: A love story. *Dame Magazine*, October 19. www.damemagazine.com/2017/10/19/twitter-and-white-supremacy-love-story/

Davidson, T., Warmsley, D., Macy, M., & Weber, I. (2017). Automated hate speech detection and the problem of offensive language. arXiv.org. https://arxiv.org/pdf/1703.04009.pdf

De Smedt, T., De Pauw, G., & Van Ostaeyen, P. (2018). *Automatic Detection of Jihadist Online Hate Speech*. CLiPS Technical Report No. 7 Computational Linguistics & Psycholinguistics Technical Report Series, Ctrs-007, February. www.uantwerpen.be/clips

Del Vigna, F., Cimino, A., Dell'Orletta, F., Petrocchi, M., & Tesconi. M. (2017). Hate me, hate me not: Hate speech detection on Facebook. In A. Armando, R. Baldoni, & R. Focardi (Eds.), *Proceedings of the First Italian Conference on Cybersecurity (ITASEC17)*, (pp. 86–95). Venice: CEUR.

Delgado, R. (1982). Words that wound: A tort action for racial insults, epithets, and name calling. *Harvard Civil Rights-Civil Liberties Review*, *17*, 133–181.

Delgado, R., & Stefancic, J. (2014). Hate speech in cyberspace. *Wake Forest Law Review*, 49, 319. https://ssrn.com/abstract=2517406

DellaVigna, S., Enikolopov, R., Mironova, V., Petrova, M., & Zhuravskaya, E. (2014). Cross-border media and nationalism: Evidence from Serbian radio in Croatia. *American Economic Journal: Applied Economics*, *6*(3), 103–132.

Dinakar, K., Reichart, R., & Lieberman, H. (2011). Modeling the detection of Textual Cyberbullying. *The Social Mobile Web*, *11*(02), 11–17.

Djuric, N., Zhou, J., Morris, R., Grbovic, M., Radosavljevic, V., & Bhamidipati, N. (2015). Hate speech detection with comment embeddings. In A. N. Joinson, K. Y. A. McKenna, T. Postmes, & U.-D. Reips (Eds.), *Proceedings of the 24th International Conference on World Wide Web* (pp. 29–30). New York: ACM.

Douglas, K. M. (2007). Psychology, discrimination and hate groups online. In *The Oxford Handbook of Internet Psychology* (pp. 155–163). Oxford: Oxford University Press.

Duarte, N., Llanso, E., & Loup, A. (2018). Mixed messages? The limits of automated social media content analysis. In *Proceedings of the 1st Conference on Fairness, Accountability and Transparency*, PMLR Vol. 81 (pp. 106). New York: Association for Computing Machinery.

ElSherief, M., Kulkarni, V., Nguyen, D., Wang, W. Y., & Belding, E. (2018). Hate lingo: A target-based linguistic analysis of hate speech in social media. arXiv.org. https://arxiv.org/abs/1804.04257

ElSherief, M., Nilizadeh, S., Nguyen, D., Vigna, G., & Belding, E. (2018). Peer to peer hate: Hate speech instigators and their targets. arXiv.org. https://arxiv.org/abs/1804.04649

Facebook. (2018). Facebook Community Standards. www.facebook.com/communitystandards/introduction

Faris, R., Ashar, A., Gasser, U., & Joo, D. (2016). Understanding harmful speech online. Berkman Klein Center Research Publication No. 2016-21. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2882824

Farkas, J., & Neumayer, C. (2017). Stop fake hate profiles on Facebook: Challenges for crowdsourced activism on social media. *First Monday*, 22(9). http://firstmonday.org/ojs/index.php/fm/article/view/8042

Fleishman, C., & Smith, A. (2016). Exposed: The secret symbol Neo-Nazis use to target Jews online. *Mic*, June 1. https://mic.com/articles/144228/echoes-exposed-the-secret-symbol-neo-nazis-use-to-target-jews-online

Flores-Yeffal, N. Y., Vidales, G., & Plemons, A. (2011). The Latino cyber-moral panic process in the United States. *Information, Communication and Society*, *14*(4), 568–589.

Fortuna, P. C. T. (2017). Automatic detection of hate speech in text: an overview of the topic and dataset annotation with hierarchical classes. U.Porto. https://repositorio-aberto.up.pt/handle/10216/106028

Fortuna, P., & Nunes, S. (2018). A survey on automatic detection of hate speech in text. *ACM Computing Surveys (CSUR)*, *51*(4), 85.

Gagliardone, I., Gal, D., Alves, T., & Martinez, G. (2015). *Countering Online Hate Speech*. Paris: UNESCO Publishing.

Gagliardone, I., Pohjonen, M., Beyene, Z. et al. (2016). Mechachal: Online debates and elections in Ethiopia: From hate speech to engagement in social media. Working paper. https://eprints.soas.ac.uk/30572/

Gagliardone, I., Patel, A., & Pohjonen, M. (2014). Mapping and analysing hate speech online: Opportunities and challenges for Ethiopia. University of Oxford Comparative Media, Law & Policy website. https://pcmlp.socleg.ox.ac.uk/mapping-and-analysing-hate-speech-online-opportunities-and-challenges-for-ethiopia/

Gerstenfeld, P. B. (2017). *Hate Crimes: Causes, Controls, and Controversies*. London: Sage.

Gitari, N. D., Zuping, Z., Damien, H., & Long, J. (2015). A lexicon-based approach for hate speech detection. *International Journal of Multimedia and Ubiquitous Engineering*, *10*(4), 215–230.

Greenawalt, K. (1996). *Fighting Words: Individuals, Communities, and Liberties of Speech*. Princeton: Princeton University Press.

Greevy, E., & Smeaton, A. F. (2004). Classifying racist texts using a support vector machine. In *Proceedings of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (pp. 468–469). New York: ACM.

Gross, T. (2017). Attacked by alt-right trolls: A Jewish journalist links Trump to the rise of hate. *NPR: Fresh Air*, March 19. www.npr.org/2018/03/19/594894657/attacked-by-alt-right-trolls-ajewish-journalist-links-trump-to-the-rise-of-hate

Haraszti, M. (2012). Foreword: Hate speech and coming death of the international standard before it was born (complaints of a watchdog). In M. Herz & P. Molnar (Eds.), *The Content and Context of Hate Speech: Rethinking Regulation and Responses* (pp. xiii–xviii). Cambridge: Cambridge University Press. https://doi.org/10.1017/CBO9781139042871.001

Hawdon, J., Oksanen, A., & Rasänen, P. (2014). Victims of online hate groups: American youths exposure to online hate speech. In J. Hawdon, J. Ryan, & M. Lucht (Eds.), *The Causes and Consequences of Group Violence: From Bullies to Terrorists* (pp. 165–182). Lanham, MD: Lexington Books.

Hawdon, J., Oksanen, A., & Rasänen, P. (2017). Exposure to online hate in four nations: A cross-national consideration. *Deviant Behavior*, *38*(3), 254–266.

Henson, B., Reyns, B. W., & Fisher, B. S. (2013). Fear of crime online? Examining the effect of risk, previous victimization, and exposure on fear of online interpersonal victimization. *Journal of Contemporary Criminal Justice*, 29(4), 475–497.

Herek, G. M., Berrill, K. T., & Berrill, K. (1992). *Hate Crimes: Confronting Violence Against Lesbians and Gay Men*. London: Sage.

Hinduja, S., & Patchin, J. W. (2007). Offline consequences of online victimization: School violence and delinquency. *Journal of School Violence*, 6(3), 89–112.

Hine, G. E., Onaolapo, J., De Cristofaro, E. et al. (2016). Kek, cucks, and god emperor Trump: A measurement study of 4chan's politically incorrect forum and its effects on the web. arXiv.org. https://arxiv.org/abs/1610.03452

Holtz, P., & Wagner, W. (2009). Essentialism and attribution of monstrosity in racist discourse: Right-wing Internet postings about Africans and Jews. *Journal of Community and Applied Social Psychology*, *19*(6), 411–425.

Howard, J. W. (2019). Free speech and hate speech. *Annual Review of Political Science*, *22*, 93–109.

Isaac, M. (2016). Twitter bars Milo Yiannopoulos in wake of Leslie Jones's reports of abuse. *New York Times*, July 20. www.nytimes.com/2016/07/20/technology/twitter-bars-milo-yiannopoulos-in-crackdown-on-abusive-comments.html

Isbister, T., Sahlgren, M., Kaati, L., Obaidi, M., & Akrami, N. (2018). Monitoring targeted hate in online environments. arXiv.org. https://arxiv.org/abs/1803.04757

Izsak, R. (2015). Hate speech and incitement to hatred against minorities in the media. UN Humans Rights Council. arXiv.org. www.ohchr.org/EN/Issues/Minorities/SRMinorities/Pages/Annual.aspx

Jackson, S. (2019). The double-edged sword of banning extremists from social media. https://osf.io/preprints/socarxiv/2g7yd/

Kaakinen, M., Räsänen, P., Näsi, M., Minkkinen, J., Keipi, T., & Oksanen, A. (2018). Social capital and online hate production: A four country survey. *Crime, Law and Social Change*, *69*(1), 25–39.

Kang, S., Kim, J., Park, K., & Cha, M. (2018). Classification of hateful comments in a Korean news portal. www.dbpia.co.kr/journal/articleDetail?nodeId=NODE07503234

Kennedy, B., Kogon, D., Coombs, K. et al. (2018). A typology and coding manual for the study of hate-based rhetoric. arXiv.org. https://psyarxiv.com/hqjxn/

Kiesler, S., Kraut, R., Resnick, P., & Kittur, A. (2012). Regulating behavior in online communities. In R. E. Kraut & P. Resnick (Eds.), *Building Successful Online Communities: Evidence-Based Social Design* (pp. 125–177). Cambridge, MA: MIT Press.

Klubicka, F., & Fernandez, R. (2018). Examining a hate speech corpus for hate speech detection and popularity prediction. arXiv.org. https://arxiv.org/abs/1805.04661

Kogen, L. (2013). Testing a media intervention in Kenya: *Vioja Mahakamani*, dangerous speech, and the Benesch guidelines. University of Pennsylvania Scholarly Commons. https://repository.upenn.edu/cgi/viewcontent.cgi?article=1000&context=africaictresearch

Kumar, S., Hamilton, W. L., Leskovec, J., & Jurafsky, D. (2018). Community interaction and conflict on the web. In *Proceedings of the 2018 World Wide Web Conference on World Wide Web* (pp. 933–943). Geneva: International World Wide Web Conferences Steering Committee.

Kwok, I., & Wang, Y. (2013). Locate the hate: Detecting Tweets against Blacks. In *AAAI'13 Proceedings of the 27th AAAI Conference on Artificial Intelligence* (pp. 1621–1622). Bellevue, WA: AAAI Press.

Lapin, T. (2018). Facebook flagged Declaration of Independence as hate speech. *New York Post*, 5 July. https://nypost.com/2018/07/05/facebook-flagged-declaration-of-independenceas-hate-speech/

Laub, Z. (2019). Hate speech on social media: Global comparisons. Council on Foreign Relations *Backgrounder*, June 7. www.cfr.org/backgrounder/hate-speech-social-media-global-comparisons

Leetaru, K. (2017). Fighting social media hate speech with AI-powered bots. *Forbes*, February 4. www.forbes.com/sites/kalevleetaru/2017/02/04/fighting-social-mediahate-speech-with-ai-powered-bots/5a22dfa327b1

Lima, L., Reis, J. C., & Melo, P. (2018). Inside the right-leaning echo chambers: Characterizing gab, an unmoderated social system. In *2018 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)* (pp. 515–522). IEEE.

Lingiardi, V., Carone, N., Semeraro, G., Musto, C., D'Amico, M., & Brena, S. (2019). Mapping Twitter hate speech towards social and sexual minorities: a lexicon-based approach to semantic content analysis. *Behaviour and Information Technology*, 1–11.

Liu, S., & Forss, T. (2014). Combining n-gram based similarity analysis with sentiment analysis in web content classification. In *Proceedings of the International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management*, Vol. 1. (pp. 530–537). Setúbal: SciTePress.

(2015). New classification models for detecting Hate and Violence web content. In *Proceedings of the 7th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management (IC3K 2015)*, Vol. *1*. (pp. 487–495). Lisbon: IEEE.

Lizza, R. (2016). Twitter's anti-Semitism problem. *The New Yorker*, October 19. www.newyorker.com/news/news-desk/twitters-anti-semitism-problem

Magdy, W., Darwish, K., Abokhodair, N., Rahimi, A., & Baldwin, T. (2016). # isisisnotislam or # deportallmuslims? Predicting unspoken views. In *Proceedings of the 8th ACM Conference on Web Science* (pp. 95–106). New York: ACM.

Magu, R., Joshi, K., & Luo, J. (2017). Detecting the hate code on social media. arXiv.org. https://arxiv.org/abs/1703.05443

Mariconti, E., Suarez-Tangil, G. Blackburn, J. et al. (2018). "You know what to do": Proactive detection of YouTube videos targeted by coordinated hate attacks. arXiv.org. https://arxiv.org/abs/1805.08168

Marwick, A. (2017). Are there limits to online free speech? Data & Society Research Institute (blog), January 5. https://points.datasociety.net/are-_there-_limits-_to-_online-_free-_speech-_14dbb7069aec

Marwick, A., & Lewis, R. (2017). *Media Manipulation and Disinformation Online*. New York: Data & Society Research Institute.

Mathew, B., Kumar, N., Goyal, P., & Mukherjee, A. (2018). Analyzing the hate and counter speech accounts on Twitter. arXiv.org. https://arxiv.org/abs/1812.02712

Mathew, B., Dutt, R., Goyal, P., & Mukherjee, A. (2019). Spread of hate speech in online social media. In *Proceedings of the 10th ACM Conference on Web Science* (pp. 173–182). New York: ACM.

Mathew, B., Saha, P., & Tharad, H. et al. (2019). Thou shalt not hate: Countering online hate speech. arXiv.org. https://arxiv.org/abs/1808.04409

Matsuda, M. J. (2018). *Words That Wound: Critical Race Theory, Assaultive Speech, and the First Amendment*. London: Routledge.

McMillin, S. E. (2014). Ironic outing: The power of hate group designations to reframe political challenges to LGBT rights and focus online advocacy efforts. *Journal of Policy Practice*, *13*(2), 85–100.

McNamee, L. G., Peterson, B. L., & Peña, J. (2010). A call to educate, participate, invoke and indict: Understanding the communication of online hate groups. *Communication Monographs*, 77(2), 257–280.

Meza, R. M. (2016). Hate-speech in the Romanian online media. *Journal of Media Research*, 9(3), 55.

Mossie, Z., & Wang, J.-H. (2018). Social network hate speech detection for Amharic language. Paper presented at the Fourth International Conference on Natural Language Computing (NATL). April 28–29, Dubai, UAE.

Muller, K., & Schwarz, C. (2017). Fanning the flames of hate: Social media and hate crime. SSRN. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3082972

Muller, K., & Schwarz, C. (2019). Making America hate again? Twitter and hate crime under Trump. SSRN. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3149103

Munger, K. (2017). Tweetment effects on the tweeted: Experimentally reducing racist harassment. *Political Behavior*, *39*(3), 629–649.

Nedig, H. (2017). Twitter launches hate speech crackdown. *The Hill*. December, 18. https://thehill.com/policy/technology/365424-twitter-to-begin-enforcing-new-hate-speech-rules

Newell, E., Jurgens, D., Saleem, H. M. et al. (2016). User migration in online social networks: A case study on Reddit during a period of community unrest. In *Proceedings of the Tenth International AAAI Conference on Web and Social Media* (pp. 279–288). Palo Alto, CA: AAAI Press.

Olteanu, A., Castillo, C., Boy, J., & Varshney, K. R. (2018). The effect of extremist violence on hateful speech online. arXiv.org. https://arxiv.org/abs/1804.05704

Paluck, E. L. (2009). Reducing intergroup prejudice and conflict using the media: A field experiment in Rwanda. *Journal of Personality and Social Psychology*, *96*(3), 574–587.

Parekh, B. (2012). Is there a case for banning hate speech? In M. Herz & P. Molnar (Eds.), *The Content and Context of Hate Speech: Rethinking Regulation and Responses* (pp. 37–56). Cambridge: Cambridge University Press.

Parenti, M. (2013). Extreme right organizations and online politics: A comparative analysis of five Western democracies. In P. Nixon, R. Rawal, & D. Mercea (Eds.), *Politics and the Internet in Comparative Context* (pp. 155–173). London: Routledge.

Phillips, W. (2015). *This Is Why We Can't Have Nice Things: Mapping the Relationship Between Online Trolling and Mainstream Culture*. Cambridge, MA: MIT Press.

Pierskalla, J. H., & Hollenbach, F. M. (2013). Technology and collective action: The effect of cell phone coverage on political violence in Africa. *American Political Science Review*, *107*(2), 207–224.

Posner, R. A. (2001). The speech market and the legacy of Schenck. In G. R. Stone & L. C. Bollinger (Eds.), *Eternally Vigilant: Free Speech in the Modern Era* (pp. 121–152). Chicago: University of Chicago Press.

Potok, M. (2015). *The Year in Hate and Extremism*. Southern Poverty Law Center intelligence report. www.splcenter.org/fighting-hate/intelligence-report/2015/year-hate-and-extremism-0

Ribeiro, M. H., Calais, P. H., Santos, Y. A., Almeida, V. A. F., & Meira, W., Jr. (2018). Characterizing and detecting hateful users on Twitter. arXiv.org. https://arxiv.org/pdf/1803.08977.pdf

Richards, R. D., & Calvert, C. (2000). Counterspeech 2000: A new look at the old remedy for bad speech. *BYU Law Review*, *2000*(2), 553–586.

Ross, B., Rist, M., Carbonell, G., Cabrera, B., Kurowsky, N., & Wojatzki, M. (2017). Measuring the reliability of hate speech annotations: The case of the European refugee crisis. arXiv.org. https://arxiv.org/abs/1701.08118

Saha, K., Chandrasekharan, E., & De Choudhury, M. (2019). Prevalence and psychological effects of hateful speech in online college communities. In *Proceedings of the 11th ACM Conference on Web Science* (pp. 255–264). New York: ACM.

Saleem, H. M., Dillon, K. P., Benesch, S., & Ruths, D. (2017). A web of hate: Tackling hateful speech in online social spaces. arXiv.org. https://arxiv.org/abs/1709.10159

Saleem, H. M., & Ruths, D. (2019). The aftermath of disbanding an online hateful community. arXiv.org. https://arxiv.org/pdf/1804.07354.pdf

Salminen, J., Almerekhi, H., Kamel, A. M., Jung, S.-g., & Jansen, B. J. (2019). Online hate ratings vary by extremes: A statistical analysis. In *Proceedings of the 2019 Conference on Human Information Interaction and Retrieval* (pp. 213–217).

Santosh, T. Y. S. S., & Aravind, K. V. S. (2019). Hate Speech Detection in Hindi-English Code-Mixed Social Media Text. In *Proceedings of the ACM India Joint International Conference on Data Science and Management of Data* (pp. 310–313). New York: ACM.

Schieb, C., & Preuss, M. (2016). Governing hate speech by means of counterspeech on Facebook. Paper presented at the 66th Annual Conference of the International Communication Association: Communicating with Power, June 9–13, Fukuoka, Japan.

Schmidt, A., & Wiegand, M. (2017). A survey on hate speech detection using natural language processing. In *Proceedings of the 5th International Workshop on Natural Language Processing for Social Media* (pp. 1–10). Stroudsburg, PA: Association for Computational Linguistics.

Selepak, A. (2010). Skinhead Super Mario Brothers: An examination of racist and violent games on White supremacist web sites. *Journal of Criminal Justice and Popular Culture*, *17*(1), 1–47.

Sellars, A. (2016). Defining hate speech. SSRN. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2882244

Siapera, E., Moreo, E., & Zhou, J. (2018). *Hate Track: Tracking and Monitoring Online Racist Speech*. Dublin: Irish Human Rights and Equality Commission.

Siegel, A. (2015). *Sectarian Twitter Wars: Sunni-Shia Conflict and Cooperation in the Digital Age*, Vol. 20. Washington, DC: Carnegie Endowment for International Peace.

Siegel, A., Nitikin, E., & Barberá, P. (2020). *Trumping Hate on Twitter: Online Hate Speech in the 2016 Presidential Election Campaign and Its Aftermath*. Unpublished manuscript.

Siegel, A., Tucker, J., Nagler, J., & Bonneau, R. (2018). *Socially Mediated Sectarianism*. Unpublished manuscript.

Siegel, A., & Badaan, V. (2020). *No2Sectarianism: Experimental Approaches to Reducing Online Hate Speech*. Forthcoming in the American Political Science Review.

Silva, L., Mondal, M., Correa, D., Benevenuto, F., & Weber, I. (2016). Analyzing the targets of hate in online social media. arXiv.org. https://arxiv.org/abs/1603.07709v1

Sonnad, N. (2016). Alt-right trolls are using these code words for racial slurs online. *Quartz*, October 1. https://qz.com/798305/alt-right-trolls-are-using-googles-yahoos-skittles-andskypes-as code-words-for-racial-slurs-on-twitter/

Soral, W., Bilewicz, M., & Winiewski, M. (2018). Exposure to hate speech increases prejudice through desensitization. *Aggressive Behavior*, 44(2), 136–146.

Staub, E., Pearlman, L. A., & Miller, V. (2003). Healing the roots of genocide in Rwanda. *Peace Review*, *15*(3), 287–294.

Stecklow, S. (2018). *Why Facebook Is Losing the War on Hate Speech in Myanmar*. Reuters Special Report, August 15. www.reuters.com/investigates/special-report/myanmar-facebook-hate/

Stephens-Davidowitz, S. (2017). *Everybody Lies: Big Data, New Data, and What the Internet Can Tell Us About Who We Really Are*. New York: HarperCollins.

Tsesis, A. (2002). *Destructive Messages: How Hate Speech Paves the Way For Harmful Social Movements*, Vol. 778. New York: New York University Press.

Tuckwood, C. (2014). The state of the field: Technology for atrocity response. *Genocide Studies and Prevention: An International Journal*, *8*(3), 81–86.

Twitter. (2017). Twitter Rules and Policies. https://help.twitter.com/en/rules-and-policies/violent-groups

(2018). The Twitter Rules. https://support.twitter.com/articles/18311

Tynes, B. M., Giang, M. T., Williams, D. R., & Thompson, G. N. (2008). Online racial discrimination and psychological adjustment among adolescents. *Journal of Adolescent Health*, *43*(6), 565–569.

Tynes, B. M., & Markoe, S. L. (2010). The role of color-blind racial attitudes in reactions to racial discrimination on social network sites. *Journal of Diversity in Higher Education*, *3*(1), 1–13.

Unsvåg, E. F., & Gambäck, B. (2018). The effects of user features on Twitter hate speech detection. In *Proceedings of the 2nd Workshop on Abusive Language Online (ALW2)* (pp. 75–85). Stroudsburg, PA: Association for Computational Linguistics.

Van Hee, C., Lefever, E., Verhoeven, B. et al. (2015). Detection and fine-grained classification of cyberbullying events. In *International Conference Recent Advances in Natural Language Processing (RANLP)* (pp. 672–680). Shumen: INCOMA.

Vidino, L., & Hughes, S. (2015). *ISIS in America: From Retweets to Raqqa*. Program on Extremism, George Washington University report. https://extremism.gwu.edu/sites/g/files/zaxdzs2191/f/downloads/ISIS%20in%20America%20-%20Full%20Report.pdf

Vindu G., Kumar, H., & Frenkel, S. (2018). In Sri Lanka, Facebook contends with shutdown after mob violence. *New York Times*, March 8. www.nytimes.com/2018/03/08/technology/sri-lanka-facebook-shutdown.html

Vollhardt, J., Coutin, M., Staub, E., Weiss, G., & Deflander, J. (2007). Deconstructing hate speech in the DRC: A psychological media sensitization campaign. *Journal of Hate Studies*, *5*(15), 15–35.

Warner, W., & Hirschberg, J. (2012). Detecting hate speech on the World Wide Web. In S. Owsley Sood, M. Nagarajan, & M. Gamon (Eds.), *Proceedings of the Second Workshop on Language in Social Media* (pp. 19–26). New York: ACL.

Waseem, Z. (2016). Are you a racist or am I seeing things? Annotator influence on hate speech detection on Twitter. In D. Bamman, A. Seza Doğruöz, J. Eisenstein et al. (Eds.), *Proceedings of the First Workshop on NLP and Computational Social Science* (pp. 138–142). Stroudsburg, PA: Association for Computational Linguistics.

Waseem, Z., & Hovy, D. (2016). Hateful symbols or hateful people? Predictive features for hate speech detection on Twitter. In M. Sahlgren & O. Knutsson (Eds.), *Proceedings of the NAACL HLT Workshop on Extracting and Using Constructions in Computational Linguistics* (pp. 88–93). Stroudsburg, PA: Association for Computational Linguistics.

Weaver, S. (2013). A rhetorical discourse analysis of online anti-Muslim and anti-Semitic jokes. *Ethnic and Racial Studies*, *36*(3), 483–499.

Weidmann, N. B. (2009). Geography as motivation and opportunity: Group concentration and ethnic conflict. *Journal of Conflict Resolution*, *53*(4), 526–543.

(2015). Communication networks and the transnational spread of ethnic conflict. *Journal of Peace Research*, *52*(3), 285–296.

Williams, M. L., Burnap, P., Javed, A., Liu, H., & Ozalp, S. (2019). Hate in the machine: Anti-Black and anti-Muslim social media posts as predictors of offline racially and religiously aggravated crime. *The British Journal of Criminology*, *60*(1), 93–117.

Yanagizawa-Drott, D. (2014). Propaganda and conflict: Evidence from the Rwandan genocide. *The Quarterly Journal of Economics*, *129*(4), 1947–1994.

YouTube. (2018). Community Guidelines. www.youtube.com/yt/ policyandsafety/communityguidelines.html

Yuan, S., Xintao W., & Xiang, Y. (2016). A two phase deep learning model for identifying discrimination from tweets. In *EDBT: 19th International Conference on Extending Database Technology* (pp. 696–697). OpenProceedings.org.

Zannettou, S., Bradlyn, B., De Cristofaro E. et al. (2018). What is Gab? A bastion of free speech or an alt-right echo chamber? arXiv.org. https://arxiv.org/abs/1802.05287

Zhang, Z., Robinson, D., & Tepper, J. (2018). Detecting hate speech on Twitter using a convolution-GRU based deep neural network. In A. Gangemi, R. Navigli, M.-E. Vidal et al. (Eds.), *The Semantic Web: ESWC 2018* (pp. 745–760). Cham: Springer.