Political Psychology

ISPP

Advances in Political Psychology, Vol. 41, Suppl. 1, 2020
doi: 10.1111/pops.12670

Hate Speech Epidemic. The Dynamic Effects of Derogatory Language on Intergroup Relations and Political Radicalization

Michał Bilewicz
University of Warsaw

Wiktor Soral
University of Warsaw

Exposure to derogatory language about immigrants and minority groups leads to political radicalization and deteriorates intergroup relations. This article addresses the psychological processes responsible for these effects as well as those involved in hate-speech proliferation in contemporary societies and discusses the factors that constrain its growth. We propose that frequent exposure to hate speech has severe effects on emotional, behavioral, and normative levels. Exposure to hate speech results in empathy being replaced by intergroup contempt as a dominant response to others—it is both a motivator and a consequence of derogatory language. The increased presence of hate speech in one's environment creates a sense of a descriptive norm that allows outgroup derogation. This leads to the erosion of existing antidiscriminatory norms. Finally, through a process of desensitization, hate speech reduces people's ability to recognize the offensive character of such language. Based on empirical evidence from the fields of social psychology and psychology of emotion and aggression, we propose a model that explains the described processes, and we trace the dynamics of this model using an agent-based modeling approach. We show that the mechanisms potentially effective in constraining hate-speech proliferation (empathy, norms) are eroded by hate speech itself. We argue that through basic psychological dynamics, societies become more accepting of derogatory language and less accepting of immigrants, as well as religious, ethnic, and sexual minorities.

KEY WORDS: hate speech, contempt, empathy, social norms, prejudice

"More troops being sent to the Southern Border to stop the attempted invasion of illegals, through large caravans, into our country"; "the U.S. is ill-prepared for this invasion, and will not stand for it. They are causing crime and big problems in Mexico. Go home!"; "Allowing the immigration to take place in Europe is a shame . . . I think it changed the fabric of Europe and, unless you act very quickly, it's never going to be what it was and I don't mean that in a positive way": These statements by the U.S. President Donald J. Trump have been widely discussed in the context of anti-immigrant and xenophobic terror acts, such as the 2019 El Paso shooting in which 22 people were killed, the 2019 Christchurch mosque shooting in New Zealand where 51 people were killed and 49 were injured, and the 2018 Pittsburgh synagogue shooting in which 11 people were killed and seven were injured.

Published by Wiley Periodicals, Inc., 350 Main Street, Malden, MA 02148, USA, 9600 Garsington Road, Oxford, OX4 2DQ,
and PO Box 378 Carlton South, 3053 Victoria, Australia

The perpetrators of all these crimes were heavily exposed to anti-immigrant hate speech and used hate speech as a justification for their actions. “HIAS [American Jewish organization] likes to bring in invaders that kill our people. I can’t sit by and watch my people get slaughtered. Screw your optics, I’m going in,” wrote the gunman from Pittsburgh in his post on social media shortly before the attack (Levenson & Sanchez, 2018), whereas the perpetrator of the El Paso crime wrote in his manifesto: “This attack is a response to the Hispanic invasion of Texas. . . . The natives didn’t take the invasion of Europeans seriously, and now what’s left is just a shadow of what was” (Aratani, 2019).

The effects of people’s exposure to hate speech (for example, portraying immigrants and refugees as “invaders”) on their racism and the tendency to engage in acts of intergroup violence or political radicalization have been a primary focus of psychological theorizing for decades. Gordon W. Allport (1954), in his seminar work on prejudice, describes such “antilocutions”—that is, verbal expressions of prejudice—as the first step on a 5-point continuum. He suggests that antilocutions can lead to avoidance, discrimination, physical attack, and extermination, although such progression is by no means inevitable. “While many people would never move from antilocution to avoidance; or from avoidance to active discrimination, or higher on the scale, still it is true that activity on one level makes transition to a more intense level easier. It was Hitler’s antilocution that led Germans to avoid their Jewish neighbors and erstwhile friend. This preparation made it easier to enact the Nürnberg laws of discrimination which, in turn, made the subsequent burning of synagogues and street attacks upon Jews seem natural. The final step in the macabre progression was the ovens at Auschwitz” (Allport, 1954, p. 15).

Although the progression from hate speech to avoidance, discrimination, and violence against outgroup members has become almost commonsensical in social sciences, a good theoretical explanation of the mechanism of such progression is still missing. The main aim of this article is to explain how and when hate-speech exposure leads to avoidance, discrimination, and intergroup aggression. We propose a model of hate-speech epidemics that brings together theoretical insights from the social psychology of emotions (findings on contempt as an empathy-reducing sentiment), aggression (research on desensitization to verbal aggression), and political psychology (the role of norms and authority in shaping behavior). We also outline the potential mechanisms that could effectively constrain the spread of hate speech (such as norms or empathy), and we show how these mechanisms become eroded by frequent exposure to hateful language. The main purpose of the model is to better explain one of the most pressing social issues of the contemporary world—the influence of derogatory language on collective violence.

Hate Speech as a Societal Problem

Hate Speech as a Harmful Language: Definitional Issues

Although there is no formal definition of hate speech, there have been several attempts to propose a working definition of such speech acts. The Cambridge Dictionary defines hate speech as “public speech that expresses hate or encourages violence towards a person or group based on something such as race, religion, sex, or sexual orientation” (Hate speech, n.d.). The most common definitions limit the scope of hate speech to speech acts that target disadvantaged social groups in a harmful way (Jacobs & Potter, 1998; Walker, 1994). Others narrow it even further by suggesting that hate speech does not target categories but rather individuals based on their ethnic, religious, or racial identity, gender, age, disability, sexual orientation, and so forth (Sedler, 1992). The commonly accepted legal definition of hate speech proposed by the Committee of Ministers of the Council of Europe reads that “hate speech covers all forms of expressions that spread, incite, promote or justify racial hatred, xenophobia, anti-Semitism or other forms of hatred based on intolerance” (Council of

Europe, 1997). Definitional problems of hate speech have been particularly pronounced in the free speech versus hate-speech debate (Gelber, 2002).

Definitions of hate speech have to acknowledge the problem of demarcation between derogatory language and nonderogatory forms of intergroup criticism. A well-studied context in which this problem occurs is that of anti-Semitic hate speech versus the criticism of Israel. Some authors create such demarcation lines based on motivation—criticism of Israel based on peace orientation and human rights consideration is considered legitimate, whereas criticism derived from far-right prejudiced attitudes is considered hate speech (Kempf, 2012). This is problematic both in legal (problems with proving the intention of the speaker) and methodological terms (studying hate speech would be highly difficult without extensive context information). Moreover, there are studies that point to the fact that criticism of Israel might be a form of socially accepted expression of anti-Semitic views (Bulska & Winiewski, 2018; Cohen, Jussim, Harber, & Bhasin, 2009). This would make all such critical statements potential instances of hate speech.

In order to overcome the problem of hate speech versus legitimate criticism demarcation, our approach is to treat discriminated minority groups' feelings as a criterion of inclusion of specific utterances to the "hate speech" category. In our empirical research (e.g., Bilewicz, Soral, Marchlewska, & Winiewski, 2017; Soral, Bilewicz, & Winiewski, 2018; Winiewski et al., 2017), we pretested all examples of hate speech. A large list of negative statements about minority groups was presented to a sample of targeted minority members (e.g., Polish Jews, Muslims, Roma, LGBT people, etc.); they were asked whether they find certain statements offensive to their minority and consider them to be an instance of hate speech. Based on that, we selected sentences that received high absolute ratings of offensiveness and hatefulness, and these sentences were used in further experimental and correlational studies. Although this way of defining hate speech is by no means objective, it allows to overcome the problems of researcher's subjectivity and speaker's intentionality. It also captures the harm caused by the hate speech among minority groups (Jacobs & Potter, 1998).

Hate Speech Promotes Prejudice: The Majority Perspective

The early claim by Allport (1954) that hate speech can lead to outgroup avoidance and discrimination has been supported by numerous studies. Greenberg and Pyszczynski (1985; see also Kirkland, Greenberg, & Pyszczynski, 1987) found that people exposed to derogatory ethnic labels become more selective in their retrieval and reconstruction of information about black professionals' (e.g., lawyer, defense attorney) performance, and this ultimately results in a lower evaluation of such professionals. This suggests that derogatory language has the power to activate negative ethnic stereotypes. Kirkland et al. (1987) further noticed that the effects of one's exposure to derogatory language are specific to ethnic-related labels—the effects disappear when nonracist derogatory labels (e.g., "shyster") are used. However, it has to be noted that further investigation into the effects of hate speech (Simon & Greenberg, 1996) has yielded much more complex results, revealing that derogatory ethnic language affects only those individuals who have preexisting racist (anti-Black) attitudes. No effects, neither negative nor positive (i.e., more positive ratings of the outgroup), can be observed in the case of individuals with pro-Black or ambivalent attitudes. This speaks to the role of hate speech as an amplifier of already existing racist sentiments. In many cases, people hold biased views and feel negative emotions toward certain outgroups, but these views and feelings are rarely expressed. The presence of hate speech in one's environment could serve as a signal of a norm allowing individuals to publicly express negative views and emotions and to act upon them.

When thinking about hate speech and derogatory labels, one might assume that their discriminatory effects are driven mostly by the emotional negativity of such words. However, both archival (Mullen & Rice, 2003) and experimental (Leader, Mullen, & Rice, 2009) studies have found that hate speech is driven not only by the negativity of the terms used to describe ethnic minorities, but also by the simplicity in their portrayal (low complexity of the terms used). Majority group members tend to avoid and discriminate against those minorities who are described in extremely simplified ways.

Research looking at the hate-speech–prejudice link is not restricted to the study of racial and ethnic contexts of discrimination. Over the last decades, more attention has been paid to homophobic language (*gay bashing*, *gay bullying*) as one of the most frequent forms of verbal derogation. The study on the effects of homophobic labels by Carnaghi and Maass (2007) revealed that, among heterosexual males, implicitly primed derogatory terms (*fag, fairy*) activate more negative associations than category labels (*gay, homosexual*). Another study (Carnaghi & Maass, 2008) revealed that such derogatory terms result in slower approach responses than category labels (no differences were found in the case of avoidance responses). Finally, Fasoli et al. (2016) found that subliminal exposure to homophobic labels increases dehumanization of homosexuals (i.e., decreases people's ascription of human-related words, such as *person* or *culture,* to gay people, in proportion to animal-related words, such as *herd* or *nature*) and physical distancing from them (i.e., sitting further apart when meeting a homosexual person) in comparison to neutral category labels and generic insults. These findings suggest that the link between hate speech and prejudice can be observed not only in conscious verbal tasks, but also on a more implicit and automatic level.

Further studies have also seemed to confirm the next steps proposed by Allport (1954) in his progression model: that hate speech leads not only to avoidance, but also to discrimination. Fasoli, Maass, and Carnaghi (2015) found that people exposed to homophobic labels tend to discriminate against gay people more in the allocation of resources (i.e., they allocate less money to LGBT-targeted AIDS prevention programs, as compared to programs directed toward the straight community). A recent study of Polish youth (Winiewski et al., 2017) found that people who are more exposed to hate speech not only tend to avoid ethnic minorities in their close environment (this has been shown in the case of Islamophobic, sexist, anti-immigrant, anti-Gypsy, anti-Ukrainian and trans-phobic language), but that they are also more supportive of harsh treatment of immigrants, including: isolation of refugees, closing national borders, refusing assistance for immigrants in need, and using invigilation methods that violate human rights. These results suggest that Allport's predictions were true: Hate speech leads to avoidance, and, in more extreme cases, it might justify acts of discrimination and physical violence.

Hate Speech Deteriorates Well-Being: The Minority Perspective

Ethnic, racial, religious, and sexual minorities often suffer from psychological problems related to their minority status. A classic study of native and foreign students in the United Kingdom (Still, 1961) observed that 14% of native British students had psychological problems, compared with 28% of Iraqi, Iranian, and Nigerian students, 22% of Turkish and Egyptian students, and 18% of students from India, Bangladesh, Pakistan, and Nepal. Similarly, sexual-minority youth report significantly higher levels of depression, suicidal ideations, and suicide attempts than heterosexual youth (Marshal et al., 2011). This discrepancy in mental health and psychological well-being between majority and minority groups could be attributed to many factors, such as low self-esteem, family rejection, minority stress, or pressure to conform. However, most researchers point to the experiences of victimization, bullying, and hate speech as possible causes of lower well-being among minority-group members (see Marshal et al., 2011).

The existing evidence showing the negative mental health effects of being exposed to hate speech includes archival, experimental, and survey/cross-sectional studies. Archival data analyses conducted by Mullen and Smyth (2004) revealed that immigrant minorities' suicide rate is strongly predicted by the negativity of the ethnophaulisms used to describe the minority (which was derived from the historical record of hate speech in the United States). An Italian study, in which gay people were exposed to homophobic labels (Bianchi, Piccoli, Zotti, Fasoli, & Carnaghi, 2017) found that homosexual individuals (particularly those who were open with their identity) reported higher levels of internalized homophobia after being confronted with such labels. Interestingly, the reverse pattern was observed for those participants who were hiding their gay identity. A recent large-scale study of international students at the University of Warsaw found that those students who were exposed to hate speech during their studies in Poland reported significantly more negative moods, lower psychological well-being, and higher levels of post-traumatic stress disorder (Kuzawińska et al., 2018).

In a nation-wide study of LGBT minorities in Poland, we asked gay and lesbian individuals about the frequency of contact with antigay or antilesbian hate speech (see Soral, Bilewicz, Winiewski, & Bulska, 2020). First, we presented participants with real-life examples of online hate speech. Then we asked participants to rate how frequently they encounter similar statements directed against gays or lesbians (depending on the individual's identity). We also measured eight symptoms of depression and asked about participants' frequency of suicidal thoughts. Our analyses revealed that the more the participants declared they encountered antigay or antilesbian hate speech, the more severe depressive symptoms they reported and the more frequently they thought about suicide. Interestingly, the strength of this association was significantly weaker among individuals declaring higher levels of acceptance among family members. Thus, our result provides some support for the hypothesis that affirmation of one's social identity may counteract the negative effects of hate speech (see, e.g., Haslam, Jetten, Postmes, & Haslam, 2009).

When analyzing the consequences of hate speech for minority groups, Leets (2002) proposed using the three-stage sequence model adapted from clinical research on traumatic experiences: disorganization, recoil, and reorganization. In the first stage, minority-group members suffer from short-term emotional crises. In the second stage, they form negative attitudes toward the group that delivered hate speech. In the third stage, they change their behavior by avoiding certain places and encounters with sources of hate speech. Most importantly, Leets (2002) found that minority group members targeted by hate speech attribute this experience to stable and enduring motives (rather than to situational causes). This ultimately enhances the divide between majority and minority group members and leads to large-scale grievances in intergroup relations.

The Epidemic Model of Hate Speech: Understanding the Dynamics

Hate speech has severe consequences for human intergroup relations and should be considered a large-scale societal issue that deteriorates living quality, increases aggression, and affects mental health and well-being of minorities. At the same time, hate speech is common in political discourse and its presence in social media has increased (Winiewski et al., 2017). In this article, we would like to propose a model that could explain the proliferation of hate speech and the effects of hate speech on contemporary societies. The model aims to incorporate findings and theorizing from social psychology of aggression, emotion, and political psychology into a potential explanation of the current epidemics of hate speech.

Hate speech is a violent language that evokes suffering among targeted groups. Most people react to such utterances with negative emotional arousal due to a "shared pain" network—a fundamental neurophysiological mechanism that forms the basis for the phenomenon of emotional empathy (Bruneau, Pluta, & Saxe, 2012; Mischkowski, Crocker, & Way, 2016). This, in turn, leads to negative responses to hate speech—willingness to stand up to such language, to punish those who use it, and to control its presence in the public sphere. Furthermore, the social norm prescribing that hateful language is socially undesirable constrains people's use of hate speech and limits its proliferation. In addition, strong social norms motivate collective action against hate speech, particularly among people who endorse such norms and consider hate speech as a source of injustice (van Zomeren, Postmes, & Spears, 2008).

Table 1. Consequences of Hate Speech in Intergroup Relations and Online Behavior

Response	Behavioral Processes	Normative Processes	Emotional Processes
Default response	• Sensitivity to hatred • Ability to recognize derogatory language and to motivate reactions to such language	• Prescriptive antidiscriminatory norms, followed mostly by normocentric individuals	• Empathy • Emotional arousal when exposed to derogatory language
Response after frequent exposure to hate speech	• Lack of sensitivity to derogatory language and to the usage of hate speech in everyday conversations and in online behavior • Avoidance • Discrimination	• Descriptive discriminatory norms • Derogation motivated by leaders (prescriptive norms)	• Contempt • Failure to empathize • Schadenfreude • Dehumanization • Lack of emotional arousal when exposed to hate speech and prejudice

Nonetheless, in the course of being exposed to hate speech, people's sensitivity to hateful language diminishes. The more hate speech people observe in their environment, the less emotional arousal they tend to feel. Moreover, frequent exposure to hate speech changes the dominant image of the outgroup targeted by such language. Being frequent targets of verbal aggression, minority group members become increasingly viewed as inferior to one's ingroup due to system justification mechanisms (Jost, 2019; Jost & Hunyady, 2003, 2005). This might modify the dominant emotional response to such groups into one of contempt. Contempt constrains empathic responses and leads to distancing from outgroup members as well as to proliferation of derogatory language (Cuddy, Fiske, & Glick, 2007). Therefore, such derogatory language should perhaps be considered "speech of contempt" rather than "hate speech," as the emotion of contempt is both the focal antecedent and outcome of such language (Bilewicz, Kamińska, Winiewski, & Soral, 2017). Finally, frequent exposure to hate speech might result in an emergence of a new norm of conduct—people might start perceiving such language as morally justified and legitimate. This could occur particularly in a situation in which important others, such as political or religious leaders, use derogatory language about minorities or immigrants.

The processes described above were the main topic of our studies that looked at the emergence of hate speech, particularly due to increased use of social media as a primary source of people's political knowledge (see also Table 1). In the next part of this article, we will outline the main findings of this research, while also discussing some potential countering mechanisms (e.g., evolution of antidiscriminatory norms, exposure to empathic messages) that could resensitize people to hate speech and constrain the presence of derogatory language in the public sphere. We will start by addressing the behavioral processes (desensitization) and then discuss the normative (normative changes affected by media consumption) and the emotional ones (dehumanization and the emergence of contempt).

Mechanisms Catalyzing the Spread of Hate Speech

Desensitization to Hate Speech

Humans who repeatedly encounter an emotional stimulus begin to display a weaker response to the stimulus, and this decrease in the response is a function of the number of stimulus presentations. This very basic behavioral phenomenon, named habituation, has been observed not only among humans, but also in nonhuman animals (Rankin et al., 2009; Thompson & Spencer, 1966). Experimental studies in which people were exposed to emotionally valenced faces showed that, after frequent exposure to such images, people's emotional reactions (amygdala activity) decreased (Breiter et al., 1996). Racial slurs, homophobic speech, or anti-Semitic tirades are highly emotional utterances that produce strong emotional stimulation. What happens when they are omnipresent? Do people become habituated to such language in a similar way as to any other strongly emotional stimulus?

The general aggression model (Anderson & Carnagey, 2004; Carnagey, Anderson, & Bartholow, 2007) proposes that repeated exposure to violent behavior leads to desensitization—i.e., people react less emotionally to subsequent acts of violence. This affects different cognitive and behavioral responses to violence: People who are desensitized perceive others' injuries as less severe, their attention to violent events and sympathy to the victims decrease, and they start believing that the violence is a norm. Desensitization has been observed at very basic physiological levels of human functioning: People frequently exposed to violence do not react with an increased heart rate (Linz, Donnerstein, & Adams, 1989) or decreased skin conductance (Cline, Croft, & Courrier, 1973) when observing subsequent acts of violence. Further, playing a violent videogame for a prolonged period of time (i.e., "Call of Duty" as compared to "Pro-skater"), results in a decrease in brain activation in response to violent images as demonstrated using encephalographic analysis (P300 event-related potential, Bartholow, Bushman, & Sestir, 2006).

In our studies of online hate speech (Soral et al., 2018), we decided to analyze whether a similar desensitization process occurs when people are immersed in a realm of online hate speech. Specifically, we tested the idea that repeated exposure to hate speech directed at various minorities would change the perceived offensiveness of similar statements. In an experimental study (Study 2), we asked participants to read either a long list of hate-speech comments (experimental condition) or a list of simply negative comments (control condition). The experiment was presented as a study on the effects of a layout of a discussion forum on the ability to memorize its content. Thus, the participants were asked to memorize the content of each page and to rate its graphical properties. In reality, the aim was to expose participants to a large amount of hate speech (vs. simply negative) content. Then, the participants rated a new list of hate-speech comments in terms of offensiveness. At the end of the study, we measured social distance toward the minorities mentioned in the hateful comments. The analyses revealed that those participants who were exposed to hate speech were less sensitive to derogatory language (i.e., they perceived hateful comments as less offensive) than the participants exposed to the simply negative content that did not include hate speech. Moreover, the participants exposed to hate speech reported a higher social distance from minorities mentioned in the hateful statements than the nonexposed participants. Finally, sensitivity to hate speech acted as a mediating variable between exposure to hate speech and social distance. These results were corroborated in analyses of a social survey of Polish adults and adolescents (Study 1 and Study 3), where we found a similar mediating process of desensitization. In addition, we found that one's exposure to hate speech is related not only to increased social distance, but also to greater support for radical anti-immigrant policies and harsh treatment of immigrants (see Figure 1). We also found that these effects cannot be explained simply by a change in one's perception of social norms regarding aggression.

Normativity of Hate Speech in Digital Media

Although the desensitization effects observed among people frequently exposed to hate speech cannot be fully explained by the normative changes, it is obvious that the mere frequency of such utterances in people's environment can produce a sense of normativity of hatred. Research on social norms (e.g., Brauer & Chaurand, 2010; Cialdini, Reno, & Kallgren, 1990) distinguishes two kinds of norms: prescriptive (i.e., judgment of whether a given behavior is desirable or undesirable) and descriptive ones (i.e., judgment of whether a given behavior is common in one's environment). This latter category of social norms seems to explain the phenomenon of hate-speech epidemics well. Simply because of the enormous frequency of derogatory language, people might get a sense that hate speech is a norm rather than a delinquent behavior. We would like to propose that this

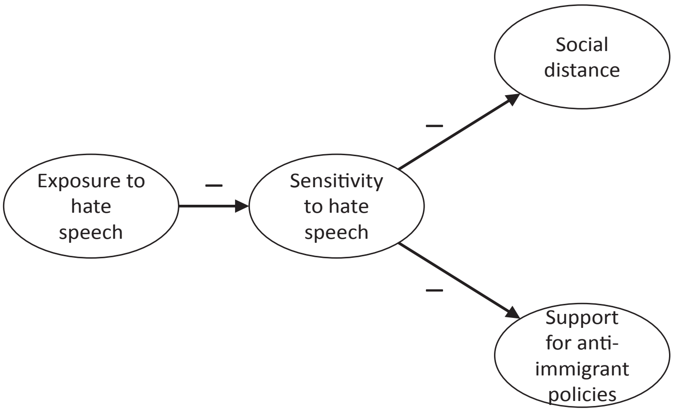
Social distance
Exposure to hate speech
–
Sensitivity to hate speech
–
–
Support for anti-immigrant policies

Figure 1. Exposure to hate speech increases outgroup prejudice and anti-immigrant attitudes through a decrease in sensitivity to hate speech. The sign '–' indicates a negative relationship. Adapted from Soral et al. (2018).

normalization process is an effect of rapid changes regarding media—the fact that people move from traditional media to social media as their primary source of information about politics and social problems.

A number of previous reports indicated that hate speech is most prevalent on social media, such as Facebook or Twitter. Recent surveys have revealed that as much as 85% of Polish adolescents had encountered hate speech on the Internet, whereas only about 45% had encountered such content on TV, and only 12% in newspaper articles (Winiewski et al., 2017). Reports from studies conducted in Finland reveal that 67% of adolescent Facebook users had been exposed to hate speech online, and 21% declared being victimized by online hate speech (Oksanen, Hawdon, Holkeri, Näsi, & Räsänen, 2014). Other studies (Hawdon, Oksanen, & Räsänen, 2017) reveal that exposure to online hate among adolescents and young adults vary from 53% in the United States to 31% in Germany, with the differences explained to some extent by the existence of anti-hate-speech laws.

Given the prevalence of hate speech in the digital media, it is likely that the normalization of hate speech would be observed mostly among individuals who use social media, or other digital forms, to gather knowledge about the world. Such individuals would also manifest higher levels of prejudice, and such an increase should be mediated by changes in their perception of hate speech. We examined this prediction in a study placed in the context of the so-called "refugee crisis" and tested how the use of digital media affects people's perception of Islamophobic hate speech and how this in turn affects their levels of Islamophobia (Soral, Liu, & Bilewicz, in press). In the first phase of the study, we classified participants based on the profiles of media use—we asked specifically about the key sources they use to obtain news about the world. We observed three different profiles of media users: traditional media users (with frequent use of TV, newspapers, and radio, but relatively low use of social media or citizen journalism sites), digital media users (with frequent use of social media, and citizen journalism sites, but relatively low usage of TV, newspapers, and radio), and highly engaged users (declaring frequent use of all types of media). In the second phase of the study, we compared levels of normative acceptance of anti-Muslim hate speech as well as Islamoprejudice across three different profiles of users. We observed that digital media users accepted anti-Muslim hate speech significantly more than highly engaged and traditional users. Also, digital media users manifested significantly higher levels of Islamoprejudice than highly engaged or digital media users. Finally, the relationship between media usage and Islamoprejudice was mediated by an increased acceptance of anti-Muslim hate speech.

Our study of media consumption (Soral et al., in press) showed that people for whom digital media is a primary source of information about politics consider hate speech to be a social norm rather than delinquent behavior. This is probably because digital media is relatively unregulated, and hateful comments are much more common in online communication—as compared to traditional media. A large frequency of derogatory language in people's online environment creates a descriptive norm that defines hate speech as common, and therefore normative. Ultimately, this process enhances prejudice and makes biased views of minorities socially accepted.

The results of our study (Soral et al., in press) indicate that descriptive (and not prescriptive) norms can very well explain the phenomenon of hate-speech epidemics. However, other research ventures might also explain how prescriptive norms make hate speech omnipresent. It should be noted that computer-mediated communication provides a psychological sense of anonymity. Although typical Internet users are almost never anonymous, the possibility of not having to disclose identifiable details may lead some individuals to believe that they cannot be traced. Although classical theories of normative influence (e.g., Deutsch & Gerard, 1955; Short, Williams, & Christie, 1976) argued that identifiability is a critical factor that facilitates social influence, more recent literature suggests that under some circumstances lack of identifiability may also promote adherence to social norms (Reicher, Spears, & Postmes, 1995). Particularly when social identity is made salient, anonymity and deindividuation can enhance the effects of an ingroup's prescriptive norms. For example, Postmes, Spears, Sakhel, and de Groot (2001) primed their participants with efficiency or prosocial norms under conditions of anonymity or identifiability. Prosocial primes resulted in more prosocial behavior, and efficiency primes resulted in more efficiency-oriented behavior only under conditions of anonymity and deindividuation. Furthermore, these effects seem to be mediated by an increased identification with an ingroup. Therefore, perceiving oneself as a group member rather than an individual makes one behave in line with what he or she believes other group members expect (i.e., not only what behavior is common among group members, but also what behavior is desired).

The effects of anonymity and deindividuation can be evident in the case of radicalized, xenophobic online discussion groups. Such groups draw—usually anonymously—individuals attracted by extremist—usually right-wing—ideologies and are used to share hateful content on immigrants and those of a different race, ethnicity, or sexual orientation. Such discussion groups—with their user anonymity and salient extremist social identity—provide optimal conditions for prescriptive social norms to affect individuals' behavior (i.e., use of hate speech). A recent investigation of one such online xenophobic forum (Bäck, Bäck, Gustafsson Sendén, & Sikström, 2018) revealed that, with increasing participation in such a collective, users' use of the word "I" decreases, while the use of the word "We" increases. This could mean that, with time, users of this xenophobic forum become more deindividualized and prone to define themselves in terms of their shared social identity. Moreover, the linguistic style of new users over time becomes more similar to the linguistic style of the whole forum. Such a change might be a result of salient prescriptive group norms shared by users of the forum and new users' adjustment to the social norms that dominate within that group.

Overall, it can be argued that descriptive social norms will affect the spread of hate speech among the mainstream of Internet users, whereas prescriptive social norms will affect the spread of hate speech among radicalized minorities of Internet users. The effects of descriptive social norms can be explained by normalization of hate speech. This process somewhat resembles the well-known "broken windows" effect (e.g., Keizer, Lindenberg, & Steg, 2008), where signs of petty criminal behavior trigger more such behavior, thus causing the spread of crime. Contrarily, the effects of prescriptive social norms can be explained by deindividuation and increased identification with extremist and xenophobic social groups.

Hate Speech Modifies Stereotype Content and Generates Dehumanization

Probably the most destructive aspect of hate speech is its consequence for intergroup emotions and stereotypes. The authors of the stereotype-content model (Cuddy et al., 2007; Fiske, Cuddy, Glick, & Xu, 2002) proposed that stereotypical images of outgroups are organized across two fundamental dimensions: competence and warmth. The first includes characteristics indicative of group's power and status (agency, competence, dominance), and the second includes characteristics indicative of benevolence (warmth, communion, and nurturance). Stereotypes of a given outgroup determine both emotions and behavioral intentions toward that outgroup. For example, groups that are perceived as cold and competent would elicit envy, and such groups would often be accused of conspiring against the ingroup (Glick, 2002; Winiewski, Soral, & Bilewicz, 2015). On the contrary, groups perceived as cold and incompetent elicit contempt, and this stereotype is often regarded as a source of dehumanization (Bilewicz & Vollhardt, 2012; Harris & Fiske, 2007).

Transformation of outgroup stereotypes has often been an important aim for political propagandists in times of genocide. The proliferation of contemptuous stereotypes creates a useful tool for the perpetrators of genocide—it allows an empathic response to the suffering of outgroup members to be limited. A brain-imaging study focused on the process of dehumanization found that people exposed to images of groups portrayed as low in competence and low in warmth have lower activation of the medial prefrontal cortex (mPFC), a brain area responsible for social cognition (Harris & Fiske, 2007). When a contemptuous stereotype of a group is activated, people cease to consider members of such groups as equally human, and this effect can be observed at a very basic level of neural encoding. Instead, the areas responsible for disgust and a fleeing reaction (amygdala, insula) were highly active when watching members of groups stereotyped as low in warmth and low in competence. This might be responsible for the phenomenon of failure of empathy (Cikara & Fiske, 2011; Zaki & Cikara, 2015), as outgroup stereotypes seem to limit empathic responses to outgroup members.

Historically, the derogatory comments and speeches preceding genocide often portrayed outgroup members as lacking warmth and competence and deprived of basic human features. The posters hung in Nazi-occupied Poland portrayed Jews as vermin, rats, and lice and connected them to typhus and other diseases (Grabowski, 2009). Tutsis were often named cockroaches, rats, and snakes in the Hutu propaganda before the genocide in Rwanda (Straus, 2006).

Importantly, dehumanizing hate speech is not limited to the genocide context. For example, it can emerge in locker-room talks where the topic of sexual orientations is discussed. Recently, an experimental study by Fasoli and colleagues (2015) showed that when people are exposed to a homophobic epithet, they associate less human-related words with gay people than when people are exposed to a mere category label ("homosexuals") or generic insult (unrelated to homosexuality). These effects were observed both after supraliminal and subliminal exposure to hate speech. Thus, even in less extreme contexts, using homophobic epithets in a conversation may lead to increased dehumanization of sexual minorities, failure of empathy, and reproduction of similar statements.

Hate or Contempt Speech?

The stereotype of low warmth and low competence is often linked to the emotions of disgust and contempt (Cuddy, Fiske, & Glick, 2008). At the same time, the term "hate speech" itself suggests that hate is the emotion implied by the usage of derogatory language about minorities or immigrants. What, then, is the difference between these two emotions?

Hate is an emotion felt by the powerless toward an ostensibly stronger target (Fischer, Halperin, Canetti, & Jasini, 2018; Sternberg, 2005). It motivates collective action toward an enemy in order to neutralize him. The goal here is the destruction of the target and an attack rather than long-lasting discrimination (Ben-Ze'ev, 2008). The emotion of hate is closely related to anger, which is known to motivate normative forms of collective action (Górska & Bilewicz, 2015; Tausch & Becker, 2013). Thus, it is clearly an approach rather than an avoidance-related emotion.

Contempt, on the other hand, is an emotion felt by those in power toward groups that are perceived as inferior in terms of status (Fischer & Giner-Sorolla, 2016). It is an emotion justifying oppression and superiority, driven by a lack of respect for certain groups (Gervais & Fessler, 2017). Here the goal is different: to discriminate, to dominate, and to exclude from one's environment (Fischer et al., 2018). Contempt is an emotion that drives nonnormative and aggressive forms of collective action (Tausch et al., 2011). It is a secondary form of a more basic emotion of disgust, with which it is often mistaken (Alvarado & Jameson, 1996; Haidt & Keltner, 1999). In contrast to disgust which can be evoked even by objects, contempt is a strictly social emotion evoked solely in reaction to other individuals or social groups (Gervais & Fessler, 2017).

As derogatory language puts certain groups in the inferior position compared to the ingroup, we propose that it is rather contempt than hate that motivates such language—and that this emotion is also an effect of such language (Bilewicz, Kamińska, et al., 2017). In recent studies (Soral, Bilewicz, & Świderska, 2020), we examined the role of various negative emotions evoked by minority groups (Jews, gays, Roma, and Muslims) in predicting the use of hate speech. We asked our participants to what extent each of the minority groups can evoke emotions of anger, hate, disgust, contempt, or envy. Although the emotional background of hate-speech usage differed depending on the target (e.g., disgust was a unique, positive, and significant predictor of using antigay hate speech; hate was a unique, positive, and significant predictor of using anti-Roma hate speech), the positive predictor common to all tested target groups was the emotion of contempt.

In another study (Soral, Bilewicz, & Świderska, 2020), we first exposed participants to a large amount of hate-speech comments (vs. more generic negative contents) and then asked them to read and rate the offensiveness of another set of hateful statements (e.g., "I detest poofs—they are degenerated human beings, they should be treated" or "Muslims are stinky cowards, they can only murder women, children, and innocent people"). During the latter phase of the study, we observed and recorded participants' facial expressions while reading hate speech. The records were then automatically coded for the presence of various emotional expressions (cf. Ekman & Friesen, 1971; Ekman & Rosenberg, 1997). Among participants previously exposed to large amounts of hate speech, we observed a gradual increase in the presence of facial expressions of contempt, but also positive emotions such as joy. Such increases were not observed among nonexposed participants in the control condition. Among participants in the control condition, we observed a gradual increase in facial expressions of sadness, which were not present among participants exposed to hate speech.

These findings suggest that derogatory language is strongly linked to the emotion of contempt, both as an antecedent (a cause of using derogatory language about minorities) and as a consequence (an emotion elicited by frequent exposure to derogatory language). Taking that into account, we propose that the term "hate speech" should be replaced with the more accurate "speech of contempt."[1] We also consider this emotional route a key mechanism linking people's exposure to derogatory language and their tendency to use such language.

Mechanisms Inhibiting the Spread of Hate Speech

So far, it has been argued that individuals immersed in hateful environments are more likely to get involved in the derogation of others than those in more friendly surroundings. Furthermore, an increasing body of evidence suggests that digital media creates platforms through which hate speech can spread. Recently, however, such alarming reports have started to motivate actions to counter-hate

[1] We believe that as social or political psychologists, we should aim to use the language mirroring observed phenomena. In line with this, we could replace the term "hate speech" with "contempt speech" in the entire article. However, it should be remembered that the term "hate speech" originated not in psychology but in legal science. There, the emotional core of this phenomenon was perhaps less important than its legal consequences (i.e., whether or not to treat it in a similar way as hate crimes). Having considered this, as well as the expectation that our article would interest not only social and political psychologist but also scholars in such fields as law, communication studies, or social sciences in general, we have decided to abstain from making such a radical change in our terminology.

speech and its consequences. In 2018, the French president Emmanuel Macron publicly condemned anti-Semitic abuse hurled at a French academic, Alain Finkielkraut, which took place during one of the "yellow vest" protests (Johnstone, 2019). Interestingly, similar counterhate actions also occur among those notorious for their controversial, extremely right-wing opinions. In November 2017, after a Polish Independence March, right-wing journalist Tomasz Terlikowski condemned racist slogans among partakers saying that "A racist cannot be a good Catholic" ("Tak prawicowcy komentowali Marsz Niepodległości," 2017). In June 2019, he similarly condemned a Polish newspaper's ("Gazeta Polska") action of issuing stickers with the slogan "LGBT-free zone." He argued that, although he believes homosexuality is a sin, "Jesus has never rejected anyone" ("Terlikowski: Czy Jezus," 2019). These and similar actions are often motivated by a strong attachment to traditional social norms. Such traditional norms—often rooted in religious beliefs—tend to place a limit on acceptable forms of criticizing others. They often prohibit counternormative actions, such as hate speech, that would cause harm to others. It is thus likely that normocentric individuals following traditional social norms would strongly support actions prohibiting the use of hate speech.

A recent review of intervention programs confronting cyberhate (Blaya, 2019) identified several existing strategies, such as strengthening the legal frame, automated identification of cyberhate to regulate and support intervention, education of an informed and ethical use of the Internet, or empowering young people to produce counterspeech. The majority of these strategies rely on making existing social norms on hate speech salient or developing and promoting new social norms. Thus, at least among scholars and practitioners, social norms are—directly or indirectly—perceived as a tool to prevent the spread of hate speech. In sum, although social norms can facilitate the spread of hate speech, at the same time they may provide a mechanism leading to its inhibition.

The potential of norms in confronting hate speech should be particularly visible among people who have a higher tendency to follow the normative influence of others. According to Oliner and Oliner (1988), the majority of people who decided to rescue Jews during the Holocaust were driven by "normocentric expectations": They responded to the appeal of an authoritative other (resistance group leaders, members of close social networks, underground political leaders, or priests). In the time of extreme hate instigated by the Nazis, some people still opposed it due to normative influences. People high in authoritarian conventionalism (Altemeyer, 1988), who actively support social norms and punishing norm violators, could paradoxically be the ones who are most resilient to hate-speech exposure and most punitive toward those who express hate speech that violates established social norms.

Authoritarianism as a Protective Factor

In classic psychological research on political personality and intergroup relations, conformity to social norms was considered to be the primary cause of prejudice (Adorno, Frenkel-Brunswik, Levinson, & Sanford, 1950; Allport, 1954; Frenkel-Brunswik, 1948). Attachment to norms mixed with aggression toward deviants is still viewed as one of the main tenets of authoritarianism (Adorno et al., 1950; Altemeyer, 1981). Furthermore, research on twins (Lewis & Bates, 2014) revealed that heritable influences on right-wing authoritarianism (RWA) are entirely shared with the heritable effects on ingroup favoritism. Indeed, the latest formulations put RWA—next to social dominance orientation (SDO)—as an ideology explaining prejudice and outgroup hostility (Duckitt, 2001).

According to Duckitt's (2001) dual-process model of ideological attitudes, both RWA and SDO are responsible for prejudiced perceptions of outgroups. However, the model predicts that the motivations and specific consequences of RWA and SDO are different. While RWA is the result of perceiving the world as a dangerous place, SDO results from perceiving the world as a competitive jungle (see Duckitt & Sibley, 2010). Consequently, individuals high in RWA prefer authorities that emphasize law and order and promote normative conduct, whereas individuals high in SDO prefer authorities that favor social inequalities and strive for ingroup dominance (Duckitt & Sibley, 2010).

As both RWA and SDO promote prejudice, it is reasonable to assume that both would also motivate the use of hate speech. While in the case of SDO such reasoning seems valid, in the case of RWA its relationship with attitudes toward hate speech is far more complex. RWA may be viewed as a factor limiting prejudice, especially when its public expression is seen as *deviant* (Crandall & Stangor, 2005). In such instances, high RWA individuals will be more likely to support prohibition of hate speech than low RWA individuals. Although low RWA individuals are less biased against outgroup members, they tend to refrain from supporting actions aimed at limiting the freedom of speech and increasing censorship. To examine such reasoning, in one of our studies (Bilewicz, Soral, et al., 2017), we examined the role of SDO and RWA in predicting prejudice toward various minority groups, as well as support for prohibiting hate speech targeting the same set of groups. In nationwide surveys of adult Poles as well as Polish adolescents, we observed that both RWA and SDO were strong and significant predictors of prejudice toward Jews, Muslims, gays, Roma, Ukrainians, or Africans. Moreover, we observed that participants with high SDO were less supportive of prohibiting the use of hate speech toward these minorities. However, we observed the reverse pattern in the case of RWA: Participants with high levels of RWA declared more support for prohibition of hate speech than participants with low RWA (see Figure 2).

The study (Bilewicz, Soral, et al., 2017) corroborated the reasoning on the complex relationship between RWA and attitudes toward hate speech. While individuals high in RWA declared higher social distance toward various minority groups than their counterparts low in RWA, those high in RWA were also more likely to defend the same minority groups against derogatory language. Such a pattern can be understood when one considers that various ways of expressing prejudice can be perceived as norm violating to various degrees. A desire to distance oneself from outgroup members (e.g., avoiding living in the same neighborhood or rejection of interracial marriages) can be seen as a subtle—non-norm-violating—way of expressing prejudice, whereas the use of hate speech is usually perceived as a blatant—norm violating—expression of prejudice. This distinction seems to be an important factor moderating the relationship between RWA and expression of prejudice. Moreover, the pattern revealed that even in an ethnically and religiously homogeneous country characterized by a relatively high social distance toward other ethnicities and religions, such as Poland, social norms can be seen as one of the factors preventing the spread of hate speech.

Social Norm as a Protective Factor

Previous findings revealed that very strong social norms enforced by an authority may reverse the relationship between authoritarianism and negative attitudes about outgroups. For example, Roets, Au, and Van Hiel (2015) found that while in a Belgian sample authoritarianism was negatively correlated with support of multiculturalism and positive attitudes about outgroups, in a Singaporean sample both correlations were reversed. Such a difference is plausibly explained by the long-lasting (over 50 years) commitment of the Singaporean government in regulating an ethnically diverse society and promoting multiculturalism. In a similar vein, Hawdon et al. (2017) suggest that anti-hate-speech laws may decrease the rate of exposure to hate speech in countries where such regulations were introduced (e.g., Germany) as opposed to countries where free speech is protected at all cost (e.g., United States). Various legal regulations may induce fear of punishment preventing the expression of derogatory language, but they also form the basis of a social norm shared by the citizens. Such a social norm restrains the individual use of hate speech, but it may also promote counterspeech (e.g., critique and ostracization of those who use derogatory language).

In modern societies, the majority usually holds extremely negative attitudes toward those who express prejudice. Crandall, Eshleman, and O'Brien (2002) found that "racists" are one of the least accepted social groups (among groups such as "rapists," "child abusers," "wife beaters," or "terrorists"). Furthermore, their findings indicate that the majority would rather accept discrimination (in

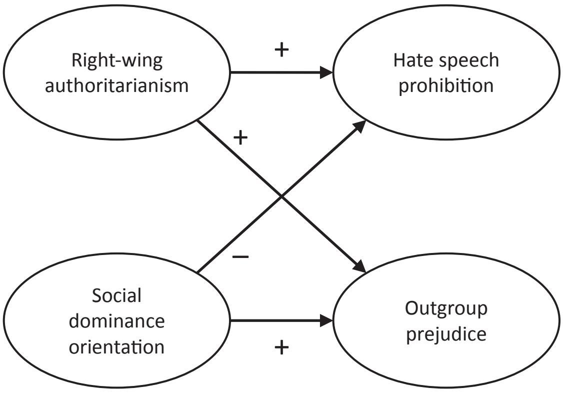

Figure 2. Effects of right-wing authoritarianism and social dominance orientation on hate-speech prohibition and outgroup prejudice. The sign (+/−) indicates the direction of a relationship. Adapted from Bilewicz, Soral, et al. (2017).

terms of dating, housing, or employment) of those labeled as "racist." Similarly, studies by Simon and Greenberg (1996) revealed that individuals, irrespective of their own attitudes toward the outgroup, tend to derogate and to express hostility toward those who use derogatory ethnic statements.

All in all, an antiracist environment enforces suppression of prejudice among those who genuinely hold such views (see justification-suppression model of prejudice; Crandall & Eshleman, 2003). However, two exceptions should be made. First, according to Crandall and Eshleman (2003), suppressing genuine prejudice requires significant mental energy and usurps attentional resources. Furthermore, it is short-lived and usually followed by a mildly negative mood. Thus, factors that deprive attentional resources and/or induce mental fatigue should make suppression of prejudice less likely to occur, even under strong external pressures. Second, individuals with high levels of genuine prejudice may find a way to express their views in a justified, socially approvable way. Such ways to justify prejudice usually include cognitions, beliefs, or ideologies. For example, belief in a just world (Lerner, 1980), that is, that "people get what they deserve," may justify prejudice against the poor, unemployed, or imprisoned.

Based on the reasoning presented in the justification-expression model of prejudice (Crandall & Eshleman, 2003), one could predict that although social norms may prevent the spread of hate speech, their efficiency is limited. First, exposure to even a single occurrence of derogatory language may provide a justification that will allow those genuinely prejudiced to express their views. Second, one could argue that repeated exposure to hate speech will make the suppression of prejudice a more difficult and demanding task. This may result from a desensitization mechanism similar to the one observed in aggression research (see Soral et al., 2018), where repeated occurrences of hate speech cease to evoke automatic, negative responses. Moreover, when hate speech permeates one's environment, the decision of which forms of expression of prejudice are acceptable and which are prohibited becomes more and more difficult. Third, when the exposure to hate speech is at its peak, hate speech can become a new social norm, replacing the more traditional norms of conduct. Of particular interest here may be the recent finding on *dynamic* or *trending* social norms (Mortensen et al., 2019; Sparkman & Walton, 2017). According to these studies, information on how people's behavior changes over time may impact one's behavior despite prevailing static norms. Thus, the mere observation that more and more people use hate speech may promote similar behavior, even though such behavior is deemed unacceptable by the majority of society.

Because of the limitations of relying on traditional social norms, preventing the spread of hate may require more than enforcing strict anti-hate-speech laws. Recent investigations have often focused on the development of new social norms in online communication and the new counter-hate-speech discourse (e.g., Mathew, Saha, et al., 2019; Schieb & Preuss, 2016). One of such new norms in online communication is the use of counterspeech. This norm encompasses a common-sense duty to correct others' misstatements and misperceptions, pointing out hypocrisy and contradictions, or simply denouncing hate speech and warning of its online and offline consequences (see Mathew, Saha, et al., 2019). Despite the lack of systematic evaluation, introducing new social norms of actively reacting to hate speech (rather than just passive ignorance) seems a promising pathway.

Empathy as a Protective Factor

In an American late-night talk show, its host Jimmy Kimmel invited famous musicians, sportsmen, and actors to read offensive comments about their talent and physical appearance ("mean tweets"). Many of them included instances of ageist, homophobic, and racist hate speech. When watching real victims reading offensive comments about them, one could easily empathize with a person harmed by such language. In Poland, 50 years after the anti-Semitic campaign orchestrated by the communist government, a local Jewish museum organized the campaign "Other history, the same hate." They invited members of the Jewish community (famous journalists, scientists, and public intellectuals) and victims of the 1968 purge to read current anti-Semitic comments found on the Internet. The video record of these readings appealed to viewers' empathy and raised awareness about the scale of hateful comments.

The underlying assumption behind such an intervention is that empathy can immunize people against hate speech. This assumption has relatively large support in psychological research on derogatory language. A study looking at gender differences in perceived harm of hate speech found that women consider hate speech more harmful because of their higher levels of empathy (Cowan & Khatchadourian, 2003). General empathy and empathic concern fully mediated gender differences in perceived harm of hate speech. This finding supports the view that empathy makes people more sensitive to hate speech. When contempt replaces empathy in group perception, the acceptance of derogatory language also becomes widespread.

Interventions based on empathy were often used in a context that is very close to hate speech, the context of cyberbullying and peer verbal violence. Studies looking at online behavior among adolescents who had received a derogatory image of a peer found that activation of cognitive empathy limits the frequency of forwarding such images to other people (Barlińska, Szuster, & Winiewski, 2013, 2015). Activation of cognitive empathy also effectively increases young people's tendency to report acts of cyberbullying—participants whose cognitive empathy was induced were more than six times more likely to report hateful content received through social media than participants who were assigned to a control task, without empathy activation (Barlińska, Szuster, & Winiewski, 2018). A systematic review of 40 studies on empathy and bullying (van Noorden, Haselager, Cillessen, & Bukowski, 2015) concluded that both cognitive and affective empathy reduce bullying and increase bystander intervention (defending a bullying victim).

There is one substantial problem with direct application of the cyberbullying studies' results to the context of hate speech. In the case of hate speech, derogatory language crosses group borders—such language targets outgroup members. Both psychological and neuroscientific research have found that empathic responses to outgroup members are scarce, and schadenfreude becomes a dominant response to outgroup members' suffering (Bruneau, Cikara, & Saxe, 2017; Levy et al., 2016; Vanman, 2016). At the same time, empathy interventions were often found to reduce racial prejudice (Batson & Ahmad, 2009; Finlay & Stephan, 2000), and possibly such interventions could also prove successful in resensitizing people to hate speech and protecting them against the negative consequences of derogatory language.

Epidemics of Hate Speech: Agent-Based Modeling Approach

Based on the theorizing presented in this review, we would like to propose a model that explains how hate speech can spread throughout society. We will start with a description of the factors that—according to our reasoning—play an important role in deciding whether an individual will engage in hate speech or not. These factors are by no means constant, but can change with time, depending specifically on an agent's social network and their previous experiences of exposure to hate speech. Thus, the presented model can be used in dynamic agent-based simulations to trace (and possibly predict) epidemics of hate speech. This approach uses sets of agents (abstract entities representing, e.g., individuals within society), each characterized by certain quantities/parameters (abstract values representing, e.g., psychological traits or behaviors) and interactions with neighboring agents (e.g., other individuals), to simulate dynamics of change in the system (abstract entity representing, e.g., society). Such an approach can reveal patterns of change not visible at lower levels of analysis.

Key Variables

The essential idea of the proposed model is that the propensity to use hate speech depends on three important quantities: (1) the level of contemptuous prejudice, (2) social norms regarding the use of hate speech, and (3) sensitivity to hate speech.

Level of Contemptuous Prejudice

Prejudiced individuals are more likely to utter hateful statements than those who are nonprejudiced. While numerous models may explain the inclination to verbally express prejudice, we would like to point to the idea that particularly contemptuous prejudice will manifest itself in hate speech. Groups portrayed in a contemptuous way are excluded from the community of human beings, and thus their members are less likely to evoke empathic reactions. We propose that in the case of target groups perceived in a noncontemptuous way, expression of prejudice is constrained by the barrier of empathy. Given the existing empathic response, just imagining the harm done by hate speech to the victim will lead to a similar unpleasant experience felt by the perpetrator and will inhibit any utterance of hate speech. Contrarily, perceiving a group in a contemptuous way will result in a lack of empathic response, and it will open the pathway to expression of prejudice. It should be noted that the level of contemptuous prejudice may characterize both an individual (i.e., it may describe individual differences in prejudice) as well as a target group (i.e., it may describe various cultural or social representations of different groups, some perceived in a more contemptuous ways than others; e.g., Eastern nations tend to be portrayed in a more contemptuous way than Western nations among Western Europeans). For simplicity, in our model we include a single target group and allow the level of contempt to vary between individuals.

In our model, we treat the level of contemptuous prejudice not as constant, but as a changing quantity. Crucially, based on our previous results, we believe that exposure to hateful statements directed at an outgroup will increase the level of contemptuous prejudice toward that outgroup. Therefore, although we assume that the level of contemptuous prejudice toward an example outgroup may initially be low, each case of contact with hate speech will lead to an increase in contemptuous prejudice by a fixed amount.

Social Norms Regarding the Use of Hate Speech

Given the lack of empathic reaction and the initial decision to express prejudice, individuals will start to monitor their social network for other similarly prejudiced attitudes (i.e., other hate-speech utterances). If one finds that the majority of one's peers tend to use hate speech, then it is highly probable that an individual will also start using hate speech. Such peer influence has previously been described in a number of studies on aggression (e.g., Boivin & Vitaro, 1995; Espelage, Holt, & Henkel, 2003; Glomb & Liao, 2003; see also previous sections for a review of the importance of social norms). An individual may start to mimic others because of fear of being ostracized and excluded from the group, or even out of fear of becoming a victim of hate speech. We assume that such peer influence will affect only individuals with a level of contemptuous prejudice high enough to cross the barrier set by empathic reaction. Thus, we propose that individuals with low levels of contemptuous prejudice (and existing empathic reaction) will not consider existing social norms regarding hateful utterances. More likely, they will engage in counterspeech (we are not modeling such behavior in our model) or will leave their social network. However, given the lack of an empathic reaction, if an individual finds that the majority of his or her peers do not use hate speech (i.e., indication of social norms prohibiting hate speech), then the individual's propensity to utter such a statement will depend on the quantity described in the next step.

Sensitivity to Hate Speech

Given the initial decision to express prejudice, an individual is constrained by existing social norms. A social norm prohibiting the use of hate speech will successfully prevent such behavior, but only to the extent that an individual is able to recognize that the behavior violates the social norm. If an individual is not able to recognize that the behavior violates the existing social norm, the individual will engage in hate speech. In contrast, successful recognition that the statement violates the social norm of not using hate speech would prevent one from using hate speech.

A crucial variable that can explain the ability to differentiate between utterances that do or do not violate existing social norms is the level of sensitivity to hate speech. Participants highly sensitive to hate speech are able to recognize that an utterance violates social norms, whereas those with low sensitivity to hate speech would not differentiate between hateful statements and those that can be regarded as permitted critique. Based on our previous findings (see Soral et al., 2018), we propose that the level of sensitivity to hate speech will depend on one's previous exposure to hate speech, such that each subsequently seen hate-speech statement will decrease sensitivity to similar statements.

Full Model

The full model is summarized on Figure 3. We propose that the pathway to expression of hate speech is a conditional process, in which each step depends on the outcomes of the previous one. The ultimate outcome—behavior—is a simple binary quantity: Either an individual would engage in hate speech or not.

An individual will not engage in hate speech if one of the following conditions are met: (1) presence of *empathic response*: a case of an individual who has a relatively low level of contempt; such a factor alone will restrain such a person from using hate speech; (2) *conformity to antidiscriminatory social norms*: a case of an individual who may have a relatively high level of contempt, but still his or her social network consists of a majority of people not using hate speech; additionally he or she is able to identify which statements can be regarded as hate speech—as a result, such an individual conforms to antidiscriminatory social norms and will not use hate speech.

Contrarily, an individual will engage in hate speech under the following conditions: (1) *radicalization*: an individual may have a relatively high level of contemptuous prejudice and his social network consists of a majority of peers using hate speech; (2) *faux pas*: an individual may have a relatively high level of contemptuous prejudice, his social network consists of a majority of peers not

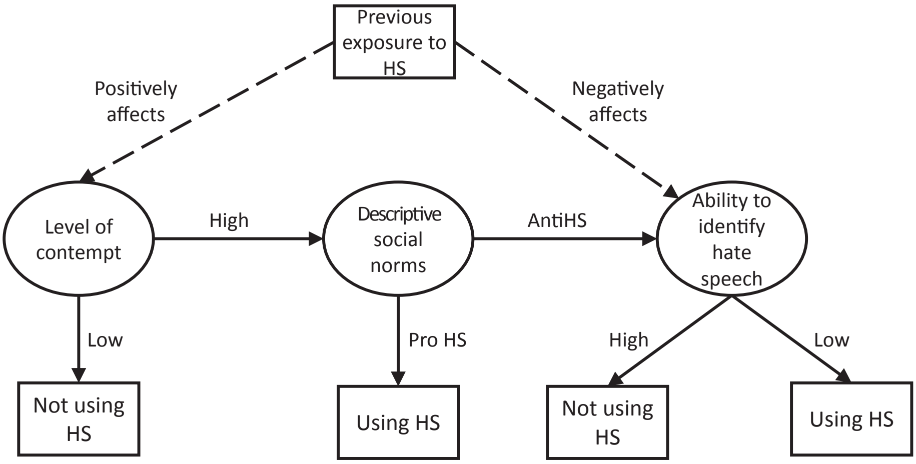

Figure 3. Model of expressing hate speech. Solid lines denote conditional steps within each iteration of the simulation, whereas dashed lines indicate consequences of the previous iteration.

using hate speech, but he or she is not able to identify that some statements can be regarded as hate speech: he or she will not conform to social norms and use hate speech.

An additional quantity that is included in our model is previous exposure to hate speech. We propose that each case of contact with hate speech will: (1) increase the level of contemptuous prejudice; (2) decrease the sensitivity to hate speech and as a result will decrease the ability to identify hate speech and conform to social norms.

As a low level of contemptuous prejudice is a factor that alone can restrain individuals from engaging in hate speech, we postulate that it does not depend on the social norms nor does it depend on the ability to identify hate speech. Obviously, an individual with a low level of contemptuous prejudice may utter statements regarded as hate speech due to some factors not included in our model, hence treated as random. To address this issue, we model the effects of contempt probabilistically and not in deterministic way (i.e., level of contempt stands for the probability of using hate speech). We also propose that an individual in a pro-hate-speech social network would engage in hate speech irrespective of whether he or she is able to identify that a statement can be regarded in such a way. Thus, we assume that pro-hate-speech networks give a sense of freedom from any constraints associated with social norms, where individuals can say whatever they want—be it hate speech or not. Finally, we assume that an individual with a high level of contempt, but within an anti-hate-speech network and possessing a high ability to identify hate speech will conform to social norms and not use hate speech. Thus, we consider the ability to identify hate speech to be a single factor that affects the use of hate speech in such a case. Obviously, there may be some other important factors (e.g., cynicism), but in our model we treat them as random noise. Therefore, we model the effect of ability to identify hate speech probabilistically.

Model Implementation and Simulation

The model proposed here describes factors responsible for engaging in hate speech at the individual level. However, our aim was to examine how hate speech can spread throughout society and to translate the individual-level processes to the societal level. In order to study the processes emerging from individual interactions, we used agent-based simulations (see, e.g., Nowak, Szamrej, & Latané, 1990). This approach is based on simulated interactions between individuals (agents) residing in a grid-like plane (e.g., chessboard). At the beginning, each agent is assigned with certain values of the parameters (e.g., certain amount of money or level of happiness). Next, a simulation starts in an iterative manner. At each step an agent interacts with its neighbors (e.g., gives away a certain amount of money) using predefined rules of change. Once the change is computed for all agents residing in a grid, the simulation advances (i.e., the values of parameters are updated) to the next step and the process repeats.

To implement our model of hate-speech epidemics, we placed 1600 agents in a 40 × 40 grid-like plane. We randomly and independently assigned several initial quantities: (1) *the use of hate speech*: We started with an initial value of 20% of agents using hate speech and 80% of agents not using hate speech; (2) *the level of contempt* was randomly drawn from a beta distribution (ranging from 0 to 1) with parameters $\alpha = 2$ and $\beta = 5$, and the mean equal to .29; (3) *sensitivity to hate speech* was randomly drawn from a beta distribution (ranging from 0 to 1) with parameters $\alpha = 5$ and $\beta = 1$, and the mean equal to .83.[2] In other words, we started with a relatively nonhateful environment, with a low level of contemptuous prejudice toward an outgroup, and relatively high ability to identify a statement as hate speech.

We programmed the change rule according to our theoretical model. At each step, the value of contemptuous prejudice was considered first. With probability equal to 1 minus the agent's level of contempt, the agent's future behavior changed to not using hate speech. Otherwise, the number of hate-speech occurrences in eight neighboring cells was computed. If the number of hate-speech occurrences was greater than four, the agent's future behavior changed to using hate speech. However, if the number of hate-speech occurrences was equal to or smaller than four, then, with probability equal to the agent's ability to identify hate speech, the future behavior changed to not using hate speech (and with probability equal to 1 minus the agent's ability, it changed to using hate speech). Furthermore, at the end of an iteration, each agent's level of contempt increased by δ times the number of hate-speech occurrences in the neighborhood (unless it was already equal to 1), and each agent's ability to identify hate speech decreased by 5δ times the number of hate-speech occurrences in the neighborhood (unless it was already equal to 0). δ refers here to an arbitrary quantity defining how fast the simulation will run. Thus, we assumed that the rate of change in the level of contempt is five times smaller than the rate of change in the sensitivity to hate speech. This proportion was assigned based on some considerations of how the attitudes and the sensitivity to hate speech can change, as well as on our previous results.[3]

The simulation program was written in Python, with the visualization produced by a JavaScript code.[4] Simulations were run several times and resulted in equilibrium usually after no more than 300 iterations.

[2] The beta distribution describes a continuous variable defined on the interval [0, 1]. It is thus especially useful for describing potential values of parameters such as proportions or probabilities (in contrast to the normal distribution where possible values are not constrained). The beta distribution is commonly described by two parameters, denoted by α and β, that specify its shape. The mean of a beta-distributed variable is $M = \alpha/(\alpha + \beta)$, and the variance of such variable is $\mathrm{Var} = \alpha\beta/[(\alpha + \beta)^2(\alpha + \beta + 1)]$. The values used in our model can be interpreted as: (1) $\alpha = 2$, $\beta = 5$, low levels of contempt with majority of agents having values around .2, and only 10% having values higher than .5; (2) $\alpha = 5$, $\beta = 1$, high levels of sensitivity to hate speech with majority of agents having values near 1, and only 3% having values lower than .5.

[3] For example, in one of the previous surveys, we computed ratios of partial correlations between exposure to hate speech and desensitization to hate speech and between exposure to hate speech and contempt toward a group targeted by hate speech. Across four target groups (Jews, Roma, Muslims, and gays), the average ratio was equal to 4.77 (values above 1 indicate that exposure to hate speech is correlated more strongly with desensitization to hate speech than with contempt toward a group targeted by hate speech; values below 1 indicate the reverse).

[4] The full source code can be found here (https://osf.io/zhqv3/).

Model Results

We will now describe the evolution of our model as it progresses through various phases.

After 10 initial iterations (see Figure 4), both the average level of contempt (.28) and the average level of sensitivity to hate speech (.80) did not differ significantly from the initially assigned values. However, the observed proportion of hate-speech occurrences was significantly smaller than the initially assigned value (5% vs. 20%). These occurrences were randomly placed with no clear spatial pattern. Furthermore, the correlation between the level of contempt and the use of hate speech was positive but relatively low, $r = .13$. This pattern may suggest that with fairly low levels of contempt in society and high levels of sensitivity to hate speech, social norms are a powerful barrier to the spread of hate speech. Although hate speech occurs among those with relatively high levels of contempt (see

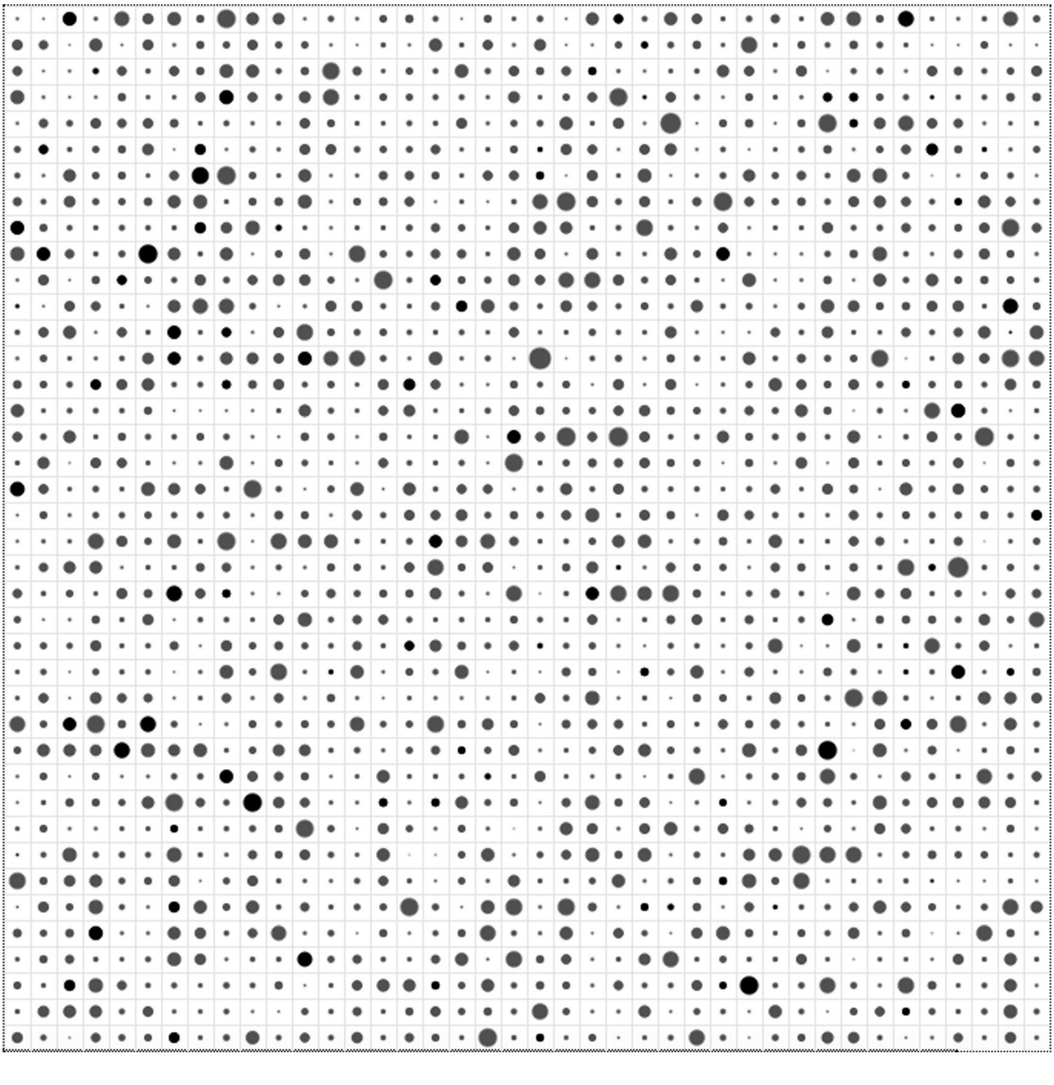

Figure 4. Model after 10 iterations. Black circles denote hate-speech occurrences, gray circles denote lack of hate speech. The size of the circle refers to the level of contempt.

large black circles), a relatively high proportion of highly prejudiced individuals refrain from using hate speech (see large gray circles).

After 100 iterations (see Figure 5), the average level of contemptuous prejudice increased slightly (from .28 to .36), but the decrease in average sensitivity to hate speech was drastic (from .80 to .43). Although the sensitivity to hate speech decreased by almost half, it was not accompanied by an increase in the proportion of hate-speech occurrences of the same magnitude (it changed from 5% to 21%). Yet, the correlation between the level of contempt and the use of hate speech increased from $r = .13$ to $r = .28$. Importantly, multiple (up to eight) small-sized (i.e., consisting of around eight agents) clusters of agents using hate speech emerged. These clusters were formed around highly prejudiced agents that tend to use hate speech. Furthermore, although the agents in these clusters' centers were highly prejudiced, agents at the borders of the clusters were usually less prejudiced, yet they still used hate speech. Furthermore, several instances of highly prejudiced agents were observed that,

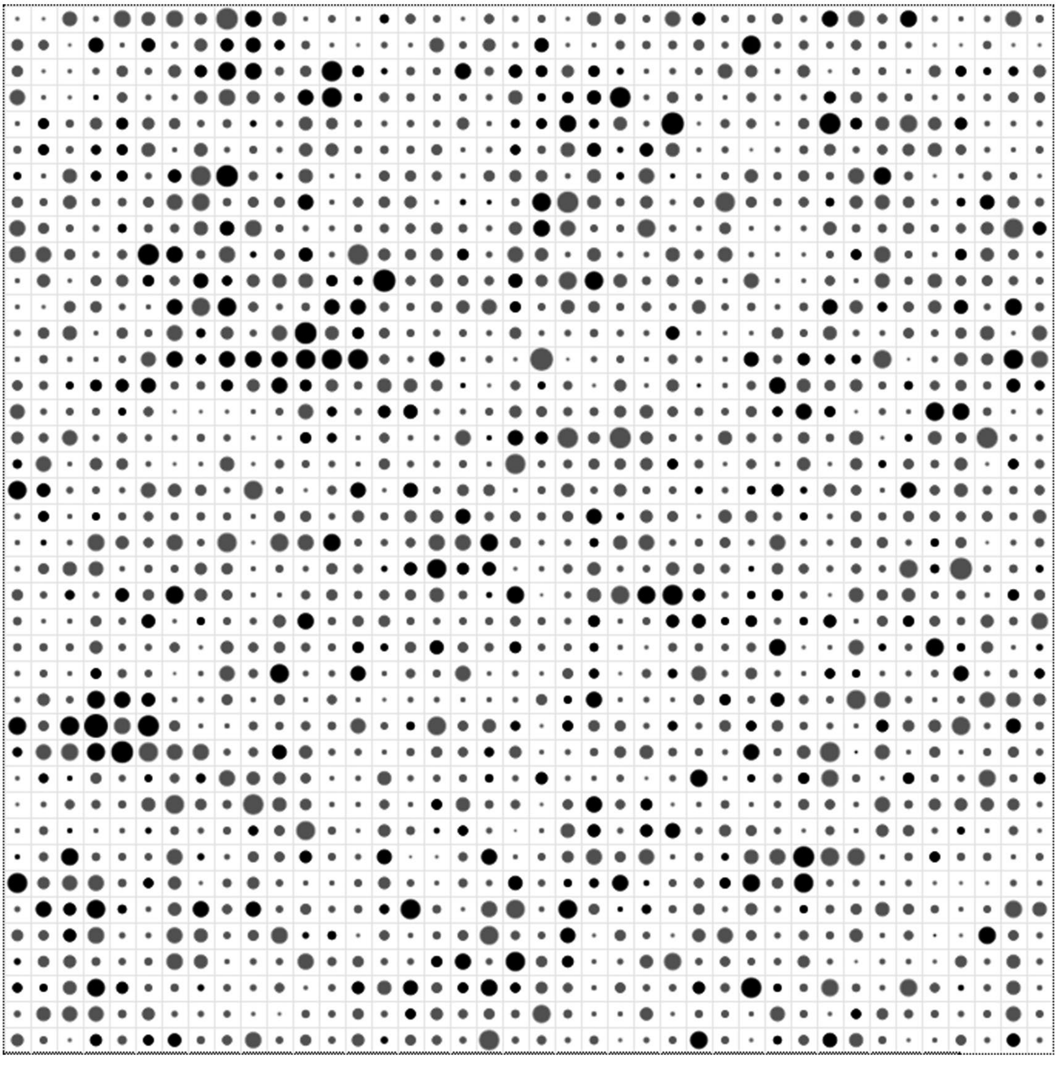

Figure 5. Model after 100 iterations. Black circles denote hate-speech occurrences, gray circles denote lack of hate speech. The size of the circle refers to the level of contempt.

despite their level of prejudice, refrained from using hate speech. In almost all cases, such agents were surrounded by a majority of low-prejudiced nonhaters.

There are three important implications of this result. First, agents did not form one large cluster, but rather multiple small clusters in which individual agents formed around one or two *leaders* who seem to counteract the existing descriptive social norms and repeatedly affect neighboring cells (by decreasing their sensitivity to hate speech and increasing the level of contempt). Second, individuals with relatively low levels of contempt may engage in hate speech given their social network consists of *radicalized* neighbors. Third, highly prejudiced individuals are strongly affected by their environments and suppress their prejudice in the presence of a multitude of other nonhateful agents.

After 200 iterations (see Figure 6), the proportion of haters increased drastically up to 65% (vs. 21%), with the average sensitivity of .07 and the average contempt of .69. The correlation between the level of contempt and the use of hate speech increased from r = .28 to r = .35. In this

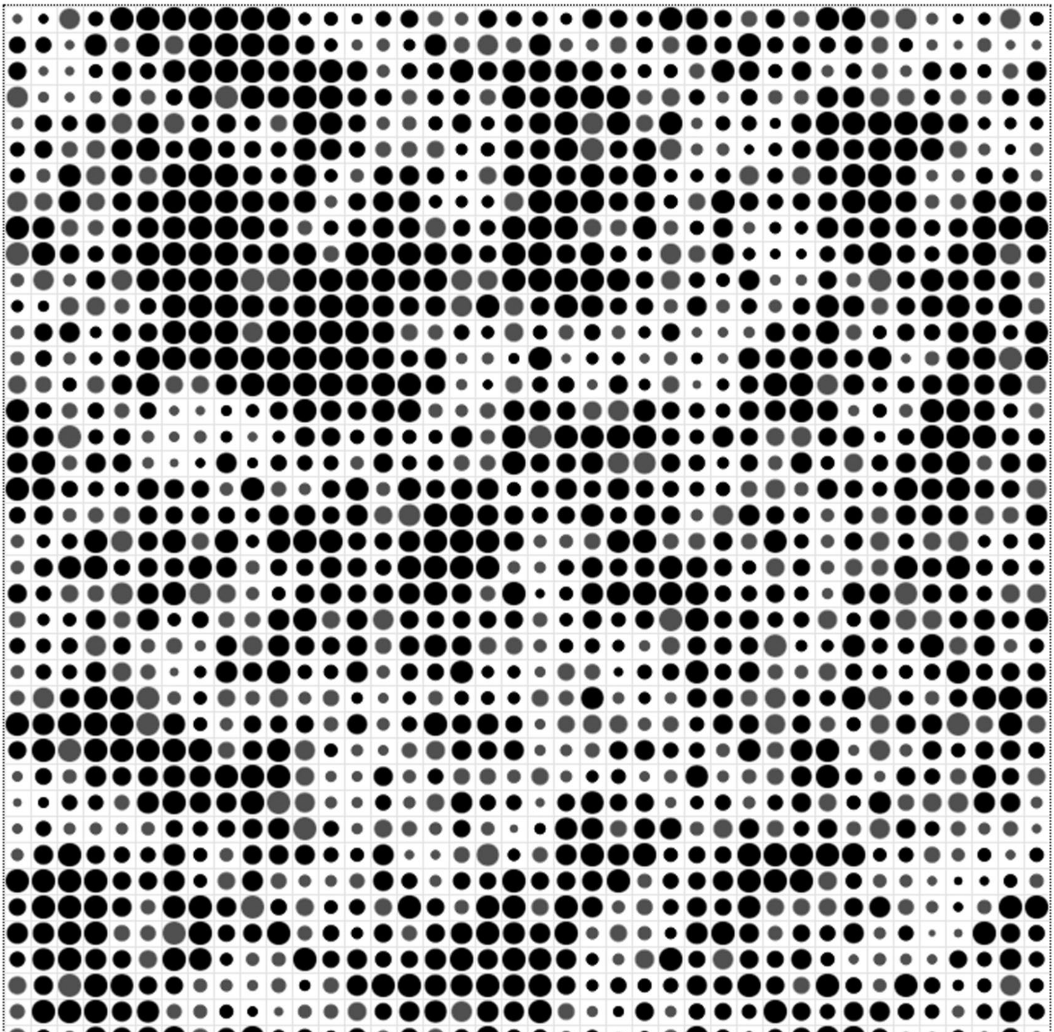

Figure 6. Model after 200 iterations. Black circles denote hate-speech occurrences, gray circles denote lack of hate speech. The size of the circle refers to the level of contempt.

phase, previously small clusters of agents using hate speech expanded into large clusters consisting mostly of highly contemptuous agents. Contrarily, clusters of agents not using hate speech consisted mostly of agents low in prejudice. Such societal polarization was observed after several runs of our simulations.

After 260 iterations (see Figure 7), agents using hate speech constitute a majority (90% of haters, average contempt = .89, average ability to identify hate speech = .07). Not surprisingly, the correlation between the level of contempt and the use of hate speech starts to decrease (*r* = .24 vs. *r* = .35). This is due to the fact that, as it evolves, the similarity between agents on both of these variables increases (and thus the variance of both variables decreases). We interpret this last phase of the simulations as the end phase that may occur only in the case of a lack of any exogenous interventions (e.g., military conflict, ethnic cleansing, or international diplomatic intervention).

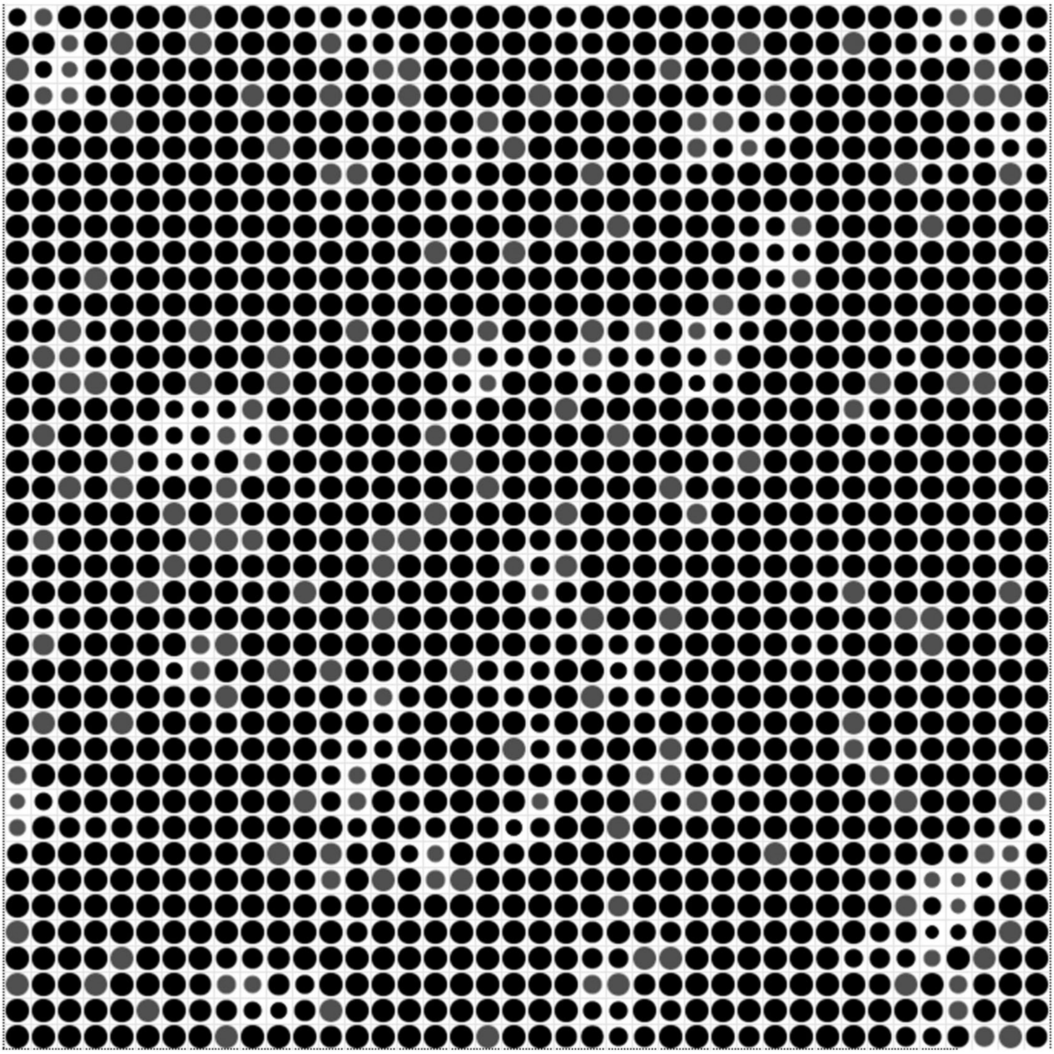

Figure 7. Model after 260 iterations. Black circles denote hate-speech occurrences, gray circles denote lack of hate speech. The size of the circle refers to the level of contempt.

Model Critique and Future Directions

Obviously, the presented results are correct insofar as one accepts the assumptions of the model. Furthermore, changing some of the parameters like the initial distribution of the level of contempt or initial distribution of the sensitivity to hate speech can drastically change the obtained results. This calls for further, more systematic explorations of the presented idea and perhaps juxtaposing the results obtained with real-life data about the dynamics of the hate-speech epidemic. There are two assumptions behind this simulation for which higher caution should be exercised and should be investigated in further analyses.

The first assumption is related to the number of peers an agent interacts with. In the present simulation, each agent interacted with exactly eight of its neighbors. This was obviously a necessary simplification, which, when translated into a real-life situation, stated that the number of interactions for each individual is the same. This is obviously far from the truth: For example, Facebook or Twitter users vary drastically in the number of Facebook friends or Twitter followers they have, that is, some may have no more than 20 while others more 2000. This can impact the outcome and dynamics of the simulation drastically: For example, those with a larger number of connections may be more exposed to hate speech than those with a small number of connections. Thus, a more rapid change in attitudes would be observed among people with a larger number of connections. In order to model such a process, the structure of existing social networks should be implemented in the structure of a much more complex simulation than the one presented above.

Another assumption of the present model is that the level of contemptuous prejudice increases monotonously and that people's sensitivity to hate speech decreases monotonously, following exposure to hate speech (down to their minimal or up to maximal levels). Thus, we neglected some factors that could decrease the level prejudice or increase the sensitivity to hate speech. Introduction of such factors would require additional research. Yet, we would like to propose one modification that is worth studying in the future. Suppose an agent with a high level of contemptuous prejudice uses hate speech within a group of peers that do not use hate speech, that is, an agent commits a faux pas. An agent may reflect on his or her inappropriate behavior or may be reprimanded by some of his or her peers. Assuming that human beings are capable of learning such social rules, the ability to identify hate speech should increase in future iterations. It is not clear yet whether such faux pas' effect would exceed the desensitizing effect of exposure to hate speech. It is also necessary to assume that agents possess the ability to reflect on their hateful behaviors or that others are sufficiently motivated to reprimand their peers for hateful language. None of these additional assumptions are guaranteed. Such possibilities deserve further examination in the future.

Altogether, it is unrealistic to assume that the proposed model could be successfully applied in every possible setting. Given the constraints and content monitoring present in modern social media (e.g., Facebook, Twitter, or YouTube), certain differences would likely occur between the model predictions and real-life data. However, we would argue that the proposed model can be reliably used to describe and explain the spread of hate speech across nonmoderated social networks, whose founders are driven by the belief that online community norms are sufficient to prevent the spread of hate speech. Mathew, Illendula, et al. (2019) has recently analyzed data from one such platform—Gab—to examine temporal dynamics of the spread of hate speech. Their analyses revealed that from October 2016 to March 2018: (1) The volume of hate-speech content on Gab increased; (2) the rate at which new Gab members were becoming hateful users was steadily increasing; and (3) the language used in the Gab community was becoming increasingly similar to the language used by users with high hate scores. Moreover, Mathew, Illendula, et al. (2019) observed that, as compared to nonhateful users, hateful users were more likely to quickly rise in ranks within their social network and become central influencers. Overall, given the similarities between our model and the results of the analyses conducted by Mathew, Illendula, et al. (2019), we argue that our model can be used to better understand psychological factors responsible for epidemics of hate speech.

Concluding Remarks

Contemporary societies are faced with a significant social change in the form of developing communication means and mass media. As an effect of moving from traditional media (newspapers, television) to new media (social media, citizen journalism) as primary sources of knowledge about politics, many constraints protecting one from abusive communication have disappeared. This has led to a significant increase in the use of derogatory language in one's social environment. People are both more often exposed to prejudiced opinions and are more openly expressing such views.

This article is an attempt to explain the psychological processes that have been activated by these technological changes in social environments. We propose that default emotional responses to other people, such as empathy, are replaced by intergroup contempt—an emotion fundamental to derogatory behaviors. Our research suggests that contempt might be both a motivator and a consequence of derogatory language. At the same time, increased presence of derogatory language might create a sense of normativity of such language. As long as derogatory language is counternormative, people who have a tendency to follow norms and obey rules (i.e., conventionalist authoritarians) are likely to oppose hate speech. However, when such language is prevalent in the environment, and when political or religious leaders support expressions of hatred, one's sense of norms might change, and hate speech might stop being a social taboo. Finally, the omnipresence of hate speech might reduce people's sensitivity to the fact that such language is offensive and humiliates other people. Through a very basic behavioral process of desensitization, societies might become more accepting toward derogatory language and less accepting toward immigrants, as well as religious, ethnic, and sexual minorities.

Based both on our research and on social psychological theorizing, we came up with a multi-level model of hate-speech epidemics and proposed and agent-based modeling approach to trace its dynamics. Considering the knowledge from various branches of social psychology and social sciences in general, we wanted to identify crucial factors (and their interactions) that facilitate or inhibit the spread of hate speech in society. The main message from this model is that the potential mechanisms that could inhibit the spread of hate speech are destroyed by hate speech itself—as the norms and empathic responses become eroded by frequent exposure to derogatory language. The model depicts this epidemic loop that ultimately generates an omnipresence of hate speech in human social environments.

The proposed model of hate-speech epidemics is obviously not an exhaustive one, and there are many other factors that might affect the spread of hate speech. For example, empathy and contempt are not the only emotions whose dynamics are important to our understanding of hate-speech dynamics. A study performed in five countries four weeks after the 2015 Paris terrorist attacks showed that people exposed to online hate speech declared higher perceived societal fear—an emotion leading to a shared sense of uncertainty (Oksanen et al., 2018). Fear and anxiety are known to be powerful intergroup emotions, leading to prejudice and intergroup violence. Therefore, these results suggest that derogatory language might cause not only contempt directed toward certain groups, but also affect a basic sense of anxiety that could lead to maladaptive prejudice reactions. Instead of constructively tackling the radicalization, hate speech might lead to further coradicalization. Although our article mostly addresses changes in perceptions of and attitudes toward outgroup members, we are confident that the study of more general feelings caused by hate speech (such as a sense of fear or uncertainty) is worth further exploration.

Obviously, our work is motivated to a great extent by concern: How can theoretical knowledge be transferred to applied settings? How can we efficiently combat (online) hatred? By no means is the presented model sufficient to deal with these issues. Recent findings suggest that hate forms a global "network-of-networks" (Johnson et al., 2019), which means that successful implementation of a certain antihate strategy in one of the networks (e.g., on Facebook) does not eliminate hatred, but rather moves it to another network (to another country, continent, or language). One example of such a process may be the launch of a new far-right social media platform in 2017—Gab—where anti-hate-speech policies do not apply, and where hate speech flourishes (Mathew, Illendula, et al., 2019). Thus, in order to make communication media a more inclusive and accepting place, we may require a policy matrix rather than one policy (see Johnson et al., 2019). According to Johnson et al. (2019), such policies could involve banning small hate-speech groups on social media, randomly banning a small proportion of hate-speech users, encouraging anti-hate-speech users to form common platforms that would counter hate speech, and utilizing competing narratives within hate-speech groups to provoke in-fights. Altogether, such matrix of policies should be implemented globally, which would require changes in domestic and international laws (e.g., making social media platforms responsible for implementing the matrix of policies).

In addition, there are many/some protective factors that go beyond basic psychological processes described in this article (social norms and empathy). Oksanen et al. (2018) suggested social trust to be one such factor, as it decreases the sense of uncertainty and establishes resilience (i.e., the capacity to recover from the shock of disruptive events). Thus, social trust could inoculate against waves of cyberhate observed after shocking events such as terrorist attacks. On the other hand, trust could be a risk factor: People who are more trusting of others may be more prone to believe in the content of hateful comments. More so, high trust and a sense of belonging increase the risk of becoming a victim of online hate speech (Kaakinen, Keipi, Oksanen, & Räsänen, 2018). Nevertheless, basic societal qualities—such as social capital, mutual trust, and cooperativeness—should be taken into account when thinking about human resilience to hate speech.

Our research and theorizing on hate speech points to the fact that one should not undervalue the role of political speeches and hateful statements expressed by political leaders in shaping people's attitudes toward immigrants and minority groups. First, through the massive presence of their words in social media, politicians' hateful utterances desensitize people to derogatory language, which, in turn, increases stereotyping and discrimination and eventually leads to a general increase of intergroup contempt in society. Second, politicians' hate speech creates a norm of conduct that replaces existing social norms that protect minorities. Therefore, we view our research as a warning against the devastating effects of political hate speech on contemporary societies.

ACKNOWLEDGMENTS

This research was supported by a grant from the Polish National Science Center [no 2017/26/E/HS6/00129] awarded to Michał Bilewicz. Correspondence concerning this article should be addressed to Michał Bilewicz, Faculty of Psychology, University of Warsaw, Stawki 5/7, 00-183 Warsaw, Poland. E-mail: bilewicz@psych.uw.edu.pl

REFERENCES

Adorno, T. W., Frenkel-Brunswik, E., Levinson, D. J., & Sanford, R. N. (1950). *The authoritarian personality*. New York, NY: Harper & Row.

Allport, G. W. (1954). *The nature of prejudice*. Reading, MA: Addison-Wesley.

Altemeyer, B. (1981). *Right-wing authoritarianism*. Winnipeg, Canada: University of Manitoba Press.

Altemeyer, B. (1988). *Enemies of freedom: Understanding right wing authoritarianism*. San Francisco, CA: Jossey-Bass.

Alvarado, N., & Jameson, K. (1996). New findings on the contempt expression. *Cognition & Emotion, 10,* 379–407. https://doi.org/10.1080/026999396380196

Anderson, C. A., & Carnagey, N. L. (2004). Violent evil and the general aggression model. In A. G. Miller (Ed.), *The social psychology of good and evil* (pp. 168–192). New York, NY: Guilford.

Aratani, L. (2019, August 5). "Invasion" and "fake news": El Paso manifesto echoes Trump language. *The Guardian*. Retrieved from https://www.theguardian.com/us-news/2019/aug/05/el-paso-shooting-suspect-trump-language-manifesto

Bäck, E. A., Bäck, H., Gustafsson Sendén, M., & Sikström, S. (2018). From I to We: Group formation and linguistic adaption in an online xenophobic forum. *Journal of Political and Social Psychology, 6,* 76–91. https://doi.org/10.5964/jspp.v6i1.741

Barlińska, J., Szuster, A., & Winiewski, M. (2013). Cyberbullying among adolescent bystanders: Role of the communication medium, form of violence, and empathy. *Journal of Community & Applied Social Psychology, 23,* 37–51. https://doi.org/10.1002/casp.2137

Barlińska, J., Szuster, A., & Winiewski, M. (2015). The role of short- and long-term cognitive empathy activation in preventing cyberbystander reinforcing cyberbullying behavior. *Cyberpsychology, Behavior, and Social Networking, 18,* 241–244. https://doi.org/10.1089/cyber.2014.0412

Barlińska, J., Szuster, A., & Winiewski, M. (2018). Cyberbullying among adolescent bystanders: Role of affective versus cognitive empathy in increasing prosocial cyberbystander behavior. *Frontiers in Psychology, 9,* 799. https://doi.org/10.3389/fpsyg.2018.00799

Bartholow, B. D., Bushman, B. J., & Sestir, M. A. (2006). Chronic violent video game exposure and desensitization to violence: Behavioral and event-related brain potential data. *Journal of Experimental Social Psychology, 42,* 532–539. https://doi.org/10.1016/j.jesp.2005.08.006

Batson, C. D., & Ahmad, N. Y. (2009). Using empathy to improve intergroup attitudes and relations. *Social Issues and Policy Review, 3,* 141–177. https://doi.org/10.1111/j.1751-2409.2009.01013.x

Ben-Ze'ev, A. (2008). Hating the one you love. *Philosophia, 36,* 277–283. https://doi.org/10.1007/s11406-007-9108-2

Bianchi, M., Piccoli, V., Zotti, D., Fasoli, F., & Carnaghi, A. (2017). The impact of homophobic labels on the internalized homophobia and body image of gay men: The moderation role of coming-out. *Journal of Language and Social Psychology, 36,* 356–367. https://doi.org/10.1177/0261927X16654735

Bilewicz, M., Kamińska, O. K., Winiewski, M., & Soral, W. (2017). From disgust to contempt-speech: The nature of contempt on the map of prejudicial emotions. *Behavioral and Brain Sciences, 40,* 20–21. https://doi.org/10.1017/S0140525X16000686

Bilewicz, M., Soral, W., Marchlewska, M., & Winiewski, M. (2017). When authoritarians confront prejudice. Differential effects of SDO and RWA on support for hate-speech prohibition. *Political Psychology, 38,* 87–99. https://doi.org/10.1111/pops.12313

Bilewicz, M., & Vollhardt, J. R. (2012). Evil transformations: Psychological processes underlying genocide and mass killing. In A. Golec de Zavala & A. Cichocka (Eds.), *Social psychology of social problems. The intergroup context* (pp. 280–307). New York, NY: Palgrave Macmillan. https://doi.org/10.1007/978-1-137-27222-5_11

Blaya, C. (2019). Cyberhate: A review and content analysis of intervention strategies. *Aggression and Violent Behavior, 45,* 163–172. https://doi.org/10.1016/j.avb.2018.05.006

Boivin, M., & Vitaro, F. (1995). The impact of peer relationships on aggression in childhood: Inhibition through coercion or promotion through peer support. In J. McCord (Ed.), *Coercion and punishment in long-term perspectives* (pp. 183–197). New York, NY: Cambridge University Press. https://doi.org/10.1017/CBO9780511527906.011

Brauer, M., & Chaurand, N. (2010). Descriptive norms, prescriptive norms, and social control: An intercultural comparison of people's reactions to uncivil behaviors. *European Journal of Social Psychology, 40,* 490–499. https://doi.org/10.1002/ejsp.640

Breiter, H. C., Etcoff, N. L., Whalen, P. J., Kennedy, W. A., Rauch, S. L., Buckner, R. L., … Rosen, B. R. (1996). Response and habituation of the human amygdala during visual processing of facial expression. *Neuron, 17,* 875–887. https://doi.org/10.1016/S0896-6273(00)80219-6

Bruneau, E. G., Cikara, M., & Saxe, R. (2017). Parochial empathy predicts reduced altruism and the endorsement of passive harm. *Social Psychological and Personality Science, 8,* 934–942. https://doi.org/10.1177/1948550617693064

Bruneau, E. G., Pluta, A., & Saxe, R. (2012). Distinct roles of the "Shared Pain" and "Theory of Mind" networks in processing others' emotional suffering. *Neuropsychologia, 50,* 219–231. https://doi.org/10.1016/j.neuropsychologia.2011.11.008

Bulska, D., & Winiewski, M. (2018). Irrational critique of Israel and Palestine: New clothes for traditional prejudice? *Social Psychological Bulletin, 13,* 1–23. https://doi.org/10.5964/spb.v13i1.25497

Carnagey, N. L., Anderson, C. A., & Bartholow, B. D. (2007). Media violence and social neuroscience: New questions and new opportunities. *Current Directions in Psychological Science, 16,* 178–182. https://doi.org/10.1111/j.1467-8721.2007.00499.x

Carnaghi, A., & Maass, A. (2007). In-group and out-group perspectives in the use of derogatory group labels: Gay versus fag. *Journal of Language and Social Psychology, 26,* 142–156. https://doi.org/10.1177/0261927X07300077

Carnaghi, A., & Maass, A. (2008). Derogatory language in intergroup context: Are "gay" and "fag" synonymous? In Y. Kashima, K. Fiedler, & P. Freytag (Eds.), *Stereotype dynamics: Language-based approaches to the formation, maintenance, and transformation of stereotypes* (pp. 117–134). Mahwah, NJ: Erlbaum.

Cialdini, R. B., Reno, R. R., & Kallgren, C. A. (1990). A focus theory of normative conduct: Recycling the concept of norms to reduce littering in public places. *Journal of Personality and Social Psychology, 58,* 1015–1026. https://doi.org/10.1037/0022-3514.58.6.1015

Cikara, M., & Fiske, S. T. (2011). Bounded empathy: Neural responses to outgroup targets' (mis)fortunes. *Journal of Cognitive Neuroscience, 23,* 3791–3803. https://doi.org/10.1162/jocn_a_00069

Cline, V. B., Croft, R. G., & Courrier, S. (1973). Desensitization of children to television violence. *Journal of Personality and Social Psychology, 27,* 360–365. https://doi.org/10.1037/h0034945

Cohen, F., Jussim, L., Harber, K. D., & Bhasin, G. (2009). Modern anti-Semitism and anti-Israeli attitudes. *Journal of Personality and Social Psychology, 97,* 290–306. https://doi.org/10.1037/a0015338

Council of Europe (1997). Recommendation No. R (97) 20 of the Committee of Ministers to Member States on "hate speech". Retrieved from https://rm.coe.int/1680505d5b

Cowan, G., & Khatchadourian, D. (2003). Empathy, ways of knowing, and interdependence as mediators of gender differences in attitudes toward hate speech and freedom of speech. *Psychology of Women Quarterly, 27,* 300–308. https://doi.org/10.1111/1471-6402.00110

Crandall, C. S., & Eshleman, A. (2003). A justification-suppression model of the expression and experience of prejudice. *Psychological Bulletin, 129,* 414–446. https://doi.org/10.1037/0033-2909.129.3.414

Crandall, C. S., Eshleman, A., & O'Brien, L. (2002). Social norms and the expression and suppression of prejudice: The struggle for internalization. *Journal of Personality and Social Psychology, 82,* 359–378. https://doi.org/10.1037/0022-3514.82.3.359

Crandall, C. S., & Stangor, C. (2005). Conformity and prejudice. In J. F. Dovidio, P. Glick, & A. L. Rudman (Eds.), *On the nature of prejudice: Fifty years after Allport* (pp. 295–309). Malden, MA: Blackwell.

Cuddy, A. J., Fiske, S. T., & Glick, P. (2007). The BIAS map: Behaviors from intergroup affect and stereotypes. *Journal of Personality and Social Psychology, 92,* 631–648. https://doi.org/10.1037/0022-3514.92.4.631

Cuddy, A. J., Fiske, S. T., & Glick, P. (2008). Warmth and competence as universal dimensions of social perception: The stereotype content model and the BIAS map. In M. P. Zanna (Ed.), *Advances in experimental social psychology* (Vol. *40,* pp. 61–149). New York, NY: Academic Press. https://doi.org/10.1016/S0065-2601(07)00002-0

Deutsch, M., & Gerard, H. B. (1955). A study of normative and informational social influences upon individual judgement. *Journal of Personality and Social Psychology, 51,* 629–636. https://doi.org/10.1037/h0046408

Duckitt, J. (2001). A dual-process cognitive-motivational theory of ideology and prejudice. In M. P. Zanna (Ed.), *Advances in experimental social psychology* (Vol. *33,* pp. 41–114). San Diego, CA: Academic Press. https://doi.org/10.1016/S0065-2601(01)80004-6

Duckitt, J., & Sibley, C. G. (2010). Personality, ideology, prejudice, and politics: A dual-process motivational model. *Journal of Personality, 78,* 1861–1894. https://doi.org/10.1111/j.1467-6494.2010.00672.x

Ekman, P., & Friesen, W. V. (1971). Constants across cultures in the face and emotion. *Journal of Personality and Social Psychology, 17,* 124–129. https://doi.org/10.1037/h0030377

Ekman, P., & Rosenberg, E. L. (Eds.). (1997). *What the face reveals: Basic and applied studies of spontaneous expression using the Facial Action Coding System (FACS)*. New York, NY: Oxford University Press.

Espelage, D. L., Holt, M. K., & Henkel, R. R. (2003). Examination of peer-group contextual effects on aggression during early adolescence. *Child Development, 74,* 205–220. https://doi.org/10.1111/1467-8624.00531

Fasoli, F., Maass, A., & Carnaghi, A. (2015). Labelling and discrimination: Do homophobic epithets undermine fair distribution of resources? *British Journal of Social Psychology, 54,* 383–393. https://doi.org/10.1111/bjso.12090

Fasoli, F., Paladino, M. P., Carnaghi, A., Jetten, J., Bastian, B., & Bain, P. G. (2016). Not "just words": Exposure to homophobic epithets leads to dehumanizing and physical distancing from gay men. *European Journal of Social Psychology, 46,* 237–248. https://doi.org/10.1002/ejsp.2148

Finlay, K. S., & Stephan, W. G. (2000). Reducing prejudice: The effects of empathy on intergroup attitudes. *Journal of Applied Social Psychology, 30,* 1720–1737. https://doi.org/10.1111/j.1559-1816.2000.tb02464.x

Fischer, A., & Giner-Sorolla, R. (2016). Contempt: Derogating others while keeping calm. *Emotion Review, 8,* 346–357. https://doi.org/10.1177/1754073915610439

Fischer, A., Halperin, E., Canetti, D., & Jasini, A. (2018). Why we hate. *Emotion Review, 10,* 309–320. https://doi.org/10.1177/1754073917751229

Fiske, S. T., Cuddy, A. J., Glick, P., & Xu, J. (2002). A model of (often mixed) stereotype content: Competence and warmth respectively follow from perceived status and competition. *Journal of Personality and Social Psychology, 82,* 878–902. https://doi.org/10.1037/0022-3514.82.6.878

Frenkel-Brunswik, E. (1948). A study of prejudice in children. *Human Relations, 1,* 295–306. https://doi.org/10.1177/001872674800100301

Gelber, K. (2002). *Speaking back: The free speech versus hate speech debate*. Amsterdam, the Netherlands: John Benjamins. https://doi.org/10.1075/dapsac.1

Gervais, M. M., & Fessler, D. M. (2017). On the deep structure of social affect: Attitudes, emotions, sentiments, and the case of “contempt.” *Behavioral and Brain Sciences, 40,* 1–18. https://doi.org/10.1017/S0140525X16000352

Glick, P. (2002). Sacrificial lambs dressed in wolves’ clothing: Envious prejudice, ideology, and the scapegoating of Jews. In L. S. Newman & R. Erber (Eds.), *Understanding genocide: The social psychology of the Holocaust* (pp. 113–142). London, United Kingdom: Oxford University Press. https://doi.org/10.1093/acprof:oso/9780195133622.003.0006

Glomb, T. M., & Liao, H. (2003). Interpersonal aggression in work groups: Social influence, reciprocal, and individual effects. *Academy of Management Journal, 46,* 486–496. https://doi.org/10.2307/30040640

Górska, P., & Bilewicz, M. (2015). When “a group in itself” becomes “a group for itself”: Overcoming inhibitory effects of superordinate categorization on LGBTQ individuals. *Journal of Social Issues, 71,* 554–575. https://doi.org/10.1111/josi.12128

Grabowski, J. (2009). German anti-Jewish propaganda in the General Government, 1939–1945: Inciting hate through posters, films, and exhibitions. *Holocaust and Genocide Studies, 23,* 381–412. https://doi.org/10.1093/hgs/dcp040

Greenberg, J., & Pyszczynski, T. (1985). The effect of an overheard ethnic slur on evaluations of the target: How to spread a social disease. *Journal of Experimental Social Psychology, 21,* 61–72. https://doi.org/10.1016/0022-1031(85)90006-X

Haidt, J., & Keltner, D. (1999). Culture and facial expression: Open-ended methods find more expressions and a gradient of recognition. *Cognition & Emotion, 13,* 225–266. https://doi.org/10.1080/026999399379267

Harris, L. T., & Fiske, S. T. (2007). Social groups that elicit disgust are differentially processed in mPFC. *Social Cognitive and Affective Neuroscience, 2,* 45–51. https://doi.org/10.1093/scan/nsl037

Haslam, S. A., Jetten, J., Postmes, T., & Haslam, C. (2009). Social identity, health and well-being: An emerging agenda for applied psychology. *Applied Psychology: An International Review, 58,* 1–23. https://doi.org/10.1111/j.1464-0597.2008.00379.x

Hate speech. (n.d.). *Cambridge advanced learner’s dictionary & thesaurus*. Retrieved from https://dictionary.cambridge.org/dictionary/english/hate-speech

Hawdon, J., Oksanen, A., & Räsänen, P. (2017). Exposure to online hate in four nations: A cross-national consideration. *Deviant Behavior, 38,* 254–266. https://doi.org/10.1080/01639625.2016.1196985

Jacobs, J., & Potter, K. (1998). *Hate crimes: Criminal law and identity politics*. New York, NY: Oxford University Press.

Johnson, N. F., Leahy, R., Johnson Restrepo, N. J., Velasquez, N., Zheng, M., Manrique, P., … Wuchty, S. (2019). Hidden resilience adaptive dynamics of the global online hate ecology. *Nature, 573,* 261–265. https://doi.org/10.1038/s41586-019-1494-7

Johnstone, L. (2019, February 17). “Yellow vests”: Macron slams abuse of French philosopher Alain Finkielkraut at protests. *Euronews*. Retrieved from https://www.euronews.com/2019/02/17/yellow-vests-macron-slams-abuse-of-french-philosopher-alain-finkielkraut-at-protests

Jost, J. T. (2019). A quarter century of system justification theory: Questions, answers, criticisms, and societal applications. *British Journal of Social Psychology, 58,* 263–314. https://doi.org/10.1111/bjso.12297

Jost, J. T., & Hunyady, O. (2003). The psychology of system justification and the palliative function of ideology. *European Review of Social Psychology, 13,* 111–153. https://doi.org/10.1080/10463280240000046

Jost, J. T., & Hunyady, O. (2005). Antecedents and consequences of system-justifying ideologies. *Current Directions in Psychological Science, 14,* 260–265. https://doi.org/10.1111/j.0963-7214.2005.00377.x

Kaakinen, M., Keipi, T., Oksanen, A., & Räsänen, P. (2018). How does social capital associate with being a victim of online hate? Survey evidence from the United States, the United Kingdom, Germany, and Finland. *Policy & Internet, 10,* 302–323. https://doi.org/10.1002/poi3.173

Keizer, K., Lindenberg, S., & Steg, L. (2008). The spreading of disorder. *Science, 322,* 1681–1685. https://doi.org/10.1126/science.1161405

Kempf, W. (2012). Antisemitism and criticism of Israel: A methodological challenge for peace research. *Journal for the Study of Antisemitism, 4,* 515–532.

Kirkland, S. L., Greenberg, J., & Pyszczynski, T. (1987). Further evidence of the deleterious effects of overheard derogatory ethnic labels: Derogation beyond the target. *Personality and Social Psychology Bulletin, 13,* 216–227. https://doi.org/10.1177/0146167287132007

Kuzawińska, O., Ferenc, K., Grzybowska, A., Jaworski, M., Tsypysheva, E., Wilk, W., … Bilewicz, M. (2018). *Mental health consequences of hate speech exposure among international students of the University of Warsaw*. Unpublished report.

Leader, T., Mullen, B., & Rice, D. (2009). Complexity and valence in ethnophaulisms and exclusion of ethnic out-groups: What puts the “hate” into hate speech? *Journal of Personality and Social Psychology, 96,* 170–182. https://doi.org/10.1037/a0013066

Leets, L. (2002). Experiencing hate speech: Perceptions and responses to anti-Semitism and antigay speech. *Journal of Social Issues, 58,* 341–361. https://doi.org/10.1111/1540-4560.00264

Lerner, M. J. (1980). *The belief in a just world: A fundamental delusion*. New York, NY: Plenum Press. https://doi.org/10.1007/978-1-4899-0448-5

Levenson, E., & Sanchez, R. (2018, October 28). Mass shooting at Pittsburgh synagogue. *CNN*. Retrieved from https://edition.cnn.com/us/live-news/pittsburgh-synagogue-shooting/index.html

Levy, J., Goldstein, A., Influs, M., Masalha, S., Zagoory-Sharon, O., & Feldman, R. (2016). Adolescents growing up amidst intractable conflict attenuate brain response to pain of outgroup. *Proceedings of the National Academy of Sciences, 113,* 13696–13701. https://doi.org/10.1073/pnas.1612903113

Lewis, G. J., & Bates, T. C. (2014). Common heritable effects underpin concerns over norm maintenance and in-group favoritism: Evidence from genetic analyses of right-wing authoritarianism and traditionalism. *Journal of Personality, 82,* 297–309. https://doi.org/10.1111/jopy.12055

Linz, D., Donnerstein, E., & Adams, S. M. (1989). Physiological desensitization and judgments about female victims of violence. *Human Communication Research, 15,* 509–522. https://doi.org/10.1111/j.1468-2958.1989.tb00197.x

Marshal, M. P., Dietz, L. J., Friedman, M. S., Stall, R., Smith, H. A., McGinley, J., … Brent, D. A. (2011). Suicidality and depression disparities between sexual minority and heterosexual youth: A meta-analytic review. *Journal of Adolescent Health, 49,* 115–123. https://doi.org/10.1016/j.jadohealth.2011.02.005

Mathew, B., Illendula, A., Saha, P., Sarkar, S., Goyal, P., & Mukherjee, A. (2019). *Temporal effects of unmoderated hate speech in Gab*. Retrieved from https://arxiv.org/pdf/1909.10966.pdf

Mathew, B., Saha, P., Tharad, H., Rajgaria, S., Singhania, P., Maity, S. K., … Mukherjee, A. (2019). Thou shalt not hate: Countering online hate speech. *Proceedings of the International AAAI Conference on Web and Social Media, 13,* 369–380.

Mischkowski, D., Crocker, J., & Way, B. M. (2016). From painkiller to empathy killer: Acetaminophen (paracetamol) reduces empathy for pain. *Social Cognitive and Affective Neuroscience, 11,* 1345–1353. https://doi.org/10.1093/scan/nsw057

Mortensen, C. R., Neel, R., Cialdini, R. B., Jaeger, C. M., Jacobson, R. P., & Ringel, M. M. (2019). Trending norms: A lever for encouraging behaviors performed by the minority. *Social Psychological and Personality Science, 10,* 201–210. https://doi.org/10.1177/1948550617734615

Mullen, B., & Rice, D. R. (2003). Ethnophaulisms and exclusion: The behavioral consequences of cognitive representation of ethnic immigrant groups. *Personality and Social Psychology Bulletin, 29,* 1056–1067. https://doi.org/10.1177/0146167203254505

Mullen, B., & Smyth, J. M. (2004). Immigrant suicide rates as a function of ethnophaulisms: Hate speech predicts death. *Psychosomatic Medicine, 66,* 343–348. https://doi.org/10.1097/01.psy.0000126197.59447.b3

Nowak, A., Szamrej, J., & Latané, B. (1990). From private attitude to public opinion: A dynamic theory of social impact. *Psychological Review, 97,* 362–376. https://doi.org/10.1037/0033-295X.97.3.362

Oksanen, A., Hawdon, J., Holkeri, E., Näsi, M., & Räsänen, P. (2014). Exposure to online hate among young social media users. *Sociological Studies of Children & Youth, 18,* 253–273. https://doi.org/10.1108/S1537-466120140000018021

Oksanen, A., Kaakinen, M., Minkkinen, J., Räsänen, P., Enjolras, B., & Steen-Johnsen, K. (2018). Perceived societal fear and cyberhate after the November 2015 Paris terrorist attacks. *Terrorism and Political Violence*. https://doi.org/10.1080/09546553.2018.1442329

Oliner, S. P., & Oliner, P. M. (1988). *The altruistic personality: Rescuers of Jews in Nazi Europe*. New York, NY: Free Press.

Postmes, T., Spears, R., Sakhel, K., & de Groot, D. (2001). Social influence in computer-mediated communication: The effects of anonymity on group behavior. *Personality and Social Psychology Bulletin, 27,* 1243–1254. https://doi.org/10.1177/01461672012710001

Rankin, C. H., Abrams, T., Barry, R. J., Bhatnagar, S., Clayton, D. F., Colombo, J., … McSweeney, F. K. (2009). Habituation revisited: An updated and revised description of the behavioral characteristics of habituation. *Neurobiology of Learning and Memory, 92,* 135–138. https://doi.org/10.1016/j.nlm.2008.09.012

Reicher, S. D., Spears, R., & Postmes, T. (1995). A social identity model of deindividuation phenomena. *European Review of Social Psychology, 6,* 161–198. https://doi.org/10.1080/14792779443000049

Roets, A., Au, E. W., & Van Hiel, A. (2015). Can authoritarianism lead to greater liking of out-groups? The intriguing case of Singapore. *Psychological Science, 26,* 1972–1974. https://doi.org/10.1177/0956797615605271

Schieb, C., & Preuss, M. (2016, June). *Governing hate speech by means of counterspeech on Facebook*. Paper presented at the 66th ICA Annual Conference, Fukuoka, Japan.

Sedler, R. (1992). The unconstitutionality of campus bans on “racist speech”: The view from without and within. *University of Pittsburgh Law Review, 53,* 631–683.

Short, J. A., Williams, E., & Christie, B. (1976). *The social psychology of telecommunications*. London, United Kingdom: John Wiley & Sons.

Simon, L., & Greenberg, J. (1996). Further progress in understanding the effects of derogatory ethnic labels: The role of preexisting attitudes toward the targeted group. *Personality and Social Psychology Bulletin, 22,* 1195–1204. https://doi.org/10.1177/01461672962212001

Soral, W., Bilewicz, M., & Świderska, A. (2020). *The speech of contempt. Emotional response to derogatory language*. Unpublished research report.

Soral, W., Bilewicz, M., & Winiewski, M. (2018). Exposure to hate speech increases prejudice through desensitization. *Aggressive Behavior, 44,* 136–146. https://doi.org/10.1002/ab.21737

Soral, W., Bilewicz, M., Winiewski, M., & Bulska, D. (2020). *Family acceptance as a buffer against hate-speech induced depression*. Manuscript submitted for publication.

Soral, W., Liu, J. H., & Bilewicz, M. (in press). Media of contempt. Social media consumption increases normativity of xenophobic verbal violence. *International Journal of Conflict and Violence*.

Sparkman, G., & Walton, G. M. (2017). Dynamic norms promote sustainable behavior, even if it is counternormative. *Psychological Science, 28,* 1663–1674. https://doi.org/10.1177/0956797617719950

Sternberg, R. J. (2005). Understanding and combating hate. In R. J. Sternberg (Ed.), *The psychology of hate* (pp. 37–50). Washington, DC: APA Books. https://doi.org/10.1037/10930-000

Still, R. J. (1961). Mental health in overseas students. *Proceedings of the British Student Health Association, 13,* 59–73.

Straus, S. (2006). Rwanda and Darfur: A comparative analysis. *Genocide Studies and Prevention, 1,* 41–56. https://doi.org/10.1353/gsp.2011.0009

Tak prawicowcy komentowali Marsz Niepodległości. "Co za ćwoki?," "patrzę i nie wierzę." (2017, November 13). *Gazeta.pl*. Retrieved from http://wiadomosci.gazeta.pl/wiadomosci/56,114884,22641248,tak-prawica-komentowala-marsz-niepodleglosci.html

Tausch, N., & Becker, J. C. (2013). Emotional reactions to success and failure of collective action as predictors of future action intentions: A longitudinal investigation in the context of student protests in Germany. *British Journal of Social Psychology, 52,* 525–542. https://doi.org/10.1111/j.2044-8309.2012.02109.x

Tausch, N., Becker, J. C., Spears, R., Christ, O., Saab, R., Singh, P., & Siddiqui, R. N. (2011). Explaining radical group behavior: Developing emotion and efficacy routes to normative and nonnormative collective action. *Journal of Personality and Social Psychology, 101,* 129–148. https://doi.org/10.1037/a0022728

Terlikowski: Czy Jezus przykleiłby na drzwiach domu Piotra naklejkę "Strefa wolna od LGBT"? (2019, July 20). *Dziennik.pl*. Retrieved from https://wiadomosci.dziennik.pl/opinie/artykuly/603078,strefa-wolna-od-lgbt-naklejki-terlikowski.html

Thompson, R. F., & Spencer, W. A. (1966). Habituation: A model phenomenon for the study of neuronal substrates of behavior. *Psychological Review, 73,* 16–43. https://doi.org/10.1037/h0022681

Vanman, E. J. (2016). The role of empathy in intergroup relations. *Current Opinion in Psychology, 11,* 59–63. https://doi.org/10.1016/j.copsyc.2016.06.007

Walker, S. (1994). *Hate speech: The history of an American controversy*. Lincoln, NE: Bison Books.

Winiewski, M., Hansen, K., Bilewicz, M., Soral, W., Świderska, A., & Bulska, D. (2017). *Contempt speech, hate speech. Report from research on verbal violence against minority groups*. Warsaw, Poland: Stefan Batory Foundation. Retrieved from http://www.ngofund.org.pl/wp-content/uploads/2017/02/Contempt_Speech_Hate_Speech_Full_Report.pdf

Winiewski, M., Soral, W., & Bilewicz, M. (2015). Conspiracy theories on the map of stereotype content: Survey and historical evidence. In M. Bilewicz, A. Cichocka, & W. Soral (Eds.), *The psychology of conspiracy* (pp. 23–41). London, United Kingdom: Routledge.

van Noorden, T. H., Haselager, G. J., Cillessen, A. H., & Bukowski, W. M. (2015). Empathy and involvement in bullying in children and adolescents: A systematic review. *Journal of Youth and Adolescence, 44,* 637–657. https://doi.org/10.1007/s10964-014-0135-6

van Zomeren, M., Postmes, T., & Spears, R. (2008). Toward an integrative social identity model of collective action: A quantitative research synthesis of three socio-psychological perspectives. *Psychological Bulletin, 134,* 504–535. https://doi.org/10.1037/0033-2909.134.4.504

Zaki, J., & Cikara, M. (2015). Addressing empathic failures. *Current Directions in Psychological Science, 24,* 471–476. https://doi.org/10.1177/0963721415599978