THE DYNAMICS OF HATE SPEECH AND COUNTER SPEECH IN THE SOCIAL MEDIA

SUMMARY OF SCIENTIFIC RESEARCH

Dr. Katarzyna Bojarska[1]

INTRODUCTION

Prejudice and hate speech have been observed throughout history and their dynamics and consequences have been scientifically researched, described and explained for decades before the digital era. Hostile portrayals and stereotyping of groups and minorities as „other, „different“ or dangerous can lead to dehumanization. This effect can escalate rapidly when hostile rhetoric reaches a large audience by means of broadcast, print or digital media and lead to real-life violent hate crimes, including genocide.

In recent years Europe has witnessed a significant increase of xenophobic, nationalist, Islamophobic, racist and anti-semitic attitudes. Their effects are not solely restricted to hostile rhetoric, instead, they turn into actual crimes against groups and individuals. As Heiko Maas, the German Minister of Justice puts it: „Hate speech is often not restricted to mere act of hateful speech. It often moves from words to deeds. The fact that "mental incitement" too often turns into violence can be seen in the surge in attacks on refugee shelters: in 2014, the number of acts tripled in comparison with the previous year“ (Maas, 2015, p. 6).

While the media dissemination channels evolve over time, the mechanisms of group-based hostility remain the same. Social media, such as Facebook or Twitter, have introduced new modes of social discursive participation and their users now contribute to dissemination of prejudice, fake news and hostility against refugees or other marginalized groups on an unprecedented scale. The dissemination of hate in digital media constitutes therefore a social emergency with real-life individual, political and social consequences.

A recent internet survey revealed that a majority (67%) of internet users reports having encountered hate speech or hateful comments online (forsa., 2017). Although new German law, Network Enforcement Act (NetzDG), which came into force in 2017, obliges social media sites to remove hate speech, fake news and illegal material within a short time after the illegal material has been reported, it does not cover all hate speech. Furthermore, owing to the fact, that different countries have different regulations and hate speech extends over any national borders, strategies are still needed to increase visibility of internet users' objection to online hate.

The primary purpose of this report is to summarize the current state of scientific knowledge that might help inform strategies to more effectively counter instances of hate speech online and make democratic message more visible. The focus of the present report is therefore to compare dynamics of hate speech and of counter-speech, which refers to any response to hateful content aiming to defy it. As Bartlett and Krasodomski-Jones define it, counter-speech is “a common, crowd-sourced response to extremism or hateful content” (2015, p. 5).

For a better understanding of the challenges of counter-speech online, hate speech will be first defined, followed by discussion of its legal status in Germany. Social and individual harms of hate speech will be discussed. We will examine available research findings on how and why hate speech spreads in social media and attempt to explain the

[1] Centre for Internet and Human Rights, Europa-Universität Viadrina

different dynamics of online hate speech and counter-speech transmission. Finally, we will point to available resources for effective counter-speech strategies.

HATE SPEECH DEFINITION AND ITS LEGAL STATUS

Hate speech can be defined as an expression of hostility toward individuals or social groups based on their perceived group membership, which can refer to their race, ethnicity, nationality, religion, disability, gender or sexual orientation. The Council of Europe defines hate speech as "all forms of expression which spread, incite, promote or justify racial hatred, xenophobia, antisemitism or other forms of hatred based on intolerance, including: intolerance expressed by aggressive nationalism and ethnocentrism, discrimination and hostility against minorities, migrants and people of immigrant origin" (Europarat Ministerkomitee, 1997)[2].

While the term hate speech (Hassrede) itself is not recognized by German law, the relevant legal term of incitement to hatred (Volksverhetzung) has been long recognized as a punishable criminal offence, regardless of whether committed online or offline. According to the § 130 Section 1 of the German Criminal Code:

„Whosoever, in a manner capable of disturbing the public peace:

1. incites hatred against a national, racial, religious group or a group defined by their ethnic origins, against segments of the population or individuals because of their belonging to one of the aforementioned groups or segments of the population or calls for violent or arbitrary measures against them; or
2. assaults the human dignity of others by insulting, maliciously maligning an aforementioned group, segments of the population or individuals because of their belonging to one of the aforementioned groups or segments of the population, or defaming segments of the population,

shall be liable to imprisonment from three months to five years."[3]

On the 1st October 2017 a new law, Network Enforcement Act, Netzwerkdurchsetzungsgesetz (NetzDG), came into force in Germany, which obliges social media sites to remove hate speech, fake news and illegal material within 24 hours after the illegal material has been reported. The new law has been however criticized for demanding social media companies, rather than courts, to decide,

[2] Der Begriff Hate Speech (Hassrede) "umfasst jegliche Ausdrucksformen, welche Rassenhass, Fremdenfeindlichkeit, Antisemitismus oder andere Formen von Hass, die auf Intoleranz gründen, propagieren, dazu anstiften, sie fördern oder rechtfertigen, einschließlich der Intoleranz, die sich in Form eines aggressiven Nationalismus und Ethnozentrismus, einer Diskriminierung und Feindseligkeit gegenüber Minderheiten, Einwanderern und der Einwanderung entstammenden Personen ausdrücken"

[3] § 130 Absatz 1 des Strafgesetzbuchs: „Wer in einer Weise, die geeignet ist, den öffentlichen Frieden zu stören,

1. gegen eine nationale, rassische, religiöse oder durch ihre ethnische Herkunft bestimmte Gruppe, gegen Teile der Bevölkerung oder gegen einen Einzelnen wegen seiner Zugehörigkeit zu einer vorbezeichneten Gruppe oder zu einem Teil der Bevölkerung zum Hass aufstachelt, zu Gewalt- oder Willkürmaßnahmen auffordert oder
2. die Menschenwürde anderer dadurch angreift, dass er eine vorbezeichnete Gruppe, Teile der Bevölkerung oder einen Einzelnen wegen seiner Zugehörigkeit zu einer vorbezeichneten Gruppe oder zu einem Teil der Bevölkerung beschimpft, böswillig verächtlich macht oder verleumdet,

wird mit Freiheitsstrafe von drei Monaten bis zu fünf Jahren bestraft.".

whether the reported content is illegal or not, without any judicial oversight. This entails the risk of unaccountable censorship and undermining free speech and can set a bad example for other countries to silence political criticism by similar laws.

The controversy over the German anti-hate crime law fits well with a wider political and academic freedom of speech vs. hate speech debate that has been running for at least a decade. In fact, the bulk of academic peer-reviewed publications[4], are of legal nature and focus on the legal status of hate speech and on whether, to which extent, by which tools and technologies and by whom hate speech should be regulated. This debate, however, falls beyond the scope of this report, which instead aims is to focus on counter-speech strategies that might be useful for individual social media users rather than legal measures to be implemented by governments.

HARMS OF HATE SPEECH

Hate speech can be considered harmful at several levels. It has potential of disturbing social peace in that exposure to hate speech shapes attitudes and influences actual behaviors (Müller & Schwarz, 2018), including serious hate crimes such as genocide (cf. Fyfe, 2017; Maravilla, 2008). Online hate may constitute a fertile ground for even more hate, in that it provides a model, a permission (Brodnig, 2016; Clay, 2017), a "social proof" of "appropriate" attitudes and behaviors (cf. Anderson, Brossard, Scheufele, Xenos, & Ladwig, 2014), desensitizes the public to verbal violence and increases prejudice (Soral, Bilewicz, & Winiewski, 2018), rewarding its followers with social acceptation while punishing and silencing voices of objection (Brodnig, 2016; Coustick-Deal, 2017). Above all, hate speech poses a threat to physical safety and psychological well-being of targeted group members (Baldauf, Banaszczuk, Koreng, Schramm, & Stefanowitsch, 2015a; Coustick-Deal, 2017; Gelber & McNamara, 2016). Several among aforementioned studies warrant more in-depth discussion.

The twentieth century has witnessed the role of mass media (e.g. broadcasting and print media) in spreading hate, resulting in escalation of dehumanization and leading to hate crimes, the most extreme of which were genocides, such as Holocaust and genocide in Rwanda (Fyfe, 2017; Maravilla, 2008). A recent study by Müller and Schwarz (2018) strongly suggests the same mechanism to be true for the role of the XXI century's digital media. The study clearly demonstrates the link between exposure to hate speech in social media and real-life violence. The authors applied a sound methodology to attempt some causative inferences on how hateful anti-refugee social media activity on the Facebook page of the German Alternative for Germany (Alternative für Deutschland, AfD) party translates into actual violent acts against refugees. The authors conclude: "Using these measures, we find that anti-refugee hate crimes increase disproportionally in areas with higher Facebook usage during periods of high anti-refugee sentiment online. This effect is especially pronounced for violent incidents against refugees, such as arson and assault. Taken at face value, this suggests a role for social media in the transmission of Germany-wide anti-refugee sentiment" (p. 3). In order to rule out potential uncontrolled factors, the researchers also provided quasi-experimental support to the interpretation of their findings. They found out that in the weeks of sizable local internet disruptions, which limited the internet access of local users, the higher anti-refugee sentiment's effect on hate crimes was significantly reduced as compared to the

[4] Publications indexed by EBSCO database as of end of May 2018.

municipalities unaffected by internet outages. Also, at the Germany-wide level, the authors observed that "the effect of refugee posts on hate crimes essentially vanishes in weeks of major Facebook outages" (p. 4). In the light of the above, the role of social media in incitement to violent hate crimes, hence in affecting violent behavior, seems indisputable.

It is also widely acknowledged that hate speech poses a serious threat to the physical safety of the members of the targeted groups. According to Amadeu Antnio Stiftung's chronicle of anti-refugee incidents („Chronik flüchtlingsfeindlicher Vorfälle", o. J.), there were 1249 reported attacks against asylum-seeking individuals or their lodgings in Germany in the year 2015, 3769 in 2016 and 1939 in 2017.

While much has been written about hate speech, its legal status, its types and its perpetrators, what is striking, is the scarcity of published research on the psychological harms of online hate speech for targeted individuals. Therefore, we need to attempt, at least partially, to extrapolate from existing research on more general psychological effects of being targeted by prejudice. The most profound effect of group-based prejudice on targeted individuals is probably elevated drainage of emotional resources in comparison to unaffected individuals, associated with constant necessity of dealing with overt discrimination as well as with microaggressions[5], both prevalent in everyday life, and augmented in the digital world due to the *online disinhibition effect*, i.e. absence of restraints in online communication in comparison to face to face communication.

Constantly increased vigilance and mental preparedness to deal with or respond to overt prejudice or microaggressions translate into chronically elevated level of stress, so called *minority stress*, which can lead to adverse health outcomes, such as depression or anxiety (Meyer, 1995, 2003).

Being affected by hate speech as a member of a targeted group is associated with significant emotional strain (Coustick-Deal, 2017; Gelber & McNamara, 2016; Mullen & Smyth, 2004). The feeling of injustice, helplessness, anxiety and threat can be listed among the psychological effects of hate speech. Since being targeted by prejudice in and of itself constitutes a source of significant distress, the decision to directly confront instances of online hateful behavior might turn out to be too much of emotional effort to endure for members of the targeted populations.

Gelber and McNamara, who conducted an exceptional qualitative study "Evidencing the harms of hate speech" (2016), list the following types of hate speech harms experienced by targeted individuals: unfairly ranking target persons as inferior, silencing the victims, distress, risk of destruction to one's self-esteem, restrictions on freedom of movement and association, harms to dignity, maintenance of power imbalances within social hierarchies of race, making onlookers to believe negative stereotypes that lead them to engage in harmful conduct, normalization of expressing negative stereotypes and discriminatory behavior and encouraging the public to imitate the hateful behavior.

[5] Sue defines microaggressions as "the brief and commonplace daily verbal, behavioral, and environmental indignities, whether intentional or unintentional, that communicate hostile, derogatory, or negative racial, gender, sexual-orientation, and religious slights and insults to the target person or group. Perpetrators are usually unaware that they have engaged in an exchange that demeans the recipient of the communication" (Sue, 2010).

SOCIAL PROOF

The recent rise in xenophobia, islamophobia and anti-refugee sentiments in numerous European countries has been hard to overlook even for casual observers. The radicalization of attitudes coincided with media reports on the rising number of refugees and asylum-seeking individuals entering European Union. The United Nations Regional Information Centre for Western Europe reports: “In some countries the refugee crisis sparked an outpouring of solidarity and many local volunteers together with central authorities were committed to making the newcomers arriving in their towns feel welcome. In other countries, however, the opposite happened and restrictive border policies combined with a toxic rhetoric have created an openly hostile environment for refugees and migrants” (United Nations Regional Information Centre for Western Europe, 2016).

The prevalent tone of media coverage on initially unfamiliar circumstances or early interpretations of their meaning by the public might dictate the prevalent response and model resulting attitudes. This might explain the divergent ways the formation of social attitudes toward the same issue can take in different countries or subpopulations. If what develops is the atmosphere of permission to hatred, it may lead to even more hatred.

This brings us to the classic research on social influence and conformity in modelling patterns of “appropriate” behavior, in other words – to normative social influence and informational social influence (social proof). We speak of *normative social influence* when we conform to the behavior of others in order to gain acceptance and to be liked (Asch & Guetzkow, 1951; Asch, 1956; Aronson, Wilson, & Akert, 2010). *Social proof* or *informative social influence* is, in turn, a psychological phenomenon which occurs in unfamiliar or ambiguous situations in which we mimic behavior of others, because we don’t know what the appropriate behavior should be and we assume that others behave in certain way, because they possess more knowledge than us (Sherif, 1935; Baron, Vandello, & Brunsman, 1996; Aronson u. a., 2010). In the context of the role of online hate speech, the normative and informational social influence seem to provide a plausible explanation for the rapid formation of attitudes and behavioral patterns in response to media coverage of initially unfamiliar topics.

EXPOSURE TO HATE SPEECH INCREASES MISTRUST, FAMILIARITY IMMUNIZES AGAINST THIS EFFECT

An interesting experiment conducted by Anderson, Brossard, Scheufele, Xenos and Ladwig (2014) demonstrates how offensive speech in digital media may contribute to shaping hostile attitudes and suggests, what could make us more immune to hate speech. The authors investigated the role of “incivility”, i.e. of offensive way of expressing opinions, on formation of risk perceptions of nanotechnology, a topic unfamiliar to most internet users. They designed an experiment, in which participants were asked to read a scientific blog entry on the benefits and risks of nanosilver, followed by artificially crafted user comments formulated either in offensive or polite language. They were then asked to assess the risk associated with the new technology.

The researchers made several interesting observations. Firstly, they found out that regardless of the offensiveness of what seemed as the apparent public reception, participants with preexisting familiarity with the topic perceived the risks of nanotechnology as lower than participants, for whom the issue was initially unfamiliar. Furthermore, those who considered themselves

less able of informed judgement of nanotechnology approached the subject with more caution and also perceived the risk as higher. The study also revealed, that participants with preexisting positive attitudes toward nanotechnology, when exposed to offensive comments, still perceived the risks of nanotechnology as lower than participants, whose initial attitudes weren't positive. Such differences were not apparent among participants exposed to comments formulated in a polite way, suggesting the latter might have been more influenced by the article's actual content than by emotional remarks of other readers. Similar observation was made in relation to participants' religiosity. More religious participants exposed to offensive comments perceived the risks of nanotechnology as higher than less religious readers. This effect disappeared among participants exposed to polite discussion. The authors suggest this effect might be explained by religious value judgments of the nanotechnology as disturbing "natural order" (which perhaps could be activated more easily by exposure to more emotional, rather than calm tone of the discussion). To conclude: preexistent familiarity with the subject, preexisting positive attitudes and a low level of (activation of) religious identification, each of them independently, contributed to participants' immunity to the influence of offensive language on attitudes and encouraged them to form more favorable perceptions of risks of nanotechnology. On the other hand, lack of preexistent knowledge, lack of initial positive attitudes (and, similarly, high religiosity) made them more susceptible to the influence of offensive comments, which facilitated development of distrust toward unfamiliar phenomenon.

Social influence might explain the various trajectories the development of attitudes toward refugees or other minority groups can take. Depending on whether our initial encounter with an unfamiliar subject is of informational or highly affective nature, depending on the sources of information we draw our knowledge from and consider reliable, depending on the kind of behaviors we observe around us and treat as a social proof to guide our own attitudes and behaviors, depending on the strength of our motivation to learn about a specific subject or challenge our beliefs, we may develop favorable or hostile attitudes, confidence or fear, and act accordingly.

FEAR FUELS HATE, FREEDOM FROM CONSEQUENCES FACILITATES HATEFUL EXPRESSION

The role of fear, threat and uncertainty in acquisition and maintaining of conservative attitudes has been demonstrated in several studies. In 2003 Jost, Glaser, Kruglanski and Sulloway concluded their meta-analytic review of political attitudes: "people embrace political conservatism (at least in part) because it serves to reduce fear, anxiety, and uncertainty; to avoid change, disruption, and ambiguity; and to explain, order, and justify inequality among groups and individuals" (Jost, Glaser, Kruglanski, & Sulloway, 2003). Nail and McGregor observed, accordingly, significant shift toward the political right both among conservatives and liberals when they tested them two months after the terrorist attacks on 9/11 (Nail & McGregor, 2009) in comparison to a year before. The same researchers in extended team experimentally instilled threat in participants and as a result obtained conservative shift in their attitudes (Nail, McGregor, Drinkwater, Steele, & Thompson, 2009). In 2017 Napier, Huang, Vonasch and Bargh experimentally demonstrated the opposite effect – having reduced fear in conservatives, they observed their social attitudes progress in a more liberal direction (Napier, Huang, Vonasch, & Bargh, 2017).

It is probably the underlying fear and uncertainty as opposed to sympathy, hope or despair, which may explain different dynamics of hate speech and

counter speech in the digital media. Scientific research provides support to casual observations that anger spreads online more effortlessly than joy or sadness (Fan, Zhao, Chen, & Xu, 2014) and that hate spreads easier than positive emotions or counter-speech do (Bartlett & Krasodomski-Jones, 2015). As already mentioned, right-wing attitudes and resulting hate speech seem to be at least partially fueled by fear of the unfamiliar and by the need for safety (Napier u. a., 2017). When an act of online hate speech is motivated by underlying fear and mistrust, it constitutes a spontaneous act arising from the need to protect oneself and from an urge to warn and convince others to do the same. Posting hateful content can help to alleviate emotional strain and evoke the feeling of satisfaction by virtue of serving in a good and righteous cause.

Furthermore, there is the *online disinhibition effect*, or reduced empathy which easily occurs when we are posting content while hidden safely behind the screen of our devices and free from consequences of our hurtful actions which would otherwise have to be faced in real-life interactions (Brodnig, 2016; Terry & Cain, 2016). Terry and Cain attribute the effect to the anonymity of both the user and the target, asynchronous communication and invisibility:

"First, the anonymity associated with computer-mediated communication may permit people to possess an alternate online identity and essentially hide behind a non-identifying pseudonym or username. This form of dissociative anonymity allows people to separate from in-person identity and moral agency, thereby freeing them to express hostility and criticism without any effect to the psyche. Similarly, online users may dissociate t hose at the other end of the communication by subconsciously viewing them merely as avatars or usernames instead of actual persons. Second, as online communication can be asynchronous, individuals do not have to manage immediate reactions to online conversations and can remove themselves from the repercussions of online discussions, even avoiding ownership for hostile and intimidating comments. Third, even in a completely non-anonymous environment (i.e., computerized medical record, e-mail correspondence, blogs), the nature of online communications is such that individuals are physically invisible to others, permitting them to disregard any type of eye contact or physical reaction of the other person(s). A significant portion of traditional face-to-face communications tends to be nonverbal (e.g., body language, tone of voice), and without these cues, online conversations lack an essential element of understanding" (2016, p. 2).

IMBALANCE IN COGNITIVE EFFORT IN EXPRESSING HATE AND COUNTERING IT

A direct online expression of hate speech or an act of sharing a hateful post is usually impulsive, careless, internally motivated and does not involve significant cognitive or emotional effort. Indeed, it might involve more effort to suppress a hateful or angry feeling than to release it. Unlike hate speech, an act of counter speech is not spontaneous, but responsive, not active, but reactive. It requires conscious decision and involves considerable cognitive and emotional effort in that, rather than with carelessness, it is more often associated with awareness of the potential consequences of direct confrontation with the hater, such as attracting their attention and being targeted by insults and even more hate personally. Hence, highly unpleasant consequences. In short, a decision to counter an act of hate speech requires usually disproportionate amount of emotional effort and resources as compared to the impulsive, self-rewarding and affective act of posting or sharing a hateful post (cf. Coustick-Deal, 2017). This might explain the restraint of many internet users who remain silent when exposed to hate speech.

According to free speech advocates as well as Facebook's official stance, counter-speech is supposed to be a more effective tool against hate speech than removing offensive content by websites administrators (Bartlett & Krasodomski-Jones, 2015, p. 4). Coustick-Deal puts her concerns this way: such rhetoric "doesn't take into account power imbalances and privilege. (…) The way 'counter speech' is advocated is as though there is some kind of balance which works like this: (…) Racists speak = racists listen to their victims. (…) However, counter speech is actually only afforded to those who have voices to begin with. It's more like: Nazi speaks -> thousands of his supporters speak with him -> his opponents are attacked. There is no balance when someone replies to your speech by threatening to kill your family" (2017). She than points out to the "unseen forces that stop a person from being able to speak at all. (…) Seeing people harassed stops members of that same group from speaking out. When we talk about surveillance, we also use the phrase 'chilling effect'—and harassment operates in much the same manner. The knowledge that we are under constant surveillance stops us from expressing ourselves freely. This same censoring effect happens through harassment, when the fear of abuse silences us".

FILTER BUBBLES ARE HARD TO BRAKE

Indeed, Bartlett and Krasodomski-Jones found out that counter-speech activity on Facebook has a strikingly lower potential to reach wider audience than hate speech (2015). They analyzed 27,886 posts uploaded over a two-month period on 150 public hate speech (124) and counter-speech (26) pages, mostly from the UK, France and Italy, with 25,522 and 2,364 posts, respectively. They also collected 8.4 million associated interactions, i.e. likes, shares and comments. Their findings are described in their report entitled: "Counter-speech: Examining content that challenges extremism online". Among four types of posts (links, photos, statuses, videos), photos were the type most interacted with. The most popular tone of posts on right wing pages was celebratory, followed by an angry one. On counter-speech pages the most popular tone was funny/satirical. This might suggest, that beside anger, which is the most virally spreading emotion in the digital media, humor might be the winner within the subset of positive emotions.

Each time a user interacts with a piece of content on a public Facebook page, it is more likely to appear in their friends' timeline (depending on the privacy settings applied). This creates opportunity for other users, who are not group members or page followers, to interact with the shared content too. Exposure to social media content, including posts shared by Facebook friends, is, however, regulated by algorithms which predict for each user and display in their newsfeed mostly the content that is most likely to be of interest to them, based on previous usage history (the same rule is true also for search engines on the internet). As a result, *filter bubbles* and *echo chambers* are created. Within them, we are mostly fed content and we witness behavior of others that resemble our own beliefs, while being isolated from views that re different from our own. This creates impression that most of other people share our own beliefs.

Bartlett and Krasodomski-Jones (2015) examined the hateful and counter-speech content's spread from hate speech and counter-speech pages to individual newsfeeds by calculating proportion of posts that had reached users who did not like the page at which given content was originally posted. They concluded: "populist right wing pages are significantly more effective at posting content which goes beyond their network of page fans. For counter-speech pages (and populist right wing pages) videos are the most effective type of content to post to reach a broader audience". Populist right-wing links, photos, videos and statuses received on average 50%, 52%, 68% and 21% of their likes from

people who didn't subscribe to the original page, as compared to only 18%, 5%, 26% and 7%, respectively, for content from the counter-speech pages. Similar applies to the proportion of comments.

Unlike negative affect which underlies hate speech and motivates users to share hateful content or opinions, messages of support for equal rights, peaceful coexistence and empathy apparently awake less interest in internet users. Aside from the fear of confrontation and of personal exposure to insults one of the barriers to speaking up, there might be other reasons preventing users from sharing pro-democratic, pro-diversity, anti-racist content as well. For instance, many regular users might perceive democratic message as so obvious that there is little point in further sharing it. Especially, living inside of the filter bubble might augment this effect – one might believe democratic values are obvious for almost everyone. While hateful content often serves as a tool to warn others from, or remind them of perceived social threats and, consequently, spreads rapidly, pro-democratic content might be perceived as less sensational, hence, less affectively loaded. Specifically, it seldom induces fear which would be a much stronger motivator for action than positive affect or sympathy (Fan u. a., 2014; Nail & McGregor, 2009; Nail u. a., 2009).

GUIDELINES FOR EFFECTIVE COUNTER-SPEECH

In their report, "Considerations for successful counterspeech", Benesch, Ruths, Dillon, Saleem and Wright (2016), provide two meanings for how "successful counterspeech" can be understood (on Twitter): "The first is speech (text or visual media) that has a favorable impact on the original (hateful) Twitter user, shifting his or her discourse if not also his or her beliefs. This is usually indicated by an apology or recanting, or the deletion of the original tweet or account. The second type of success is to positively affect the discourse norms of the 'audience' of a counterspeech conversation: all of the other Twitter users or 'cyberbystanders' who read one or more of the relevant exchange of tweets. This impact is difficult to assess when studying counterspeech "in the wild" as we have, but it may be indicated by long conversations that remain civil, and by counterspeech that leads to others counterspeaking" (p. 2). Depending on the approach, different strategies might prove useful.

Strikingly little academic research on successful counter-speech strategies has been published to date. Most available (and valuable) guidelines, DOs and DON'Ts, are collections of experience-based observations and conclusions made by digital media activists operating within various anti-hate speech projects, rather than effects of rigorously controlled scientific analyses. They are available in form of downloadable publications or on websites.

I will now try to draw some research-based suggestions for successful counter-speech and for prevention of the online spread of hate speech. Finally, I will point to some more useful resources and recommendations.

TO AFFECT THE HATER'S BEHAVIOR

ENCOURAGE ONLINE CIVILITY BY REMEMBERING AND REMINDING OTHER USERS ABOUT THE HUMANITY OF PERSON(S) TARGETED BY HATE: Munger (2017) tested experimentally, how online sanctioning of racial hatred affects hateful white male Twitter users. He used bots to tweet a reminder: "Hey man, just remember that there are real people who are hurt when you harass them with that kind of language" after selected users used racial slurs, while manipulating the bots' perceived race and social status and keeping the bots' perceived male gender constant. Then he analyzed racist activity of the treated users over a period of one month. Bots appearing as white men of high status (as indicated by large number of followers) achieved the longest

lasting change by significantly reducing future racial slur usage among treated users. Unfortunately, the effects achieved by bots whose profile picture depicted a black male and by bots with lower social status were less lasting. The author explains the effect by attributing it to "in-group" identification, without mentioning white male privilege as an alternative hypothesis. Exclusion of women from the experiment and the lack of women-treating-women, women-treating-men and men-treating-women conditions makes it difficult to determine the exact nature of the effect.

USE YOUR PRIVILEGE FOR A GOOD CAUSE: In accordance with the results mentioned above, and being mindful of Coustick-Deal's argumentation (2017) that not everyone has power to defend themselves from hate speech, whether you are white or male or straight or cisgender, use your white or male or straight or cis privilege to counter hatred by clearly expressing expectation that social norm of treating other users with respect be observed. In other words, if you are not from one of the disadvantaged groups, you can help the cause by reminding hateful users of the humanity of targeted persons you help to reduce the online disinhibition effect and promote healthy culture of online debate.

CONFRONT PROBLEMATIC ONLINE ACTIVITY BY PRIVATE MESSAGING: This kind of approach has been mentioned by Rafael, Dinar and Heyken (Dinar & Rafael, 2017; Rafael, Dinar, & Heyken, 2017). Public shaming poses a threat to the self-esteem of the shamed person, which might induce strong motivation to publicly resist the confrontation to restore the self-esteem. The aim becomes – to defend one's initial stand, as publicly changing mind might be interpreted as losing face, especially if the tone of the public confrontation is aggressive. The intervention might become easier and might bring more lasting effect if it is more personal, concerned and private. One way to achieve this would be to express concern by sending the problematic user a private message, sparing them public shame. This could be done in a similar manner as in the experiment above, by reminding them about the humanity of the targeted persons and informing about real-life harm in being targeted by hate. For instance: "Hi X, I've seen you have posted a meme [describe the meme]. I was wondering, why you shared it. I know you are a mindful person and just wanted to let you know that sharing such content hurts actual people. It costs nothing to post it, but there are people in the real world who are beaten and insulted because of others spreading this kind of memes. It negatively affects their/our everyday life. I would appreciate it if you could remove that post and wouldn't post anything similar in the future". This technique should work better for users who are not (yet) radical right-wing advocates.

TO AFFECT THE ONLINE-BYSTANDERS

REDUCE FEAR: It has been experimentally demonstrated that reduction of fear is able to shift attitudes of conservatives toward more liberal stances, and vice versa. This is quite a general suggestion, for fear reduction can take various forms. Fear reduction as a strategy to prevent spread of hate is likely to be more effective among users who do not (yet) strongly identify with right wing values and beliefs.

PROVIDE SOCIAL PROOF BY ENCOURAGING FRIENDS TO VISIBLY SUPPORT COUNTER-SPEAKERS: Filter bubbles aren't helpful in making counter-speech efforts viral. Unfortunately, social media spread of anti-hate content is less likely than hateful content to reach audience outside of the filter bubble of like-minded individuals (Bartlett & Krasodomski-Jones, 2015). As the result, counter-speech becomes less visible. To increase the chances of the content leaving the filter bubble, setting an example or providing a social proof might be a strategy worth trying, especially – but not exclusively – on public Facebook pages.

It has been argued that only some users have enough courage, privilege, power or motivation to be able to directly confront online hate. It is therefore important to encourage silent cyber-bystanders to show their support for those who do discuss with the haters. Counter-speech can be supported by liking the comments and posts of counter-speakers (which increases the comments' visibility), but even better, by writing few words of support. The power balance in discussions between haters and counter-speakers often involves a group of haters against a single opponent. The likes the opponent gets might not be very visible signs of support for other silent bystanders to notice.

Apart from giving likes, it could make even more sense to express support by posting short comments: "X is right", "I agree with X", "I am on your side, X", "X makes a good point", "X, your words convince me". Such short responses, while unlikely to attract insults from haters, more focused on responding to the main thread, might provide a social proof for other silent readers (cyber-bystanders) and encourage them to gain more confidence in showing even more support. Such support can also reduce the confidence in the haters, who might otherwise feel strong having to argue with only a single user.

Another way to show support to pro-democratic views is to include democratic message as a part of one's profile photo, for instance by adding a supportive photo frame (available on Facebook) or using a supportive hashtag. While Facebook's filter bubbles filter out much of activity of our contacts, especially if their beliefs are different from ours, the change of a profile photo might be able to reach outside the filter bubble, because it is often displayed in the friends' newsfeed.

POST AND PRODUCE COUNTER-SPEECH MEMES: Bartlett and Krasodomski-Jones (2015) determined, that the most successful counter-speech type of content was visual (a photo), the most successful counter-speech tone was satirical and the most successful type of counter-speech message was constructive (rather than insulting).

JOIN COUNTER-SPEECH AND ANTI-DISCRIMINATION GROUPS AND PAGES: They might provide useful resources to boost your confidence in argumentation (for instance, Everyday Feminism provides daily powerful articles on gender, sexual orientation, race and intersectionality). They can also be useful for gaining immediate support whenever social proof might be needed. One example of such a Facebook group is #ichbinhier. Would you like to follow Facebook pages with useful content, but are unsure where to find them? Ask your friends for recommendations of anti-discrimination, anti-racism, LGBTQ-support or feminist pages they subscribe and find useful. Once there, ask again among their members and you will likely receive even more recommendations.

WHAT IS OBVIOUS FOR YOU MIGHT NOT BE OBVIOUS FOR YOUR FRIENDS AND FOR MEMBERS OF GROUPS YOU BELONG TO: Remember that familiarity with a given subject appears to make us immune to negative effects of online hate? Probably by reducing fear from the unknown. Boost familiarity among your social media contacts by sharing helpful articles that had helped you grow: for instance, on how to be less racist, how to better support transgender community or how to better oppose sexism? Or maybe you already know a lot about various forms of discrimination and you assume others know them too? Share the links with others, adding an interesting excerpt from the text. Have you read an article in a foreign language? Link the article and provide a short summary in your native language. Some among your friends might get interested in the subject, too.

READ USEFUL EXPERIENCE-BASED RESOURCES AND USE RECOMMENDED INTERVENTION METHODS: Among useful, experience-based publications containing valuable suggestions are: "„Geh

sterben!“ Umgang mit Hate Speech und Kommentaren im Internet” by Amadeu Antonio Stiftung (Baldauf, Banaszczuk, Koreng, Schramm, & Stefanowitsch, 2015b), „Considerations for successful counterspeech“ by Benesch, Ruths, Dillon, Saleem and Wright (2016), “Digitale Antidiskriminierungsarbeit” by Rafael, Dinar and Heyken (2017) and “Hass und Hetze im Internet – Analyse und Intervention”, also by Dinar and Rafael. There is also an informative and comprehensive book, “Hass im Netz: Was wir gegen Hetze, Mobbing und Lügen tun können” by Ingrid Brodnig (2016).

SUMMARY

This article was an attempt to concisely summarize some of the most recent scientific literature on the psychological dynamics of hate-speech and counter-speech on the internet. While many internet users spontaneously engage in various types of counter-speech activity, mostly by commenting hateful posts, but also by creating memes, some even joining online anti-hate communities; while various anti-hate projects are being carried out, making counter-speech tools, videos, articles and other resources available to the internet users: only few helpful scientific evaluations of counter speech strategies appear to have been published to date. Most available suggestions and resources are valuable experience-based observations by activists involved in the said anti-hate-speech projects, rather than by scientists.

Research has demonstrated, that online hate has much more viral potential than joy or sympathy, and another body of research has shown the possible mechanism underlying viral spread of hate and prejudice: fear and psychological need for safety. Other researchers have shown that familiarity with what we might otherwise fear of immunizes us from perceiving a given issue as threatening, even when we are exposed to social proof guiding us otherwise. Fear makes us motivated to share hateful content and to warn others. Hate and ridicule are our weapons against what we are afraid of and what we perceive as unfamiliar. Our online invisibility and the “facelessness” of our online adversaries combined with asynchronicity of internet discussions and perceived lack of consequences of our actions make us prone to online disinhibition effect. A reminder of humanity of people we tend to mistreat might have a sobering effect on our online behavior, especially when coming from either an in-group member with a high status or from a privileged white man.Satirical and visual form of counter-speech is the most popular one, yet it is still unlikely to reach the audience outside of our filter bubble of like-minded individuals. Filter bubbles and echo chambers deceive us, creating an impression that most people share our values, beliefs and fears.
Online hate and prejudice threaten members of targeted groups. They transform into real-life violence, endangering the physical safety and psychological well-being of the victims. Fear for one’s own safety is one of the factors that silences the victims and suppresses active resistance. As the result, only some people are privileged enough to be able to challenge online hate. On the other hand, the privileged status of being free of oppression and from its consequences reduces motivation to actively counter hate.

Countering hate is responsive in nature and involves effort, as opposed to expressing hate, which is impulsive, spontaneous and effortless. All these factors contribute to marked asymmetry that helps hate speech spread and renders counter speech relatively invisible.

The unbelievable success of the #metoo movement proves, however, that crowd-initiated social media counter-action driven by anger can go viral and reach far beyond the original filter bubble, leading to actual social change on an inter-continental scale.

REFERENCES

Anderson, A. A., Brossard, D., Scheufele, D. A., Xenos, M. A., & Ladwig, P. (2014). The "Nasty Effect:" Online Incivility and Risk Perceptions of Emerging Technologies. *Journal of Computer-Mediated Communication*, *19*(3), 373–387. https://doi.org/10.1111/jcc4.12009

Aronson, E., Wilson, T. D., & Akert, R. M. (2010). *Social psychology*. Upper Saddle River, NJ : Prentice Hall. Retrieved http://archive.org/details/Social_Psychology_7th_edition_by_Elliot_Aronson_Timothy_D._Wilson_R_M._Akert

Asch, S. E. (1956). Studies of independence and conformity: A minority of one against a unanimous majority. *Psychological Monographs*, *70*(9).

Asch, S. E., & Guetzkow, H. (1951). Effects of group pressure upon the modification and distortion of judgement. In *Groups, leadership and men* (S. 177–190). Pittsburgh, PA: Carnegie Press.

Baldauf, J., Banaszczuk, Y., Koreng, A., Schramm, J., & Stefanowitsch, A. (2015a). Die direkte Bedrohung durch Hate Speech darf nicht unterschätzt werden! Interview mit Dorothee Scholz, Diplompsychologin. In J. Schramm & A. Lanzke (Hrsg.), *„Geh sterben!" Umgang mit Hate Speech und Kommentaren im Internet* (S. 25–29). Berlin: Amadeu Antonio Stiftung. Retrieved http://www.amadeu-antonio-stiftung.de/w/files/pdfs/hatespeech.pdf

Baldauf, J., Banaszczuk, Y., Koreng, A., Schramm, J., & Stefanowitsch, A. (2015b). *„Geh sterben!" Umgang mit Hate Speech und Kommentaren im Internet*. (J. Schramm & A. Lanzke, Hrsg.). Berlin: Amadeu Antonio Stiftung. Retrieved http://www.amadeu-antonio-stiftung.de/w/files/pdfs/hatespeech.pdf

Baron, R. S., Vandello, J. A., & Brunsman, B. (1996). The forgotten variable in conformity research: Impact of task importance on social influence. *Journal of Personality and Social Psychology*, *71*(5), 915–927. https://doi.org/10.1037/0022-3514.71.5.915

Bartlett, J., & Krasodomski-Jones, A. (2015). *Counter-speech: Examining content that challenges extremism online* (S. 21). Demos. Retrieved https://www.demos.co.uk/wp-content/uploads/2015/10/Counter-speech.pdf

Benesch, S., Ruths, D., Dillon, K. P., & Saleem, H. M. (2016). Considerations for Successful Counterspeech, 1–9.

Brodnig, I. (2016). *Hass im Netz: Was wir gegen Hetze, Mobbing und Lügen tun können*. Wien: Brandstätter Verlag.

Chronik flüchtlingsfeindlicher Vorfälle. (o. J.). Retrieved 12. Juni 2018, von https://www.mut-gegen-rechte-gewalt.de/service/chronik-vorfaelle

Clay, R. A. (2017). Islamophobia: Psychologists are studying the impact of anti-Muslim sentiment and exploring ways to prevent it. *Monitor on Psychology*, *48*(4), 34. Retrieved http://www.apa.org/monitor/2017/04/islamophobia.aspx

Coustick-Deal, R. (2017, Februar 6). What's wrong with counter speech? Retrieved 14. Mai 2018, von https://medium.com/@ruthcoustickdeal/https-medium-com-whats-wrong-with-counter-speech-f5e972b13e5e

Europarat Ministerkomitee. Empfehlung R (97) 20

des Ministerkomitees an die Mitgliedstaaten über die „Hassrede“ (1997).

Fan, R., Zhao, J., Chen, Y., & Xu, K. (2014). Anger Is More Influential than Joy: Sentiment Correlation in Weibo. *PLoS ONE*, *9*(10), e110184. https://doi.org/10.1371/journal.pone.0110184

forsa. (2017). *Hate Speech*. Landesanstalt für Medien Nordrhein-Westfalen. Retrieved https://www.lfm-nrw.de/fileadmin/user_upload/lfm-nrw/Service/Pressemitteilungen/Dokumente/2017/Ergebnisbericht_Hate-Speech_forsa-Mai-2017.pdf

Fyfe, S. (2017). Tracking Hate Speech Acts as Incitement to Genocide in International Criminal Law. *Leiden Journal of International Law*, *30*(2), 523–548. https://doi.org/10.1017/S0922156516000753

Gelber, K., & McNamara, L. (2016). Evidencing the harms of hate speech. *Social Identities*, *22*(3), 324–341. https://doi.org/10.1080/13504630.2015.1128810

Jost, J. T., Glaser, J., Kruglanski, A. W., & Sulloway, F. J. (2003). Political Conservatism as Motivated Social Cognition. *PSYCHOLOGICAL BULLETIN*, *129*(3), 339–375. Retrieved https://search.ebscohost.com/login.aspx?direct=true&db=edsbl&AN=RN131422234&lang=pl&site=eds-live

Maas, H. (2015). Geleitwort. In Amadeu Antonio Stiftung (Hrsg.), *„Geh sterben!“ Umgang mit Hate Speech und Kommentaren im Internet* (S. 6). Berlin: Amadeu Antonio Stiftung. Retrieved http://www.amadeu-antonio-stiftung.de/w/files/pdfs/hatespeech.pdf

Maravilla, C. S. (2008). Hate Speech as a War Crime: Public and Direct Incitement to Genocide in International Law. *Tulane Journal of International & Comparative Law*, *17*(1), 113–144. Retrieved https://search.ebscohost.com/login.aspx?direct=true&db=lgs&AN=502065087&lang=pl&site=eds-live

Meyer, I. H. (1995). Minority stress and mental health in gay men. *Journal of Health & Social Behavior*, *36*(1), 38–56.

Meyer, I. H. (2003). Prejudice, social stress, and mental health in lesbian, gay, and bisexual populations: Conceptual issues and research evidence. *Psychological Bulletin*, *129*(5), 674–697. https://doi.org/10.1037/0033-2909.129.5.674

Mullen, B., & Smyth, J. M. (2004). Immigrant suicide rates as a function of ethnophaulisms: hate speech predicts death. *Psychosomatic Medicine*, *66*(3), 343–348.

Müller, K., & Schwarz, C. (2018). *Fanning the Flames of Hate: Social Media and Hate Crime* (SSRN Scholarly Paper No. ID 3082972). Rochester, NY: Social Science Research Network. Retrieved https://papers.ssrn.com/abstract=3082972

Munger, K. (2017). Tweetment Effects on the Tweeted: Experimentally Reducing Racist Harassment. *Political Behavior*, *39*(3), 629–649. https://doi.org/10.1007/s11109-016-9373-5

Nail, P. R., & McGregor, I. (2009). Conservative Shift among Liberals and Conservatives Following 9/11/01. *Social Justice Research*, *22*(2–3), 231–240. https://doi.org/10.1007/s11211-009-0098-z

Nail, P. R., McGregor, I., Drinkwater, A. E., Steele, G. M., & Thompson, A. W. (2009). Threat

causes liberals to think like conservatives. *Journal of Experimental Social Psychology*, *45*, 901–907. Retrieved https://www.academia.edu/23822628/Threat_causes_liberals_to_think_like_conservatives

Napier, J. L., Huang, J., Vonasch, A. J., & Bargh, J. A. (2017). Superheroes for change: Physical safety promotes socially (but not economically) progressive attitudes among conservatives. *European Journal of Social Psychology*, *48*(2), 187–195. https://doi.org/10.1002/ejsp.2315

Rafael, S., Dinar, C., & Heyken, C. (2017). Digitale Antidiskriminierungsarbeit. *Wissen schafft Demokratie, Vol 1, Iss 2, Pp 160-171 (2017)*, (2), 160. https://doi.org/10.19222/201702/15

Sherif, M. (1935). A study of some social factors in perception. *Archives of Psychology (Columbia University)*, *187*, 60–60.

Soral, W., Bilewicz, M., & Winiewski, M. (2018). Exposure to hate speech increases prejudice through desensitization. *Aggressive Behavior*, *44*(2), 136–146. https://doi.org/10.1002/ab.21737

Sue, D. W. (2010). *Microaggressions in Everyday Life: Race, Gender, and Sexual Orientation*. John Wiley & Sons.

Terry, C., & Cain, J. (2016). The Emerging Issue of Digital Empathy. *American Journal Of Pharmaceutical Education*, *80*(4), 58–58. https://doi.org/10.5688/ajpe80458

United Nations Regional Information Centre for Western Europe. (2016). Intolerance and xenophobia on the rise in Europe. Retrieved 12. Juni 2018, https://www.unric.org/en/latest-un-buzz/30377-intolerance-and-xenophobia-on-the-rise-in-europe