Received: 28 September 2020 | Revised: 16 December 2020 | Accepted: 28 December 2020

DOI: 10.1002/ab.21948

RESEARCH ARTICLE

AGGRESSIVE BEHAVIOR | WILEY

Artificial intelligence against hate: Intervention reducing verbal aggression in the social network environment

Michał Bilewicz[1] | Patrycja Tempska[2] | Gniewosz Leliwa[2] | Maria Dowgiałło[2] | Michalina Tańska[3] | Rafał Urbaniak[4] | Michał Wroczyński[2]

[1]Faculty of Psychology, University of Warsaw, Warsaw, Poland

[2]Samurai Labs, Gdynia, Poland

[3]Institute of Psychology, Jagiellonian University, Kraków, Poland

[4]Faculty of Social Sciences, University of Gdańsk, Gdańsk, Poland

Correspondence
Michał Bilewicz, Faculty of Psychology, University of Warsaw, Stawki 5/7, 00-183 Warsaw, Poland.
Email: bilewicz@psych.uw.edu.pl

Funding information
Narodowe Centrum Nauki, Grant/Award Number: 2017/26/ E/HS6/ 00129

Abstract

This article presents a quasi-experimental intervention study designed to reduce the level of verbal aggression on a social networking service (Reddit). The interventions were based on three psychological mechanisms: induction of a descriptive norm, induction of a prescriptive norm, and empathy induction. Each intervention was generated using a communicating bot. Participants exposed to these interventions were compared with a control group that received no intervention. The bot-generated normative communications (both the ones priming descriptive and the ones priming prescriptive norms), as well as the empathizing intervention, reduced the proportion of verbal aggression posted by Reddit accounts. All three interventions proved effective in reducing verbal violence when compared with the control condition.

KEYWORDS

artificial intelligence, empathy, hate speech, social media, verbal aggression

1 | INTRODUCTION

Online violence including hate speech, personal attacks, and bullying is a growing societal concern. Online violence often accompanies, precedes or escalates into offline violence, both physical and psychological (Anderson et al., 2017; Holtz & Appel, 2011; Ybarra et al., 2008). In 2020, several global social networking services (including Facebook, Twitter, Reddit, and Twitch) have introduced more restrictive hate speech policies and increased suspensions and bans of accounts promoting hateful language and violence. Some of the companies have directly addressed aggressive acts in their codes on conduct (Twitter, for instance, states: "You may not promote violence against or directly attack or threaten other people on the basis of race, ethnicity, national origin, caste, sexual orientation, gender, gender identity, religious affiliation, age, disability, or serious disease. We also do not allow accounts whose primary purpose is inciting harm towards others on the basis of these categories," whereas Facebook says: "We do not allow hate speech on Facebook because it creates an environment of intimidation and exclusion and in some cases may promote real-world violence"), in some cases banning the accounts linked to the President of the United States of America (Timberg & Dwoskin, 2020).

Although in the case of hate speech, defined as an offensive language against disadvantaged social groups (Jacobs & Potter, 1998), such content monitoring and punitive actions might sometimes prove effective, they seem insufficient in eradicating more general verbal violence from social networking sites. Prevalence of hate speech and other forms of verbal violence creates a pressing need for identifying more effective strategies and tools to mitigate such behaviors. Moderation practices generally begin with reporting of obvious abuses by other community members and culminate in a decision regarding disciplinary action by a human moderator. These strategies have proven insufficient to the largest and most well-resourced online platforms such as Facebook and Twitter, and are not available to smaller, newer, and less-resourced online sites due to factors such as high costs or complexity.

There are several methods with the potential to reduce verbal aggression on social network platforms and other content and community sites. These methods include promoting counter-narratives (Poole et al., in press), psychoeducation (Mishna et al., 2011), automated identification of cyberhate (Blaya, 2019), and bystander intervention

(Kazerooni et al., 2018). When analyzing the psychological mechanisms underlying such interventions, two paths that might lead to effective de-radicalization can be identified: the elicitation of social norms and induction of empathy.

2 | THE ROLE OF SOCIAL NORMS IN REDUCING ONLINE VERBAL AGGRESSION

The introduction and reinforcement of social norms is a common way of reducing aggression in many contexts, including sexual assault (Gidycz et al., 2011) and school bullying (Perkins et al., 2011). It can also be applied to the context of online hate speech: The development of new social norms (e.g., promoting the common use of counter-speech rather than hate speech) can effectively reduce verbal violence on the Internet (Mathew et al., 2019). Social norms can be also communicated in the form of legal regulations in a given country. Countries where anti-hate speech laws have been introduced (e.g., Germany) have witnessed a greater decrease in hate speech than countries where no such regulations exist (e.g., the United States; Hawdon et al., 2017). This corresponds to the distinction between descriptive and prescriptive norms. Descriptive norms (Cialdini et al., 1990; Prentice, 2007) describe what constitutes a typical or common behavior in a specific environment (often within a group to which one belongs). Prescriptive norms (Brauer & Chaurand, 2010), on the contrary, are judgments about the desirability of a certain behavior. The first type of norms is based on observations of the reactions and behaviors of other people. The second comes from the knowledge of laws, regulations, and commandments.

It is plausible that interventions inducing these two kinds of norms could prove effective in reducing people's tendency to use derogatory language on social networking sites. People, when being informed by other users of Facebook, Twitter, or Reddit that their behavior is hurtful, might gain the sense of a descriptive norm countering verbal aggression. On the contrary, being reminded about abstract norms and standards could make one realize that their behavior violates the prescriptive norm of not using hurtful language. Research on normative acceptance of online hate speech has found that frequent exposure to hate speech creates a sense of normativity of such language, and this, in turn, worsens intergroup attitudes (Bilewicz & Soral, 2020; Soral et al., 2020). Therefore, changing the norm could improve one's attitude and reduce their willingness to use derogatory language. Recent studies have found that denouncing hate speech and warning of its online and offline consequences (Mathew et al., 2019) could reduce the amount of verbal aggression on the Internet. This corresponds with a prescriptive norm violation and could be applied in the form of direct intervention.

3 | THE ROLE OF EMPATHY IN REDUCING ONLINE VERBAL AGGRESSION

The second psychological mechanism that has been previously found to reduce aggression and violence is empathy. Both the cognitive and affective components of empathy have been reported to reduce the amount of bullying behavior a person engages in, and to increase the probability of a bystander helping the victim of an aggressive act (van Noorden et al., 2015). When it comes to online aggression (cyberbullying), induction of cognitive empathy has been found to reduce the frequency of forwarding hateful contents (Barlińska et al., 2013, 2015) and to increase the tendency to report such acts (Barlińska et al., 2018). Although empathic reactions are rather rare (Bruneau et al., 2017; Levy et al., 2016), it is commonly believed that kindness and empathy tend to reduce the tendency to act in a hateful and derogatory way (Zaki, 2014).

When thinking of an intervention targeting users of aggressive language on social media, one could propose employing empathy and kindness as an interesting (though paradoxical) treatment. Eliciting empathy in an online hater could interfere with the script of aggressive behavior of such an individual. When communicated in a kind and direct way, empathy might potentially reduce one's willingness to harm other users of social networking sites.

At the same time, the empathy-driven strategy of de-escalation and aggression reduction has faced substantial criticism (Bloom, 2017a, 2017b). Already Adam Smith (2006) suggested that empathy-driven reactions often paradoxically lead to aggression toward the perpetrator, in the name of the victim ("When we see one man oppressed or injured by another, the sympathy which we feel with the distress of the sufferer seems to serve only to animate our fellow-feeling with his resentment against the offender", p. 69). More recently, Paul Bloom (2017a) suggested that there is a contradiction between empathy and morality, as empathy limits people's capacity to make correct moral judgements. He proposed that compassion, understood as valuing other people and caring about their welfare without necessarily feeling their pain, might be more effective in aggression reduction than eliciting empathy.

To assess the effects of normative and empathy-based interventions on reducing verbal aggression in social media, we performed a systematic analysis of behavior among users of social networking site who engaged in online personal attacks. After performing such attack, these users were confronted by a programmed artificial account. The account was systematically interacting with them using one of three strategies based on the psychological mechanisms described above.

4 | THE CURRENT STUDY

In this quasi-experimental intervention study, we aimed at examining which of the three psychological methods (i.e., descriptive norms vs. prescriptive norms vs. empathy) could be effective in reducing online verbal aggression. We employed an innovative approach that involved artificial intelligence (a bot using artificially generated natural language) approaching other social media users in a systematic way.

The study estimated the impact of various automatic counter-speech interventions on user's engagement in personal attacks (acts of verbal aggression) among two selected Reddit communities. For around 6 months, an online bot monitored these communities and performed the interventions whenever a personal attack was detected. A control group was composed of similar users of other

subreddits selected at approximately the same time—but not targeted with any intervention.

5 | METHODS

5.1 | Participants

The selection of participants (Reddit accounts) for intervention was based on three criteria. Participants were users of two communities (r/MensRights and r/TooAfraidToAsk) who produced (1) at least 30 comments in the period before the intervention (at least one comment per 2 days on average), (2) at least one violent comment in the period before the intervention, and (3) the same or higher number of comments before and after the intervention. Overall, 454 users fulfilling the criteria were selected for the intervention. In addition, a control group of 437 users was recruited from other subreddits based on the same criteria (a randomized timestamp was generated to indicate the moment used for pretest–posttest comparisons).

5.2 | Detection of verbal aggression

The detection of verbal aggression was performed automatically using Samurai Labs' proprietary technology. The technology combines statistical learning from data with symbolic reasoning based on deep linguistic analysis and experts' knowledge (e.g., linguists, psychologists, cyberbullying researchers). It employs a bottom-up approach where each problem is divided into a set of corresponding phenomena (e.g., speech acts) and detected independently using precise contextual models, which enables automatic interventions.

In practice, it means that a whole variety of structures could be detected while conducting the interventions without the need to construct an a priori fixed list of dictionary words. Furthermore, the model recognized complex linguistic phenomena such as indirect speech, rhetorical figures or counter-factual expressions, greatly reducing the number of false alarms. The model consisted of 43 contextual submodels designed to detect various phenomena related to personal attacks. It was validated on 477,851 Reddit comments from r/MensRights and r/TooAfraidToAsk collected between June 01, 2019 and August 31, 2019, and achieved over 93% precision in the detection of personal attacks on this data set.

5.3 | Bot account (intervention agent)

The bot conducting the interventions was disguised as a regular Reddit user. As in the study by Munger (2017), we had to ensure that participants would not become aware that they were interacting with an automated bot to maintain the natural character of online conversations. To increase the bot's credibility, we purchased an account with a history of posts and comments, as well as 500+ karma points (karma is a score that reflects how much a user has contributed to the Reddit community—karma points are received for positive votes given to user's submissions and comments).

The bot's username was generated so that it would indicate a male account. It was crucial, because one of the targeted subreddits, r/MensRights, is frequented mostly by men, who gather there to discuss their legal rights, their relationship to society, and social roles, while often referring to alleged discrimination of men by women.

To contribute further to the bot's credibility, we made sure that its online activity would resemble an activity of a regular Reddit user. Once a user would receive an intervention comment, a notification appeared on their profile, just like in the case of any response posted to a comment on Reddit. Any of the targeted individuals could click on the bot's profile and see its activity history (with the few most recent comments displayed on one page, see Figure 1). Nonetheless, an account that generates only peaceful comments can raise suspicion even if the messages are very diverse. Thus, while the experiment lasted, one of the experimenters was posting a couple of comments daily from the bot's account to make it look like an actual human user. These comments were made on subreddits other than the two selected for the study and included discussions on octopuses' habits, help with other users' homework, recommending jazz albums, and so forth. Consequently, users would see a mix of comments on various topics in the bot's activity, which would likely dispel any suspicions regarding our bot being artificial.

Another aspect that could lead to skepticism when it comes to the bot being a human is response timing. Although each personal attack in the study group was treated with an intervention in almost real-time, responses were delayed randomly between 3 and 6 min.

5.4 | Intervention

The interventions performed by the bot had the form of a direct reply to the user attacking others on the forum, addressing the hostile behavior and motivating change. The interventions would appear in the comment thread under the targeted comment. Each intervention consisted of two distinct parts. The first part was a kind introduction built from randomly selected modules. The structure of such an introduction is presented below (Figure 2) with a sample of the module contents.

The second part of each intervention had a form of a message utilizing one of the psychological methods identified as potentially effective in reducing online aggression: induction of a descriptive norm ($n = 208$), induction of a prescriptive norm ($n = 141$), and empathy induction ($n = 105$). Also, this message was built of randomly generated sentences consisting of modules.

5.4.1 | Disapproval message

This message was meant to induce a descriptive norm—it was creating an impression that verbal aggression is not accepted by fellow community members. It consisted of two parts. The first was an acknowledgment of negative emotions or other factors that might have led the individual to engage in the personal attack, and an expression of

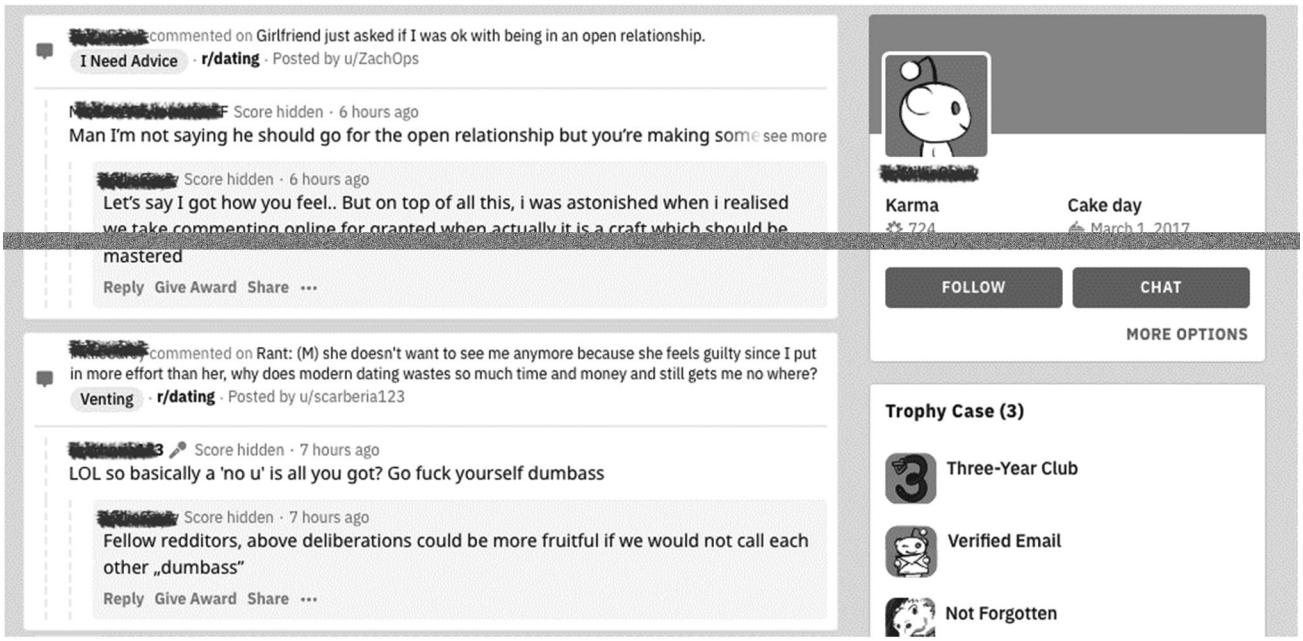

FIGURE 1 The profile of the artificial bot used in the interventions (with sample comments/messages)

understanding. The function of this introductory part was to prevent the individual from potentially rejecting the second (negative) part of the message: a direct disapproval of their aggressive behavior. The disapproval was expressed in a few different ways. For instance, it could have been a comment regarding the individual's use of verbal aggression, a remark underlining the importance of respect, or a negative opinion on the general use of hateful comments. Sometimes the disapproval involved presenting a short imaginative story or referring to an authority figure important to Reddit users. Examples of the entire disapproval message: "Hi:) I somewhat commiserate with what you're feelin', but let's try to express our points without hurtful language," "i hear you, bro I guess I am in tune with the feeling, but using kinder words might be a way to go," "Actually, I kind of get how you feel. but please don't offend anybody."

5.4.2 | Abstract norms message

This message was meant to induce a prescriptive norm. It referred to abstract values and virtues by reminding users about the desired forms of communication. It had either a utilitarian form (explaining why it is useful to adhere to certain ways of communication), or an empowering form (suggesting that certain standard of communication can be upheld if all users are dedicated to following them). Examples of the abstract norms message: "Good day sir, have you ever thought about how this discussion could be more enjoyable for all if we would treat each other with respect?," "ability to express ourselves but with respect to others is a wonderful sign of character and takes lots of courage," "capability to imagine a different point of view is a wonderful quality and requires hard work."

5.4.3 | Empathy message

The empathizing message was focused on the target of verbal aggression. It aimed to bring to attention the emotions and feelings that the attacked individual could be experiencing. Following on the argument of Bloom (2017b), this strategy sought to also utilize the effects of more general compassion by stressing the shared humanity of the victim and the perpetrator of the verbal attack. Examples of the empathy message: "Some behaviors might be hard to get for some people but let's keep in mind there are people of flesh and

part 1
Greeting patch = ["Hmm...", "alright," , "All right. ", "ah.. ", "ok," , "Ok.", "Gotcha mate ","yeah,", "Well.. ","Right,", "oh, I'm sorry.", "fine,", "Actually, ", "Hey mate! ", "My friend ", "hey bro ", "Hola compadre!", "hey there, partner" , "Howdy!", "howdy ho!", "Hi :) ", "Good Day, sir!", "i hear you, bro ", "dear friend ", "Fellow redditor,"]
a = ["I kind of", " I suppose I ", "I think I ", "Let's say I ", "I somewhat ", "i guess I "]
b = ["understand ", "got ", "get" , "can sympathize with ", "can empathize with ", "can relate to ", "commiserate with ", "am in tune with"]
c = ["you. ", "the feeling. ", " this feeling..", "your emotions.", "how you feel..", "what you're feelin' "]

FIGURE 2 The structure of the introductory part of an intervention. Each targeted account received a comment starting with a sentence built from the greeting patch and the components a, b, and c

blood on the other side of the screen," "certain behaviors may be hard for us to understand still try to remember to be gentle cause you never know what's going on in someone's life," "What other people are saying or doing can be hard to get for some people but let's keep in mind to be gentle, people are fragile."

5.5 | Key measure

5.5.1 | Verbal aggression

To measure changes in verbal aggression, a pre- and a post-intervention index of verbal aggression was created. They were calculated as a proportion of comments that had a form of a personal attack to all the comments produced by a given account 60 days before the intervention or the timestamp and 60 days after the intervention or the timestamp, respectively.

6 | RESULTS

A two-way mixed-design analysis of variance was conducted with the type of intervention as a between-subject factor (four levels: disapproval vs. norms-inducing vs. empathizing vs. no intervention) and the measurement time as a within-subject factor (two levels: before vs. after the intervention).

The analyses revealed a significant effect of the measurement time, $F(1, 887) = 23.47$, $p < .001$, $\eta^2_p = 0.03$. Overall, participants displayed a lower proportion of verbal aggression after the intervention, $M = 0.018$, $SD = 0.026$, than before the intervention, $M = 0.021$, $SD = 0.025$. There was no significant effect of the intervention type, $F(3, 887) = 0.001$, $p = .45$, $\eta^2_p = 0.003$. There was, however, a significant interaction between the intervention type and the measurement time, $F(3, 887) = 8.11$, $p < .001$, $\eta^2_p = 0.03$ (see Figure 3).

To examine this interaction, we performed a planned contrast analysis. The analysis revealed that the biggest difference between the measurement times was observed in the case of the disapproval ($p < .001$) and norm-inducing ($p < .001$) interventions. In the case of the disapproval intervention, the proportion of verbal aggression was reduced from $M = 0.023$, $SD = 0.027$ to $M = 0.017$, $SD = 0.027$, whereas in the case of the norm-inducing intervention, it was reduced from $M = 0.021$, $SD = 0.022$ to $M = 0.014$, $SD = 0.019$. In the case of the empathizing intervention, a significant but smaller decrease in verbal aggression was observed ($p = .032$), that is, the empathizing intervention reduced the proportion of verbal aggression from $M = 0.025$, $SD = 0.031$ to $M = 0.020$, $SD = 0.036$. In the control condition (no intervention) the proportion of verbal aggression remained at the same level ($M = 0.019$, $SD = 0.023$ before and $M = 0.020$, $SD = 0.025$ after the timestamp; $p = .265$). The change of verbal aggression in the control group was significantly lower than the change observed in the disapproval group, $F(1, 643) = 13.939$, $p < .001$, $\eta^2_p = 0.021$; lower than the change observed in the norm-inducing intervention, $F(1, 576) = 13.548$, $p < .001$, $\eta^2_p = 0.023$ and lower than the change observed in the empathizing intervention, $F(1, 540) = 5.233$, $p = .023$, $\eta^2_p = 0.010$.

To test the differences between the alternative types of intervention, we contrasted the intervention types against each other. The effectiveness of interventions did not differ significantly between the intervention types, $F(2, 451) = 0.388$, $p = .68$, $\eta^2_p = 0.002$. The norm-inducing intervention was equally effective as disapproval intervention, $F(1, 347) = 0.099$, $p = .75$, $\eta^2_p < 0.001$, and as empathizing intervention, $F(1, 311) = 0.374$, $p = .54$, $\eta^2_p = 0.001$. The effectiveness of empathizing intervention did not differ significantly from disapproval intervention, $F(1, 244) = 0.904$, $p = .34$, $\eta^2_p = 0.004$.

7 | DISCUSSION

The goal of the present study was to examine a set of interventions designed to reduce the level of verbal aggression on a social networking service. The interventions were based on three psychological mechanisms: induction of a descriptive norm (disapproval message), induction of a prescriptive norm (abstract norm-inducing message), and empathy (empathizing message). Each intervention was generated using a communicating bot. Participants exposed to these interventions were compared with a control group that

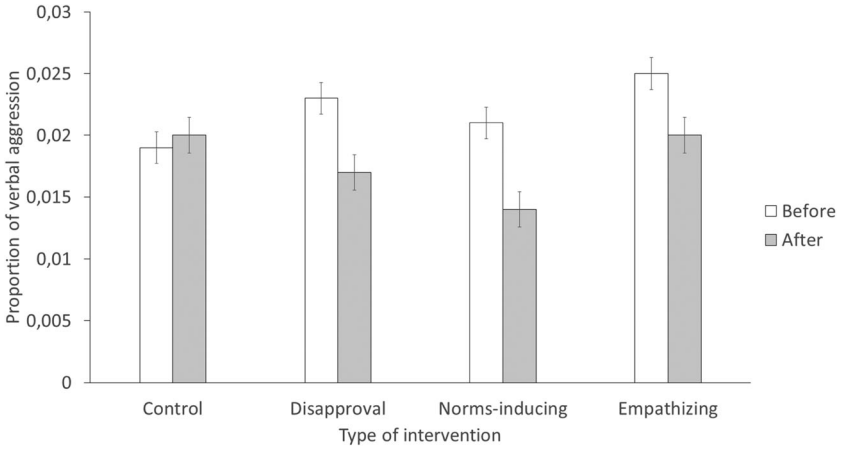

FIGURE 3 Proportion of verbal aggression to all comments before and after disapproval, norms-inducing and empathizing interventions, as well as in the control condition (without intervention)

received no intervention. The results of this study showed that the bot-generated normative communicates (both the ones priming descriptive and the ones priming prescriptive norms), as well as empathizing interventions significantly reduced the proportion of verbal aggression posted by Reddit accounts. All three interventions proved effective in reducing verbal aggression when compared with the control condition.[1]

These results add to the evidence for the role of social norms in eliciting behavior change (Brauer & Chaurand, 2010; Prentice, 2007), while also pointing to a new context for norm-based interventions in social network environment. Although the effectiveness of empathy induction did not significantly differ from the normative interventions, the statistical effects observed in the case of this intervention were not particularly strong. The limited effects of the empathizing intervention could be attributed to the fact that empathizing messages are often focused on the target of a specific behavior rather than on promoting general moral standards (Bloom, 2017a), which can limit the capacity of such interventions to modify aggressive behaviors toward other targets. Our study also shows that aggression-targeting interventions can be performed almost entirely in the absence of human researchers. Our interventions were generated by an automated device—a bot that created statements from pre-programmed modules. The use of such a bot accompanied by automated recognition of verbal aggression provides an interesting alternative to costly moderation procedures employed by social networking companies.

The study had some obvious limitations. The first limitation is related to the fact that the study had a form of a quasi-experiment. The accounts selected for the interventions were not fully random (the bot appeared only on specific subreddits and its activity was observed by other users—therefore, the interventions had to be nested within subreddits). This resulted in the initial verbal aggression level being different between the conditions. In the future, similar interventions should be tested in a fully randomized experimental design. Another limitation is that there was a possibility of users from the control group coming across the bot account (e.g., if they observed other accounts interacting with the bot). This possibility was rather unlikely, as accounts in the control condition were selected from other Reddit groups than the ones targeted by the bot.

This quasi-experiment constituted a first attempt to employ automatically generated contents in a behavioral social media intervention using psychologically developed procedures. Our results suggest that artificial intelligence could be potentially used as a means to limit the proliferation of hate speech and verbal aggression online, thus addressing one of the most pressing problems of contemporary media.

ACKNOWLEDGMENTS

We wish to thank our academic partners, Dr. Michal Ptaszynski from the Kitami Institute of Technology and Dr. Marek Kisiel-Dorohinicki, Dr. Aleksander Smywinski-Pohl, Krzysztof Wrobel, and Mateusz Piech from AGH University of Science and Technology for their substantive support, expertise, and insights that greatly assisted our work. We would also like to thank the whole Samurai Labs team for cheering us on in this study, providing valuable feedback and putting their effort into making our bot more human. We would like to give a special thanks to Maciej Brochocki, our VP of Engineering, who developed and integrated the bot and ensured that it worked flawlessly during the whole experiment and to Marta Beneda for her comments on the first version of this manuscript. This study was funded by the National Science Centre Sonata Bis (Grant 2017/26/E/HS6/00129) to the first author.

DATA AVAILABILITY STATEMENT

The data that support the findings of this study are available on request from the corresponding author.

ENDNOTE

[1]This result is supported by the comparison of decrease in verbal aggression on subreddit, on which intervention took place (r/MensRights) with the dynamics of verbal aggression on other subreddits of similar content that were not the target of our intervention. Supplementary analyses can be found at https://osf.io/ebw3j/

ORCID

Michał Bilewicz https://orcid.org/0000-0001-5027-1691

REFERENCES

Anderson, C. A., Bushman, B. J., Bartholow, B. D., Cantor, J., Christakis, D., Coyne, S. M., Donnerstein, E., Brockmyer, J. F., Gentile, D. A., Green, C. S., Huesmann, R., Hummer, T., Krahé, B., Strasburger, V. C., Warburton, W., Wilson, B. J., & Ybarra, M. (2017). Screen violence and youth behavior. *Pediatrics*, *140*(S2), S142–S147.

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.

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.

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.

Bilewicz, M., & Soral, W. (2020). Hate speech epidemic. The dynamic effects of derogatory language on intergroup relations and political radicalization. *Advances in Political Psychology*, *41*(S1), 3–33.

Blaya, C. (2019). Cyberhate: A review and content analysis of intervention strategies. *Aggression and Violent Behavior*, *45*, 163–172.

Bloom, P. (2017a). *Against empathy: The case for rational compassion*. Random House.

Bloom, P. (2017b). Empathy and its discontents. *Trends in Cognitive Sciences*, *21*, 24–31.

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.

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.

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.

Gidycz, C. A., Orchowski, L. M., & Berkowitz, A. D. (2011). Preventing sexual aggression among college men: An evaluation of a social norms and bystander intervention program. *Violence Against Women*, *17*, 720–742.

Hawdon, J., Oksanen, A., & Räsänen, P. (2017). Exposure to online hate in four nations: A cross-national consideration. *Deviant Behavior*, *38*(3), 254–266. https://doi.org/10.1080/01639625.2016.1196985

Holtz, P., & Appel, M. (2011). Internet use and video gaming predict problem behavior in early adolescence. *Journal of Adolescence*, *34*, 49–58.

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

Kazerooni, F., Taylor, S. H., Bazarova, N. N., & Whitlock, J. (2018). Cyberbullying bystander intervention: The number of offenders and retweeting predict likelihood of helping a cyberbullying victim. *Journal of Computer-Mediated Communication*, *23*, 146–162.

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 of the United States of America*, *113*, 13696–13701.

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

Mishna, F., Cook, C., Saini, M., Wu, M. J., & MacFadden, R. (2011). Interventions to prevent and reduce cyber abuse of youth: A systematic review. *Research on Social Work Practice*, *21*, 5–14.

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

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.

Perkins, H. W., Craig, D. W., & Perkins, J. M. (2011). Using social norms to reduce bullying: A research intervention among adolescents in five middle schools. *Group Processes & Intergroup Relations*, *14*, 703–722.

Poole, E., Giraud, E. H., & de Quincey, E. (in press). Tactical interventions in online hate speech: The case of #stopIslam. *New Media & Society*.

Prentice, D. A. (2007). Prescriptive vs. descriptive norms. In R. Baumeister, & K. D. Vohs (Eds.), *Encyclopedia of social psychology* (Vol. 2, pp. 629–630). Sage Publications.

Smith, A. (2006). *The theory of moral sentiments*. Dover.

Soral, W., Liu, J. H., & Bilewicz, M. (2020). Media of contempt: Social media consumption increases normativity of xenophobic verbal violence. *International Journal of Conflict and Violence*, *14*, 1–13.

Timberg, C., & Dwoskin, E. (2020). Silicon Valley is getting tougher on Trump and his supporters over hate speech and disinformation. *Washington Post*. https://www.washingtonpost.com/technology/2020/07/10/hate-speech-trump-tech/

Ybarra, M. L., Diener-West, M., Markow, D., Leaf, P. J., Hamburger, M., & Boxer, P. (2008). Linkages between internet and other media violence with seriously violent behavior by youth. *Pediatrics*, *122*, 929–937.

Zaki, J. (2014). Empathy: A motivated account. *Psychological Bulletin*, *140*, 1608–1647.

How to cite this article: Bilewicz M, Tempska P, Leliwa G, et al. Artificial intelligence against hate: Intervention reducing verbal aggression in the social network environment. *Aggressive Behavior*. 2021;47:260–266. https://doi.org/10.1002/ab.21948