Graphical Abstract

We need to beat namespotting, no idea how

Us

Not sure what graphical abstract is supposed to be, but it goes here.

Highlights

We need to beat namespotting, no idea how

Us

- Research highlight 1
- Research highlight 2

We need to beat namespotting, no idea how

Us

[a], , , , ,

Abstract

During a six-month experiment conducted on Reddit, we studied the impact of counter-speech interventions against personal attacks on the number of personal attacks sent by 440 users regularly attacking others. We used two types of interventions, normative—which referred to social norms, and empathetic—which referred to emotions and encouraged perspective-taking. We employed a collective intelligence approach—the collaboration between human and machine intelligence. Artificial Intelligence was used to detect verbal aggression and notify human volunteers, who then performed the interventions. We analyzed the data from three perspectives. We used time series models (1) of the short-term impact of individual interventions, and of (2) the cumulative impact of interventions received as the experiment progressed. We also (3) used aggregated data for a long-term before/after analysis. The short-term effect of interventions is damaging: users tend to be on average around 26% more aggressive the next day, but the effect does not last beyond two days. The cumulative effect of interventions is helpful: each intervention (up to around 8-10 total, the effectiveness of more interventions tends to be lower) decreases daily aggression by 4% on average and the effects accumulate and balance out the short-term effect in the long run. The effectiveness of normative interventions seems overall higher, except for the less aggressive offenders, for which empathetic interventions might be equally or more useful.

Keywords: counter-speech, artificial intelligence, collective intelligence, peaceful interventions, Reddit, no i nie wiem jeszcze który do przemocy słownej: online aggression, online abuse, personal attacks
PACS: PACScode, PACScode
2000 MSC: MSCcode, MSCcode

Contents

1. Introduction

Much effort has been made to tackle the problem of verbal aggression and harassment online, but looking at various reports and surveys [47, 86, 101],

it remains a common hindrance for people who engage with social media in their everyday lives. The problem was exacerbated during the COVID-19 pandemic, during which much of our social life moved to cyberspace. This shift was followed by an increase in cyberbullying attitudes and perpetration [3], 90% increase in public reports of illegal online content [34], including 114% increase in non-consensual sharing of intimate images, 30% increase in cyberbullying, as well as 40% of increase in adults reporting online harassment. According to a report by company L1ght [45], hate speech directed towards China and the Chinese went up by 900% on Twitter. Gaming platforms were in the spotlight as well, with a 40% increase in toxicity on Discord.[1]

Unfortunately, the growing need for even more efficient and proactive moderation was not followed by an increased capacity to execute it, forcing companies and policymakers to rethink the current model of moderation processes and workforce. Due to the COVID-19 restrictions, including social distancing, a lot of those serving the role of moderators had to be sent home [6] without the ability to work remotely because of the constraints affiliated with restrictive non-disclosure agreements (NDA), among others. Curtailing the moderators' workforce was accompanied by more agency given to algorithms and AI-based moderation. As argued by Gerrard [31], those changes are a serious safety red flag for all the users of online platforms. At the same time, looking at the prevalence of online violence and the scope of the problem, cooperation between humans and algorithms seems to be inevitable. The question is how to execute such a partnership, taking into consideration the technological constraints and human limitations, and the strong suits of both.

The matter in question is of great significance, as online violence (including cyberbullying, sexual harassment, hate speech, or any other type of abuse online) poses a great risk to individuals and groups alike with its harrowing consequences. On an individual level, online abuse was found to correlate with anxiety, lower self-esteem, feeling of hopelessness, isolation, substance abuse, decreased academic performance, and in more extreme cases self-harm, suicidal ideation, or suicide [109, 37, 1, 44]. On a group level, it has been linked to group polarisation and radicalization [51, 103].

[1] In both reports, the increase is reported as a relative change between the years 2019 and 2020, with no absolute indicators.

Online hatred running rampant on social media platforms can incite atrocities such as genocide—The Rohingya genocide was preceded by an increase in anti-Muslim rhetoric and a genocidal campaign occurring on Facebook [27]. There is also research on how the prevalence of hate speech can lead to desensitization—a decreased ability to judge hateful statements as hateful, and in turn to an increase of prejudice [84], and of social acceptability of hate speech [2], or increase of offline hate crime [104]. Online abuse is also a threat to the stability of online services and their communities. It can contribute to lower user engagement or churn [97], directly impacting the profitability of companies. In some countries, there are new regulations that introduce severe fines for companies not reacting to reported hate speech or other types of online harm [96, 95]. In the past, there were also instances in which companies pulled their ads from online platforms, where violent speech was prevalent [73]. On the more extreme side, some companies were sued or even shut down because of online abuse and suicides related to it [63]. Cyberbullying and gun-related threats have been found to play a major role in YikYak decline and shutdown [26]. Snapchat was sued by a mother of a teen who committed suicide after being bullied on a platform [22], while Formspring closed down in 2013 after a series of suicides among its users [82].

In this paper, we describe a novel experiment in the collective intelligence paradigm designed to reduce the level of verbal aggression on the Q&A forum Reddit. The collective intelligence approach in its underlying principles entails cooperation between humans and machines to solve a particular problem. Detection algorithms were used to automatically detect verbal aggression in the form of personal attacks and to notify human volunteers, who then responded with counter-speech interventions aimed at the reduction of verbal aggression in study participants.

This paper is organized as follows: first, we will sketch the landscape of online content moderation and locate the techniques we used within it. Second, we will outline the theoretical underpinnings of counter-speech interventions utilized in the experiment. Then, we describe the experiment itself. Having done this, we wrap up with a short discussion of volunteer engagement in this experiment.

2. Automated versus human-based moderation

The hindrances and threats that go along with the Artificial Intelligence-based methods for moderation have been widely debated, with the most

critical discussions revolving around the technology performance [55, 81]. State-of-the-art solutions are mostly governed by statistical methods, including deep learning and machine learning [39, 50, 83]. Their performance is inherently tied to the amount of data being fed to the system and annotation quality.

At different stages of the process, from data gathering and preparation, annotation, to the training or algorithms themselves, biases seem to be omnipresent [9, 32, 58]. For instance, Sap et al. [79] examined how the lack of knowledge of data annotators about the dialect of minorities can lead to the amplification of bias against those minorities. When improperly trained, annotators repeatedly make incorrect judgments about comments expressed in African American English dialect, those mistakes are repeated by algorithms once the model learns to identify hate speech using such datasets. Since algorithms are getting more control in the decision-making process of who can participate in discussions online, such errors can have detrimental effects on people of African-American descent, who can be excluded from certain communities.

Yet another bias can result from the technical shortcomings of statistical methods such as deep learning and machine learning. For instance, Perspective API,[2] a tool for detection of verbal aggression online, was repeatedly judging neutral sentences such as: "I am a woman", "I am a Jew", or "I am a black gay woman", as highly toxic [100]. Such errors are inextricably related to the foundations of deep learning and machine learning models, which learn directly from the annotated data. Since women and minorities are frequently a target of attacks, such algorithms, which have no understanding about the structure of the sentence and the grammar behind it, make statistical approximation and incorrectly give a high toxicity score because of the occurrence of "sensitive" keywords—woman, Jew, black, gay, and so on.

Users of online services are also creative in their strategies of circumventing automated content moderation systems, and as shown by Gröndahl et al. [35], current techniques are vulnerable to the most common evasions such as word changes (inserting typos and leetspeak), word-boundary changes (inserting or removing whitespace), or word appending (appending common or non-hateful words). For instance, in the evaluation of seven state-of-the-art models trained for hate speech detection, adding "love" to abusive

[2] `https://perspectiveapi.com`

sentences—a word which is negatively correlated with hate speech—resulted in the F1 score[3] preformance decrease to almost 0% in the majority of models tested.

Lack of generalizability of the models—the ability to perform well on datasets coming from sources other than the one used for training—is a severe shortcoming as well [77, 112]. As shown again by Gröndahl et al. [35], models trained on a dataset coming from Wikipedia, which achieve 86% of F1 score, perform below 50% F1 on Twitter (between 24% to 50% depending on the model and the Twitter sample). The adaptability from Twitter to Wikipedia looks even worse. One of the models trained on a Twitter sample with a performance of 83% F1 dropped to 14% on a Wikipedia sample.

As shown by Wu et al. [107], Lipton and Steinhardt [54], Musgrave et al. [62] in practice, the development of such models often lacks thorough error analysis and legitimate experimental methodology, which can result in non-reproducibility. This is also connected with a potential lack of thorough understanding of the limitations of the models and spurious conclusions being announced to a wider public. Specifically, Lipton and Steinhardt [54] distinguishes four dysfunctional patterns occurring in the current research paradigm in the industry and academia alike:

1. First, the inability to draw a clear distinction between speculation and explanation, with the former often being disguised as the latter. For instance, in Steinhardt et al. [91] the author admitted to stating that "the high dimensionality and abundance of irrelevant features...give the attacker more room to construct attacks"—although no experiments were conducted to measure the effect of dimensionality of the neural network on its attackability.
2. Second, the inability of successful identification of the sources of performance improvement (whether it was problem formulation, optimization of the heuristics, data-preprocessing, hyperparameter tuning, or perhaps yet another aspect). As was shown by Melis et al. [59], some improvements in language modeling, which originally were ascribed to complex innovations in the architecture of the network, stem from hyperparameter tuning. As mentioned by Lipton and Steinhardt [54], there is a tremendous value coming from the thorough understanding of a particular method, and a variety of techniques are vital in the

[3]This is an overall measure of a model's accuracy that combines precision and recall.

process (like ablation, robustness checks, qualitative error analysis) for the benefit of the whole community.

3. Third, "mathiness"—the use of obscure language and often covering weak argumentation with the alluring but often apparent depth of technical jargon. Again Jacob Steinhardt admitted infusing his 2015 paper co-authored with Percy Liang Steinhardt and Liang [92] with an irrelevant theorem to amplify the empirical results. They discussed "staged strong Doeblin chains" which actually had limited pertinence to the learning algorithm, which was the main subject of a paper.
4. Last but not least, misuse of language. This includes suggestive definitions without proper explanations of what they mean in the context (e.g., calling good performance in simple NLP tasks a human-level understanding), overloading the papers with technical terminology, or suitcase words (words that can encompass a variety of meanings, e.g. consciousness). In one of the examples described, empirical results reported by Esteva et al. [25] in which a classifier achieved low error on independent and identically distributed test data was described as "dermatologist-level classification of skin cancer," omitting the fact that dermatologists judge cases of much wider variety.

Yet another obstacle in the process is the lack of gold standard in dataset construction and taxonomies of abusive language being used, for instance, in the process of annotating different datasets. Frequently people obtain data from various sources and do not follow any universally used instructions when it comes to annotation, what results in lack of annotation coherency across various datasets related to one phenomenon (e.g., hate speech). The lack of expert annotators and proper annotation criteria and instructions are also widespread, with the common practice of hiring untrained workers from Mechanical Turk or other crowdsourcing platforms.

Although there are some initiatives developed in response, most notably, functional tests for Hate Speech Detection Models created by Röttger et al. [78], or the Online Safety Data Initiative (OSDI) [21], focused on projects related to improving access to data, standardizing the description of online harms, as well as creating tools and benchmarks for evaluation of technologies focused on safety, much effort must be made before wider adoption of such solutions comes into force.

At the same time, only automated methods can scan through the massive amount of content being generated every day on different platforms. On

Facebook, there are more than 3B comments and likes daily [57], 500M tweets are sent daily on Twitter [53], and over 2B comments made by users of Reddit in 2020 [72], which is almost 3M comments made daily. With this amount of content, it's either impossible or extremely costly to scale the moderation workforce. One can also have doubts about the ethical aspects of hiring workers who are often unaware of how this kind of task will affect their well-being. Being submersed in the cyber-Augean stables takes a toll on many—as examined by Roberts [75, 76]. Workers hired for such tasks are often low-status and low-wage, isolated, and asked to keep what they've seen in secret under restrictive NDAs. This in turn, makes the research in the area extremely difficult, since moderators are not allowed to talk about their work conditions or any other related subject. Those who decide to break the NDA risk a penalty.

Screening through the reported user-generated content is connected with exposure to violent and deeply disturbing materials, with child pornography, murders, or suicides as examples of the most extreme cases. This can lead to serious psychological damage, such as depression or PTSD [75]. Although there are certain initiatives being developed or introduced to reduce the emotional impact of the moderation, such as stylistic alterations to content (for instance, applying grayscale or blurring to images [41], workplace wellness programs, clinical support, or psychological training [89], none of the methods can eliminate the psychological distress completely. Some of the employees filed a lawsuit against Facebook, and as a result, the company agreed to pay $52M in compensation for mental health issues developed during the job [64]. Also, as described by Parks [67], the work is often performed under time pressure, reviewing 25K pieces of content per day. Spending on average three to five seconds on each image reported for moderation might not lead to the most thoughtful decisions, and as shown by Stepanikova [93], high time pressure can amplify human biases. Considering that Facebook employs 15K moderators [43] and most likely more are needed to keep up with the growing amount of content, with the parallel considerations about the negative effects of content moderation on mental health, a carefully thought-out collaboration between humans and machines in this area seems indispensable and highly demanded.

3. Different approaches to content moderation

There are different approaches when it comes to the moderation of online content. One can follow the workflow of reactive moderation, which happens once the content is published. Harmful messages can be either reported by the users of the platforms or identified by automated methods and then sent for review to moderators. A set of actions can then be taken depending on the platform, its community guidelines, on the content, or on the user level. A harmful message can be deleted, made invisible to other users, or certain profanities can be altered with special signs to censor them. Depending on the type and amount of infraction, a particular user can be warned, muted, shadowbanned, or banned from further participation in the community for a period of time. The weakness of the reactive method is that the damage is already done. Whoever is the recipient of the abusive message has the chance to see it and potentially suffer [38, 42, 102]. Yet another weakness connected with relying solely on human reports is the content that is harmful but unreported by its recipient or bystanders. Although the exact scale of unreported negative content is not known, various self-report studies show that a lot of children, teens, or even adults do not report cyberbullying or harassment online [29, 49, 46].

Yet another type of moderation is pro-active moderation. On this approach, automated methods used are either based on Artificial Intelligence or other less sophisticated tools (e.g., blacklists) and can screen the content before it gets published. If a harmful message is detected, it can be removed before reaching the recipient. Due to the aforementioned dubious performance of state-of-the-art statistical methods, particularly low precision, such tools are rarely used autonomously, and serve as additional support to content moderators. Alternatively, all content can be reviewed by moderators before it gets published on a website, although such a process is rarely implemented, as it stifles the discussion.

Another categorization can be obtained by looking at who is engaged in the moderation of the content. In a distributed or decentralized moderation system, the community members decide which content gets taken down, as opposed to a centralized moderation system, in which professionals hired by a particular service are in charge. Some services use a hybrid approach, with Reddit being one of such cases. Consisting of numerous communities initiated by Reddit's users, the owners of the so-called subreddits (mods) are in control of the moderation within the subreddit. At the same time, Reddit

hires a small centralized team of moderators (admins) that enforce content policy on a broader scale.

4. Counter-speech interventions

The term counter-speech refers to the use of means of communication to counteract hate speech or other forms of harmful or extremist speech. It is a form of counter-narrative that seeks to challenge and undermine the messages of hate and extremism, and to promote alternative, positive, and inclusive narratives. It can be broadly defined as citizens' response to hate speech aimed to stop it and reduce its consequences [48, 30]. Hate speech is a type of speech used to insult, offend, or intimidate a person based on their identity (such as race, religion, sexual orientation, national origin, or disability) [24]. This type of speech is often used to spread hateful or discriminatory ideas, both offline and online, and can take many forms, such as name-calling, threats, or slurs. Social media platforms often recommend their users to employ counter speech as a way to stop hate speech from spreading, conveniently handing the problem out to dedicated users to avoid creating new technologies or investing in manual moderation [80].

Benesch et al. [5] defined successful counter-speech in two ways. The first type has a favorable impact on the hateful user, for example by shifting their discourse or beliefs, and the second one impacts the whole audience of counter-speech, by shifting the norm of the conversation. Rieger et al. [74] proposed a three-layered classification of counter-speech: primary prevention, where counter-voices are being disseminated to the general public to educate and promote civic behavior; secondary prevention, where counter-speech is being directed at specific groups, for example, those more vulnerable to harmful effects of online hate, and tertiary prevention where counter-voices respond directly to those spreading hate. Our responses fulfilled all these functions, as they were publicly visible comments sent as direct responses to hateful users.

When it comes to the effectiveness of online counter-speech there is just a handful of large-scale studies, and even fewer of those who distinguish between different types thereof. One of them used a special situation in Germany, where groups that label themselves as hateful or counter-speaking talked about immigration and elections. Results based on analysis of 130,000 Twitter conversations suggest that counter-speech can help make discussions

less divisive, encourage more counter-speech, and lead to fewer hateful responses [30].

Another study by Ziegele et al. [114] looked at almost 10,000 comments on the Facebook pages of 15 news organizations in Germany to investigate the patterns, determinants, and potential effects of interactive moderation. The authors found that the moderation style of initial comments was related to the presence of incivility in users' subsequent reply comments. Specifically, a sociable moderation style was associated with a decrease in incivility in reply comments, while a regulative style was associated with an increase in incivility. This shows that the style of counter-responses can also determine their effectiveness.

Moreover, Ziems et al. [115] looked at the dynamics of the spread of racial hate during a pandemic and the role of counter-speech in mitigating this spread. Using a hand-labeled data set of 3,355 tweets, the authors trained a text classifier to identify hate and counter-speech tweets and conducted a longitudinal analysis of tweets and users. Interestingly, hateful and counter-speech users interacted and engaged extensively with one another, rather than staying in isolated, polarized communities. Results revealed that Twitter users were more likely to become hateful after being exposed to hateful content and that counter-speech messages may discourage users from becoming hateful, suggesting that consistent counter-speech could be a potential remedy for hate on social media platforms.

Xie et al. [108] examined the influence of committed minorities and looked at how a small group of people with an opinion different from the majority can quickly change the whole group's opinion. They found that if more than 10 percent of the group expresses another viewpoint, there is a dramatic decrease in the time taken for the entire population to adopt it. In other words, a relatively small group of committed individuals can have a significant impact on the opinion of the majority.

Similarly, Schieb and Preuss [80] used a computer simulation to study whether counter-speech can curb hate speech on Facebook. They found that even a small group of users speaking out against hate can influence a much larger group, as long as there are enough people in the audience who do not hold extreme opinions. They also emphasized the potential negative effects of counter-voices, especially when they do not hold to the same standards of civic discussion but instead act in an authoritative or mocking way towards those expressing extremist ideologies.

Although some studies suggest censorship and elimination of hateful con-

tent to be a more effective strategy of toxicity reduction [2], such methods also collide with civil liberties and diffuse rather than reduce hate [14]. Therefore, at least *prima facie*, counter-speech seems to be a golden means for reducing antisocial behavior while preserving civil liberties.

In our study, we have applied counter-speech interventions to a slightly different phenomenon, namely personal attacks. Personal attacks are instances of speech that are intended to hurt, offend, or belittle a specific person or group. These can include name-calling, insults, or other types of derogatory language that is directed at an individual or group [8]. While personal attacks and hate speech can sometimes overlap, the main difference between the two is that personal attacks are focused on an individual or a specific group of people, whereas hate speech is aimed at an entire group of people based on their identity.

Although counter-speech can be an effective strategy for toxic discourse mitigation, it usually occurs as a spontaneous act, as it is determined by personal motivations of the counter-speaking party. For example, Brauer and Chaurand [12] found that being personally implicated by a counter-normative behavior was a primary determinant of the bystander reaction. Apart from that, the more public the space, the less likely it is that someone will step up, as studies on bystander interventions consistently show that the presence of other people in critical situations reduces the probability of assistance from bystanders [28]. Moreover, even if a bystander decides to react, they will not necessarily choose the most effective strategy, which can lead to potentially undesirable side effects. For instance, counter-voices that do not hold up to civic standards themselves but act in an authoritative manner or mock those falling for extremist ideologies can actually strengthen the underlying attitudes [114, 80, 52].

To maximize the potential impact of our interventions and to minimize accidental instigation of more verbal aggression, we decided to educate the volunteers on scientifically grounded strategies for normative influence and persuasive messaging. Drawing from research on social compliance and aggressive behavior, we have created two intervention categories—empathetic and normative—each based on a different mechanism of influence. We have used the distinction from our previous experiment where normative and empathetic interventions were automatically generated by a bot in response to personal attacks on Reddit. Both types of responses significantly decreased the prevalence of personal attacks, although the study had its limitations. For example, the interventions (in a form of public comments) were sent

within certain subreddits, so we could not control if the users from one intervention group have not seen interventions from another group [8]. The current study addresses these limitations by employing a fully randomized experimental design and extending the study to the entire Reddit platform. Coming back to our example, this means that the probability that a subject encouters an intervention directed at someone else is much lower.

Another difference is that in the previous study, the interventions were generated automatically, and in this study replies were written and sent by human volunteers. We have engaged volunteers personally committed to lowering online violence and created an opportunity structure for quick response, by providing real-time notifications about personal attacks that should receive an intervention. We hoped that choosing intrinsically motivated volunteers, and providing them with guidance and AI assistance would solve the issue of bystander apathy (see however our section on volunteer engagement for a discussion of difficulties with this strategy).

Empathetic interventions. Empathetic interventions were thoughtful responses delivered in a gentle and straightforward manner, empathizing with either the sender of the personal attack, the user who received it, or the both of them. We hypothesized that presenting hostile individuals with empathy and kindness would reduce further attacks by interfering with their script of aggressive behavior. However, we had no way of verifying the direct effects of our interventions other than the change in the number of personal attacks generated by users.

According to literature, empathy can decrease the tendency to act in a hateful and derogatory manner [113]. Studies on desensitization to hate-speech against refugees conducted by Bilewicz [7] and Soral et al. [85] found empathy to be an effective countermeasure to this phenomenon. Hate-speech desensitization is a lowered sensitivity to hate speech caused by frequent exposition to it. As a result, the agent stops perceiving such speech as harmful, which in turn increases their prejudice towards the victims, and, in extreme cases, leads to the dehumanization of stigmatized groups. Empathy-inducing interventions made subjects more sensitive to hateful content and more supportive of hate-speech prohibition.

Research on peer aggression has shown that both cognitive and affective empathy can reduce bullying behavior and increase the likelihood of a bystander helping a victim [99]. A study looking at predictors of either engaging in defending or remaining an outsider in bullying situations, found empathy to be a strong predictor of defending behaviors [65]. Empathy was also inves-

tigated as a factor that can mitigate the negative effects of high social status on bullying behavior amongst 461 primary and secondary school students. The results confirmed that high levels of empathy suppressed involvement in bullying and moderated the relationship between bullying and social status [13].

Similar results were obtained regarding online aggression. For instance, online aggression towards celebrities was shown to be linked to low levels of empathy and high levels of moral disengagement [66], and a survey of 2070 secondary school students found that cyberbullies had less empathy [88]. A study [4] looking at empathy and cyber-bystanding behavior among adolescents, found both affective and cognitive empathy to be positively related to prosocial cyber bystander behavior, with affective empathy being a stronger predictor than cognitive empathy. Moreover, the study investigating the bystander effect in cyberbullying has shown that factors such as empathic concern and immediate empathic response increased provided support [56].

Wright and Wachs [106] conducted a longitudinal analysis of the consequences of being a bystander to homophobic cyberbullying on a sample of 1067 adolescent students. According to their findings, witnessing homophobic cyberbullying led to increased perpetration of homophobic cyberbullying, but this effect was moderated by empathy and toxic online disinhibition. In other words, high empathy level was one of the factors preventing students from perpetrating online aggression.

Finally, Chin et al. [16] investigated how conversational agents (such as chatbots and smart speakers) should respond to being verbally abused by their users. Specifically, they wanted to understand what kind of response style has the most positive impact on emotions that mitigate users' aggressive behaviors. The authors hypothesized that empathetic responses addressing the harm caused by the users will reduce further abuse of the agent by eliciting feelings of guilt in the abusers. Indeed, the empathetic response style was most effective in reducing verbal abuse towards the conversational agent, although the feeling of guilt was not directly measured in this study.

Normative interventions. The normative interventions were based on the principle of normative influence. Their objective was to reduce personal attacks by lowering the perceived acceptability of such behavior in a particular online setting. The normative approach stems from the psychological research of normative social influence, stating that in ambiguous situations, people tend to rely on social cues, to determine what kinds of behaviors will be accepted.

Research has repeatedly shown that perceptions of what others are doing cause us to behave similarly. Some scholars propose that it is because perceived social norms serve as shortcuts helping us to quickly adapt to new environments [19]. Others say that we are only motivated by norms we have internalized during socialization, so the normative influence can only be exerted if the norm resonates with our internal value system [60]. Either way, social norms are agreed to be powerful determinants of human behavior that can be applied to solve social problems, such as littering or resource waste [17, 19, 33, 60].

One important distinction made within this approach is between two types of norms: descriptive norms that refer to what is typically done and are related to how others behave, and prescriptive (or injunctive) norms that explicitly describe what is commonly approved, for example, the rules of proper conduct and moral beliefs [19]. In other words, a descriptive norm motivates compliance by providing evidence about what is likely to be effective and adaptive action, while a prescriptive norm motivates by informing the agent which actions are approved and which will result in social sanctions [110, 18, 40].

In a study on littering in public spaces conducted by Cialdini et al. [19], study subjects were exposed to either descriptive or injunctive anti-littering norms, and the impact on the littering behavior was observed. In the descriptive condition, the subjects were exposed to a clean or littered environment, serving as proof of choices made by others. In the injunctive condition, the participants received messages stating the anti-littering norm. Results have shown that prescriptive messages directly asking not to litter had significantly reduced littering, and the subjects littered more in a highly littered environment and littered less in a clean environment, although these effects were not statistically significant.[4]

Descriptive norms can also be evoked by providing written information about what others usually do. Goldstein et al. [33] distributed flyers with normative messages that contained descriptive information about the reuse of hotel towels ("the majority of guests reuse their towels"). This strategy has been proven superior to traditional appeals focusing on environmental

[4]It is also worth noticing that seeing litter in an otherwise clean environment could have also engaged a prescriptive norm, as it might have reminded subjects about societal objections to littering.

protection. Moreover, normative appeals were most effective when the described behavior occurred in the same location ("in this hotel room"), and so referred to, so-called, local norms.

The information about how others behave or think can also have an impact on prejudice and its expression. For example, positive information about the racial beliefs of others produced stereotype change in message recipients [87], and public expression of prejudice, as well as reactions to hostile jokes were highly correlated with social approval of these expressions [20].

When it comes to the use of normative influence to reduce uncivil behavior online, several norms were activated trough social sanctions, peer presure and context. In a study by Munger [61] social sanctioning of racist tweets in the form of messages reminding about the human and labeling the behavior, resulted in the reduction of racist tweets, although only in the condition in which a white male with high followers approached another white male [61]. In addition, types of statements made by people in online political discussions influenced the way others expressed their opinions [68], prior exposure to troll content was a strong predictor for trolling behavior [15], and the expression of online hate-speech in [2] was strongly dependent on the amount of hate-speech already present in a given environment.

We have included both descriptive and prescriptive norms among normative interventions. Descriptive norms were expressed by informing the attacker about the prevalence of civil behavior in this community, and prescriptive norms were expressed by informing about the rules of this community.

5. Collective Intelligence and moderation

The main objective of the experiment was to test whether the level of verbal aggression (personal attacks) of a group of users regularly attacking others on Reddit can be significantly decreased by community-driven, counter-speech interventions conducted by volunteers in partnership with Artificial Intelligence (collective intelligence approach).

Verbal aggression was defined in this experiment as a personal attack—any kind of verbal harassment, insult, or threat directed against the interlocutor in a text-based conversation online. Those were detected using Samurai Labs' cyberviolence detection system, describe in detail in Section 6.

Traditionally, collective intelligence has been defined as "a group or a team's combined capacity and capability to perform a wide variety of tasks

and solve diverse problems" [105]. In our paper and in the theoretical underpinnings of the experiment itself, we will be relying on a collective intelligence scope proposed by Nesta, an innovation foundation (`https://www.nesta.org.uk`), which focuses on a collaboration between human and machine intelligence to develop innovative solutions to social challenges.

Instead of using a negative motivation system, the assumption was to test a positive one—convincing verbally violent users to refrain from using cyberviolence based on civic education—peer-pressure regulation and experiential learning of a positive set of norms and empathy. Algorithms developed for the detection of personal attacks were used to monitor the activity of experimental groups and notify volunteers about all attacks generated by its members. Volunteers, after receiving a notification on Slack, could then react with an appropriate intervention. Such an approach served as a distributed bottom-up voluntary model of moderation based on collective intelligence—utilizing human + machine intelligence to develop solutions to social challenges.

Normative interventions can refer either to general social norms of civility and respect, community standards, a particular subreddit rule, or a descriptive norm among others. Empathetic interventions can refer to the emotional state of the recipient or the sender of the attack or both. They are designed to evoke an empathetic response. Here are some examples of interventions that were used during the experiment:

- *Insulting others is against Reddit's policy.* (normative)
- *Hey there, we do not call each other like that here.* (normative)
- *There is a human being on the other side who might be hurt by your words.* (empathetic)
- *I see you are frustrated but remember the human.* (empathetic)

Our hypotheses were as follows:

H1 If a group of human volunteers, notified by an AI-based detection system about verbal aggression generated by the treatment group users, responds with counter-speech interventions, this will result in a decreased verbal aggression level for the whole group after the intervention period.

H2 If two groups receive different types of interventions (empathy-based or normative), then the decrease in verbal aggression will be larger in the case of normative interventions in comparison to the empathy-based ones.

Although empathy has been found to be helpful in the reduction of aggressive behavior [10], and the improvement of intergroup relations [94], there are also important arguments and evidence concerning its limits. Specifically, Paul Bloom in his book *Against Empathy: The Case for Rational Compassion* [11], stresses that empathy can have a very narrow focus and can be prone to biases. For instance, it's easier to be empathetic towards those who are close or similar to us. Since conversations on Reddit are anonymous and usually occur between strangers, we expected empathetic interventions to be less effective than normative interventions in curbing personal attacks.

6. Technology applied for personal attack detection

To perform the research at a scale sufficient for statistically reliable analysis, we needed to apply a technology for the initial automatic detection of personal attacks, to be further forwarded to human moderators for deeper analysis and making decisions regarding the outcome of the moderation (e.g., deletion of the post, sending a warning to an abusive user, banning the user, etc.). As we already mentioned in sections 2, 3 and 5, performing the moderation relying on the assumption that the moderator will read through all of the posts and comments is unrealistic, while moderation based solely on other users' reporting could lead to a large number of abusive messages remaining unreported. This problem, present in the literature [71, 97] and in the mass media[5,6] has lead to an increase in the number or automatic tools

[5] https://www.japantimes.co.jp/news/2022/12/03/business/tech/twitter-harmful-content/

[6] https://www.forbes.com/sites/forbestechcouncil/2022/06/14/the-growing-role-of-ai-in-content-moderation/

being developed and deployed to support content moderators in their work.[7]

In our research we applied Samurai detection software, a proprietary technology developed Samurai Labs.[8] The technology comprises a combination of symbolic and statistical methods, where each statistical component (e.g., a deep learning model) is governed by a symbolic component utilizing a variety of natural language processing methods (e.g., tokenization, syntactic parsing, etc.). Symbolic components are used to determine if a potentially abusive utterance is not a part of a broader utterance indicating that the first one should not be considered abusive. For example, an utterance "you are an idiot" is potentially abusive, but in reality it is not, if it appears as a part of some broader context such as "I cannot believe he said you are an idiot." Another example of using symbolic components is determining if an abusive phrase is targeted against an interlocutor (e.g. using a linking verb to assign the abusive phrase with a second person as in the "you are an idiot" example).

Samurai has already been applied in a variety of studies, including suicidal text detection [70], detection of toxic usernames [98], and detection of personal attacks on Reddit [97]. For the research presented in this paper we used the module of the system developed for the detection of personal attacks.

Personal attacks are defined as any kind of abusive remark made in relation to a person (*ad hominem*) rather than to the content of the argument expressed by that person in a discussion. The definition of 'personal attack' subsumes the use of specific terms which compare other people to animals or objects or make nasty insinuations without providing evidence. Three examples of typical personal attacks are as follows:

- *You sound like a whiney bitch.*
- *Proof my dick in your ass u flairless fucken hoe.*
- *Go cuddle your anime pillow dipshit.*

Figure 1 illustrates how an input text (such as "ccant believ he sad ur an id10+...!") is processed step-by-step utilizing both statistical and symbolic

[7] `https://influencermarketinghub.com/content-moderation-tools/`.

[8] `https://www.samurailabs.ai/`, described in (Ptaszyński, Leliwa, Piech, & Smywiński-Pohl, 2018; Wroczynski & Leliwa, 2019).

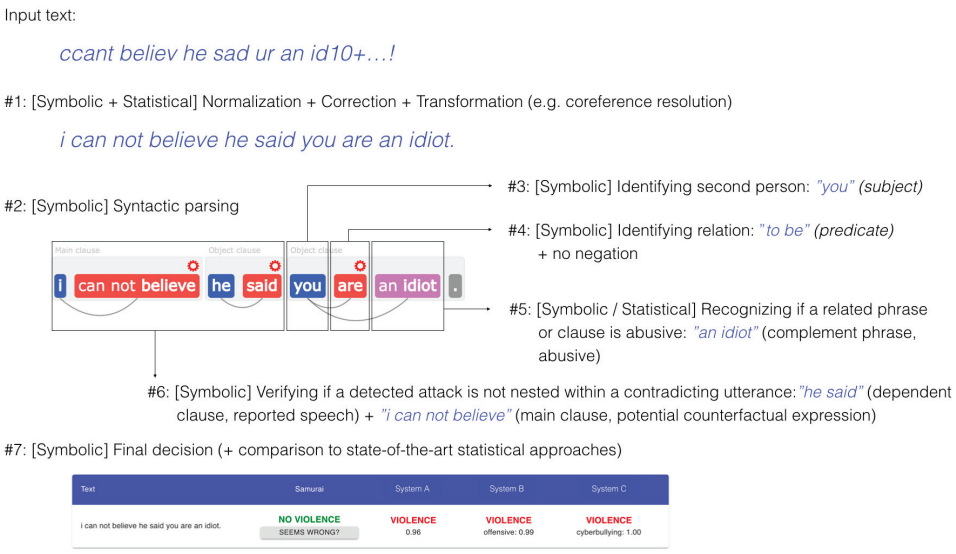

Figure 1: Example of processing of one sentence by the applied Samurai technology.

methods. The performance of Samurai has been measured on a number of occasions and evaluated as sufficiently reliable for practical use [69, 70, 97, 98].

7. Data analysis

7.1. Data collection and exploration

We conducted a 6-month experiment in a digital setting conducted on a popular Q&A and news forum, Reddit (`www.reddit.com`). We formed treatment and control groups based on three main criteria:

1. During the intervention period, we have expected to have 20 active volunteers at any given time, each willing to conduct approximately 10 interventions a day. So, we needed approximately 200 attacks daily generated by the treatment groups.
2. We first conducted a preliminary identification of users who regularly attack others to test interventions on them. The key issue here is that such users are not too common, and attacks are rather rare among other users. If instead we simply picked a random group of users with the intention of intervening in response to their online attacks, the sample

size would have to be extremely large, and still the chance of observing enough attacks to draw any interesing conclusions from the reactions to our interventions would be low. Thus, we need to keep in mind that the group studied is not just users of Reddit, but rather users who for a least a few weeks tended to systematically attack others.

3. The identification of users who were active during the whole preliminary monitoring period was necessary to minimize the risk of attrition during the study.

The user identification process was as follows:

1. First, we obtained 1 week of real-time umoderated data from Reddit (February 15-22. 2020), unmoderated data from Reddit. The content was downloaded from the data stream provided by pushshift.io.
2. Samurai Labs Artificial Intelligence for personal attacks detection was applied to identify those users who attacked others at least once within this period. This resulted in the identification of 93966 users.
3. We removed all accounts which we suspected not to be run by humans (AutoModerator and all users which had "bot" in the username string). This resulted in the removal of 388 users, thus 93578 were left on our list.
4. Next, we removed those users who generated only 1 personal attack during the week. This step resulted in the removal of users generating the number of personal attacks below the third quartile (Q3). 20124 users were left in our group.
5. We removed users who generated less than 14 comments during this week. It was not a lot, since 14 comments were below the 1st quartile (Q1: 28 comments), but such a criterion moved us in the direction of identification of those who are relatively active. This resulted in the removal of 2192 users, so 17932 were left.
6. We discarded users whose personal attacks to all comments ratio was below 2%. This means the inclusion in the sample of users above the 1st quartile of personal attacks to all comments ratio. 4422 users were removed, leaving us with a group of 13510.
7. The next step of the process begun on March 9, 2020, and lasted until May 5, 2020 (9 weeks). During this period we have monitored the activity of the identified group of 13510 users and applied further selection criteria to make sure we select those who were regularly active and attacked other users.

8. The period of monitoring was divided into weeks. We have discarded those weeks during which technical difficulties occurred with `pushshift.io` (resulting in missing data). So we have taken into consideration only 6 full weeks for the period.
9. Users who generated at least 1 attack per week during 5 out of 6 weeks were identified. Initially, we planned to restrict the list to only those users who generate at least 1 attack during each week (6/6), but such a restrictive criterion led to only 255 users left, which was not enough for the study. The less restrictive criterion (at least 1 attack generated during 5/6 weeks) resulted in 694 people.
10. Next, we calculated the daily average number of personal attacks generated by the resulting group (357 attacks per day).
11. Knowing that we need around 200 attacks per day per treatment group (just enough for volunteers to keep up according to our assumption), we have randomly selected 195 users per each treatment group (normative and empathetic). The rest was delegated as a control group (304 users).
12. Some of those were further removed as t data visualisation and content inspection helped to identify them as bots, or they ceased their activity during the period. We were left with the data on 440 users.

The duration of the experiment, 6 months, was divided into three 2-months periods. The first two months served as a monitoring period to properly select groups and establish baselines. The next 2 months served as a treatment period, during which groups received counter-speech comments from volunteers in response to personal attacks detected by the Artificial Intelligence-based system. The last 2-months served as the post-treatment monitoring period to gather the data needed to evaluate the effectiveness of interventions. We started with an observation period on March 9, 2020, leading to the intervention period starting on July 8, 2020, following with a further observation period starting on September 9, 2020, and ending on November 20, 2020. The time series of attacks observed and of interventions conducted can be inspected in Figure 2, along with a quick search for weekly patterns. Some of the users turned out to be bots, a few ceased to be active during the experiment (with no strong reason to think this happened due to them receiving an intervention) and a few received treatment of two different types by accident (we relied on multiple volunteers and such mistakes were likely to happen). Ultimately, in the time series data, we ended up with data on 440 users.

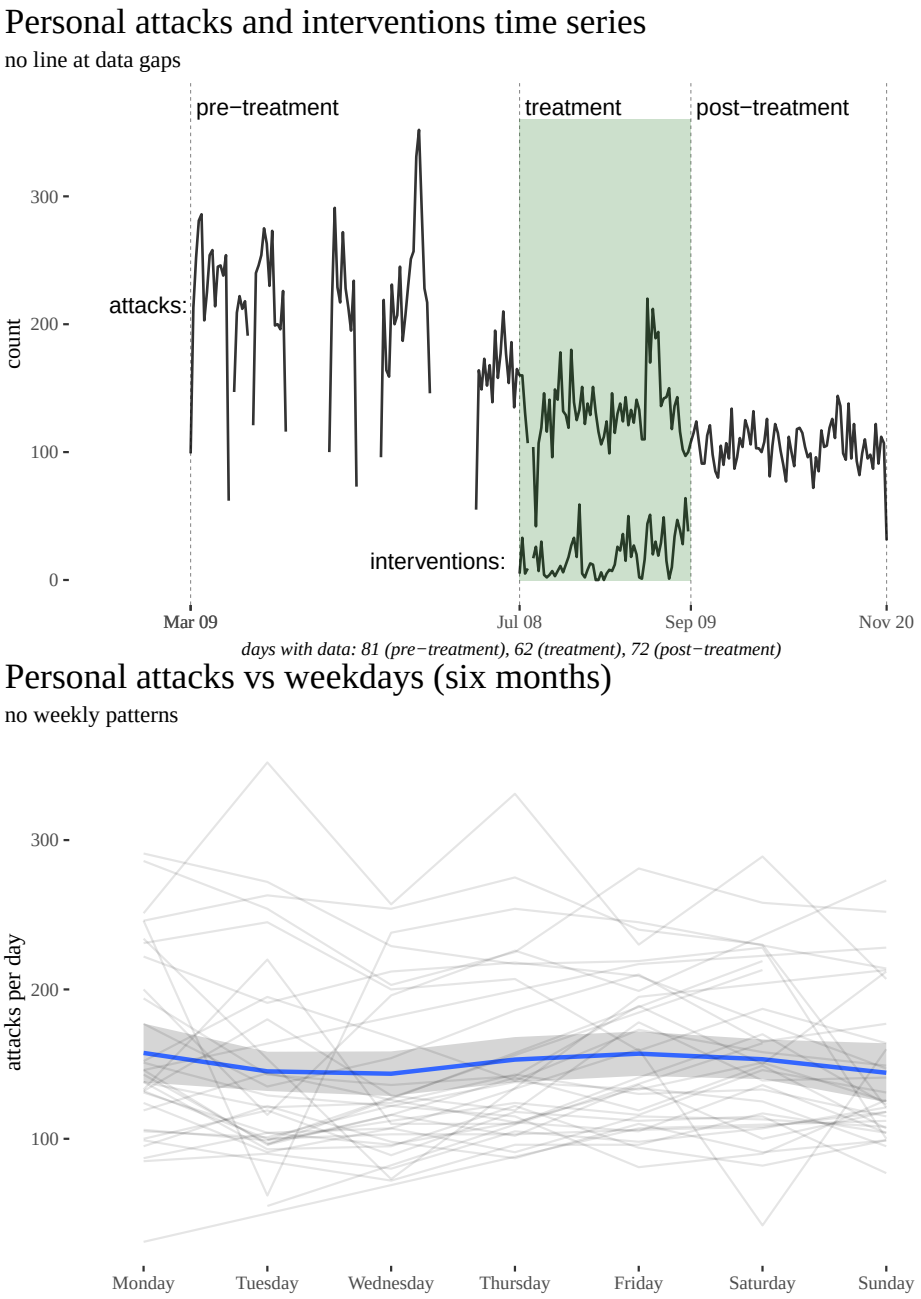

Figure 2: Daily sums of attacks and interventions throughout the three experimental periods, with GAM smoothing (left) and daily attack sums from all weeks in the experimental period plotted against week days (right)—no pattern seems to arise.

variable explanation
AB attacks before (pre-treatment)
AD attacks during (the treatment period)
AA attacks after (post-treatment)
CB comments before
CD comments during
CA comments after
group treatment group
IC intervention count

Table 1: Variables involved in the before-and-after analysis.

We analyzed the data from three perspectives: we used the daily data to (1) build seven time series models estimating the impact of individual interventions at lags 1-7 days, and (2) to study the impact of the cumulative number of interventions received as the experiment progressed, and (3) we used aggregated data to run a long term before-and after analysis, comparing the summarized aggression levels before and after the intervention period.

Before we move to the analysis, let us inspect the data. First, at the aggregated level, the data involve the variables listed in Table 1.[9] The distribution of IC in the treatment groups is visualized in Figure 3. Note that the distributions are somewhat different, even though the total intervention counts are similar. The issue is discussed in Section XXXXX.

add ref

Second, in the distribution of standardized difference in attacks, the peaks of distributions are shifted a bit between the groups, with lowest median for the normative group, but the differences seem minor (Figure 4). This might suggest no impact of the interventions. This conclusion would be too hasty, as the impact of other predictor variables and interactions involved can mask actual associations.

To see how this masking can occur, let us inspect changes in attacks against intervention counts. It turns out that restricting attention to various aggression levels in the before period results in fairly strong changes

[9]Further variables were defined in terms of those, in particular, we will be predicting AdiffS which is the standardized difference AA-AB, and AdiffS, which is the standardized difference CA-CB.The standardized variables are systematically named ⟨variable˙name⟩S.

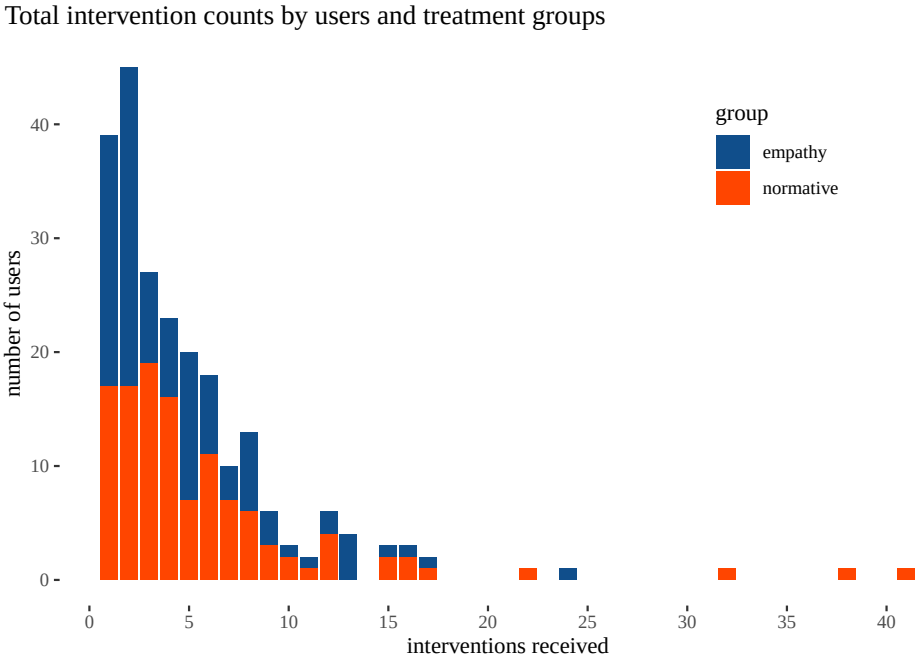

Figure 3: Distribution of daily interventions (by treatment group).

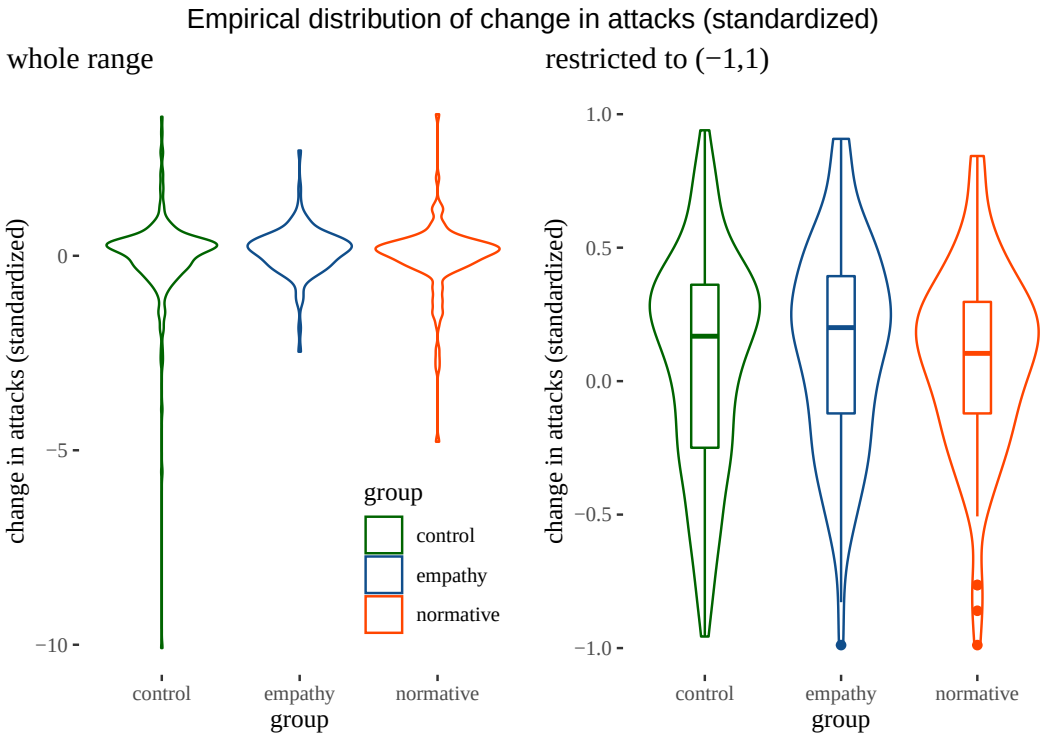

Figure 4: Empirical distribution of change in attacks (by treatment group).

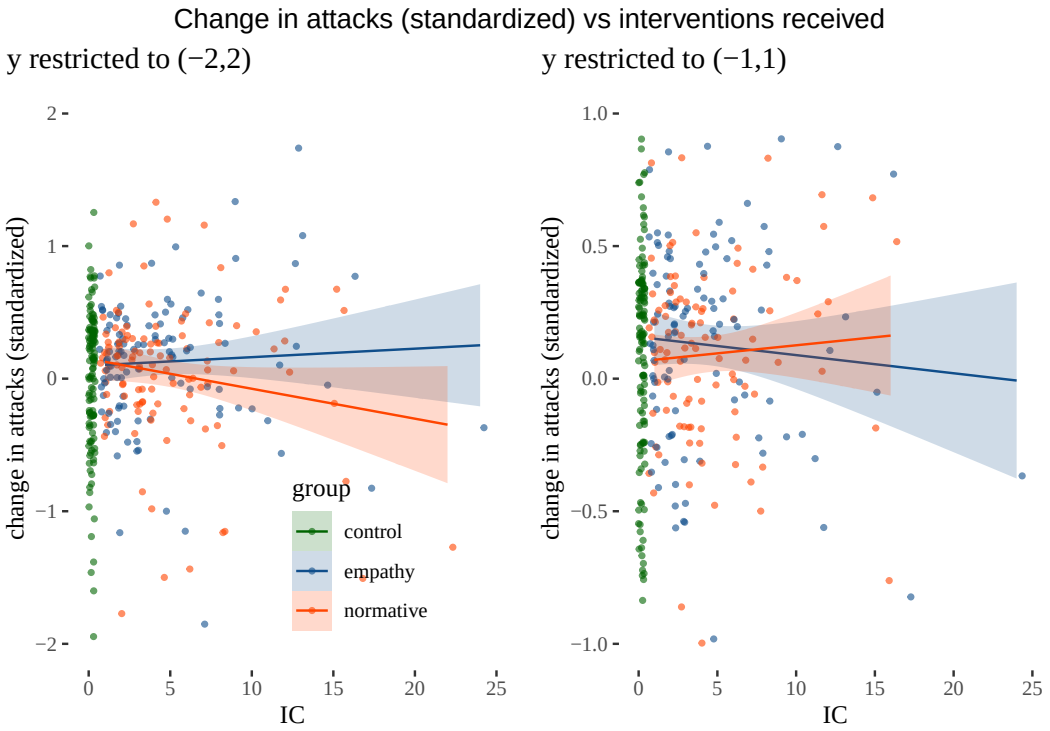

Figure 5: Change in attacks vs the number of interventions received by treatment group, jittered with linear smoothing.

to the regression lines (Figure 5). This suggests we should keep an eye out for interactions with aggression before in the analysis, and that the initial comparison of means or medians between groups might be misleading if the effects in different volume groups are different and to some extent cancel each other.

Further insights, undermining the initial impression suggested by Figure 4, can be obtained by visualizing individual time series. Figure 6 contains six fairly typical examples.

The general phenomenon is that while in the control group attacks tend not to diminish, unless activity itself diminishes, they tend to diminish in the normative group (although the more aggressive the user is, the less of an impact can be observed), and in the empathetic group if the user is not very aggressive. Of course, visualization of individual cases (which the reader might suspect to be cherry-picked) is no replacement for statistical analysis,

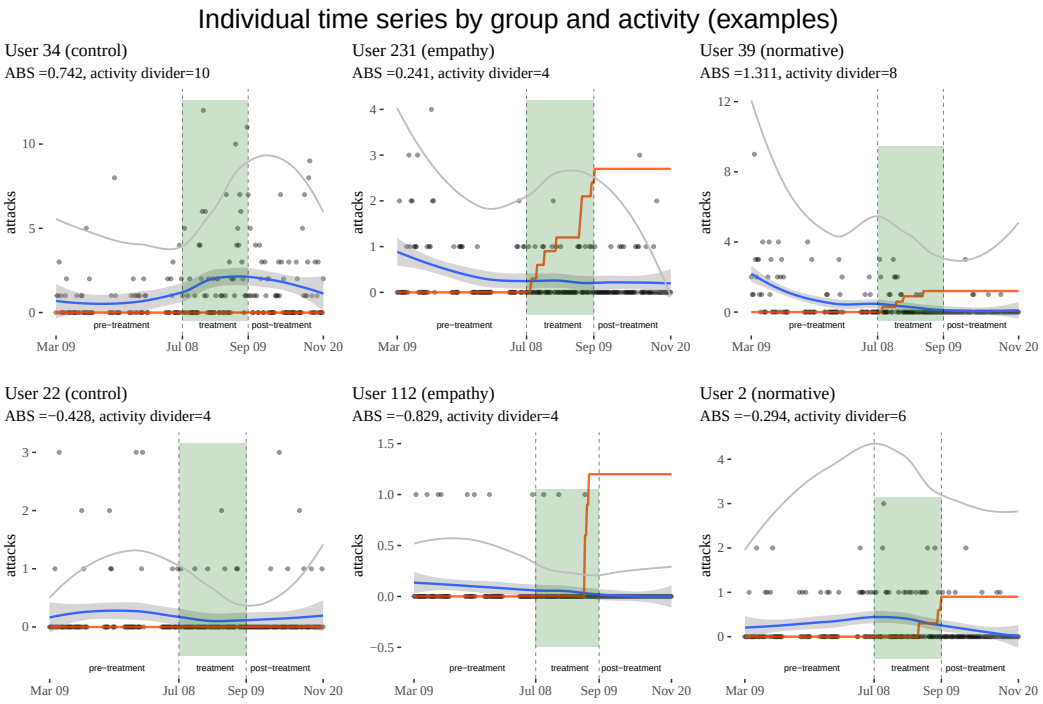

Figure 6: Examples of individual time series. Black points are attacks (smoothed in blue), red lines represent the cumulative number of interventions received (lag 3) divided by 10, gray lines represent overall activity level divided by a variable divider listed in the subtitle. Divisions introduced for visual comparability of general trends.

to which we will now move.

7.2. Causal thinking, choice of variables and models

First, we inspect correlations between predictors to avoid multicolinearity, as highly correlated predictors do not improve predictive performance and artificially inflate uncertainty in their corresponding coefficients in the models. We then develop a plausible causal model of the situation (Figure 7). It turns out that to avoid multicolinearity we cannot condition on CDS (comments during, standardized) if we condition on CAS (comments after, standardized) or CBS (comments before, standardized). Similarly, for the time series data, since activity levels in particular day slices are correlated, it will not be useful to condition on more than one auto-regressive element (and since the predictive power is the highest for lag 1 with no discoverable weekly patterns, we will not go further than lag 1).

To choose the right variables to condition (or not condition) on to identify the causal effect of the interventions, we need to think about the causal structure of the problem. Comments during (the intervention period) impact attacks during, which trigger interventions. Unmeasured user features cause comments before (the intervention period), which impact attacks before directly. Comments during (their impact on ADS is already included) impact attacks during directly and comments after, which impact attacks after directly. Intervention count impacts attacks after and comments after. The same directions of impact are included for intervention type. Finally, comments through time are connected causally, and so are attacks. The structure for the time series data is analogous, except now instead of before and after, we have multiple daily indices.

What do we learn from causal considerations? IT has no backdoor paths, but IC does, so we need to make sure these are closed to avoid including spurious correlations in our analysis. There are in fact 65 different paths from IC to AAC. Crucially, all backdoor paths go through ADS, which then becomes either a fork or a pipe, so all backdoor paths can be closed by conditioning on ADS. Moreover there is only one directed indirect path, it goes through CAS, so we should not condition on CAS if we are to identify total causal effect of IC on attacks, including the impact mediated by its impact on comments. We might be interested in the direct effect of IC and IT on AAS, but then we also need to block indirect causal paths from the intervention to the outcome. For such an evaluation we would need to also condition on CAS and block all backdoor paths from CAS to AAS. This,

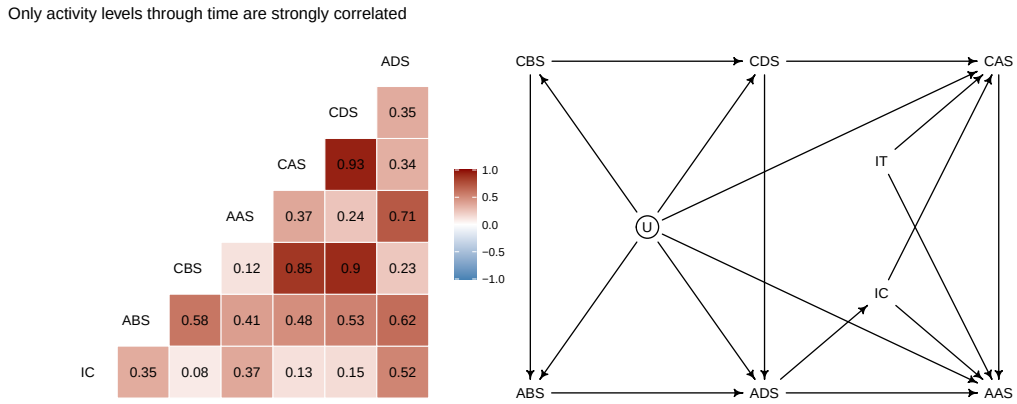

Figure 7: Correlations between predictors (left) and a plausible causal model (right) used in the before-and-after analysis.

however, given the causal model, cannot be achieved, as it would required conditioning on unobserved user features. That is, we do not think direct causal effect is identifiable.

Analogous considerations apply to the time series model for the total impact of individual interventions received exactly k days before (in our case, $k \in \{1, \ldots, 7\}$). The situation, however, is somewhat different for the impact of the total number of cumulative interventions received so far. The trouble is, for example, that if a user received so far a number of interventions until yesterday, some of them had been received before yesterday and those had already impacted their aggression level yesterday. In other words, conditioning on lagged attacks leads to the post-treatment bias and should be avoided.[10]

Otherwise, it's open season for the other variables and interactions between them, and our decision to include or exclude them in the model will

[10]In fact, in the aggregated data analysis, we will be predicting the standardized difference between attacks before and after (ADiffS), and the standardized difference between comments, before and after (CDiffS), but the general points about the nodes involved apply also to defined nodes. As already discussed, we do not include CDS because of its strong correlation with CBS. We also do not condition on ABS when modeling ADiffS (or on CBS when modeling CDiffS)—not only because it has a pretty strong correlation with another predictor (ADS), but rather also because it is used to define the output variable. In such a set-up, it is clear that a model including ABS would have better predictive power, but since a definitional connection is present, thinking that its inclusion in the model tells us something about causality would be misled.

be guided by information-theoretic criterion of predictive power, whose more detailed explanation is included in the appendix, the so-called Widely Acceptable Information Criterion:

$$\mathsf{WAIC}(\mathsf{y}, \Theta) = -2(\mathsf{lppd} - \overbrace{\sum_i var_\theta \mathsf{log} p(y_i|\theta)}^{\text{penalty}})$$

We also use posterior predictive checks in cases in which the likelihood functions used by the models to be compared are different and information-theoretic calculations might be misleading. In such cases we investigated the ratio of actual observations included in the 50% and in the 89% posterior predictive distribution, and the models for which higher ratios were observed in both were selected (no case of diverging evaluation for the two criteria has been observed).

In our model building we used the rethinking package, except for the cumulative impact time series models, where it becomes computationally unfeasible, in which case we built models in Rstan directly. Moreover, for the time series analysis we will build hierarchical Bayesian models which tend to have around $u \times 2p + 2p$ parameters (we will explain later why), which means that for 440 users our final model with interaction with six predictors would have $440 \times 12 + 12 = 5292$ parameters, and have to be trained on daily data for 7 variables collected for six months. The building of such a model on a modern computer takes days. Since in reaching this model we needed to build multiple somewhat simpler models or models with different structures and test their performance, model selection on the full data set was unfeasible. That is, in the time series analysis in model selection at each step we compared models (sometimes built with quadratic approximation) with respect to three independent samples for $40, 60$ and 60 users. We made the decision only if a given model structure performed better on all these subsets (which was usually the case, so the model selection criteria gave us pretty robust answers). For the most complicated model of the impact of cumulative number of interventions, building a single model for the whole data set was not computationally feasible (computation time does not increase linearly with the number of users included in the dataset), so we randomly split the dataset and provided results for the subgrous—the results were not very divergent and the highest posterior density intervals were not very wide.

For the time series, the model that the procedure led us to is as follows (see the appendix for a detailed explanation of how this model has been

reached):

$$
\begin{gathered}
\mathsf{attacks}_i \sim \mathsf{NegativeBinomial}(\lambda_i, \phi_{\mathsf{userID[i]}}) \\
log(\lambda_i) = l_{\mathsf{userID[i]}} + a_{\mathsf{userID[i]}} \times \mathsf{attacksL1} + c_{\mathsf{userID[i]}} \times \mathsf{act} + \\
+ i1control_{\mathsf{userID[i]}} \times \mathsf{control} \times \mathsf{intL1D} + \\
+ i1emp_{\mathsf{userID[i]}} \times \mathsf{emp} \times \mathsf{intL1D} + \\
+ i1norm_{\mathsf{userID[i]}} \times \mathsf{norm} \times \mathsf{intL1D} \\
l_{\mathsf{userID[i]}} \sim \mathsf{Norm}(\bar{l}, \bar{\sigma_l}) \\
a_{\mathsf{userID[i]}} \sim \mathsf{Norm}(\bar{a}, \bar{\sigma_a}) \\
c_{\mathsf{userID[i]}} \sim \mathsf{Norm}(\bar{c}, \bar{\sigma_c}) \\
i1control_{\mathsf{userID[i]}} \sim \mathsf{Norm}(i1controlOverall, \bar{\sigma_{i1}}) \\
i1emp_{\mathsf{userID[i]}} \sim \mathsf{Norm}(i1empOverall, \bar{\sigma_{i1}}) \\
i1norm_{\mathsf{userID[i]}} \sim \mathsf{Norm}(i1normOverall, \bar{\sigma_{i1}}) \\
i1controlOverall \sim \mathsf{Norm}(0, .2) \\
i1empOverall \sim \mathsf{Norm}(0, .2) \\
i1normOverall \sim \mathsf{Norm}(0, .2) \\
\bar{\lambda} \sim \mathsf{Norm}(.00001, 2.5) \\
\bar{\sigma_l} \sim \mathsf{Exp}(1.5) \\
\bar{a} \sim \mathsf{Norm}(0, .2) \\
\bar{\sigma_a} \sim \mathsf{Exp}(5) \\
\bar{c} \sim \mathsf{Norm}(0, .2) \\
\bar{\sigma_c} \sim \mathsf{Exp}(5) \\
\bar{\sigma_{i1}} \sim \mathsf{Exp}(5)
\end{gathered}
$$

This might seem somewhat confusing, so let us disentangle this maze:

- Each user has their own baseline aggression level, $l_{\mathsf{userID[i]}}$.
- However, these individual aggression levels are not disconnected, they come from a distribution themselves, $\mathsf{Norm}(\bar{l}, \bar{\sigma_l})$. $\bar{l}$ is the mean baseline aggression level for the whole population, and $\bar{\sigma_l}$ is the standard deviation of this distribution. These general parameters are to be estimated along with the individual ones.
- Then there are individual auto regression coefficients $a_{\mathsf{userID[i]}}$, which capture the correlation between yesterday's attacks with today's attacks, so to speak. These also come from a general distribution $\mathsf{Norm}(\bar{a}, \bar{\sigma_a})$, with its own general parameters to be estimated.
- Next, there are individual user's coefficients connecting the user's activity on a given day with their aggression on the same day, $c_{\mathsf{userID[i]}}$, all coming from a general distribution $\mathsf{Norm}(\bar{c}, \bar{\sigma_c})$ whose parameters are also to be estimated.

- For any particular treatment group, say, empathy, we have a user level coefficient $i1emp_{\mathsf{userID[i]}}$, which is activated if the user is in the empathy group (that is, we multiply by the indicator variable emp) and then applied to the number of interventions received the day before (lag 1). Similarly for the two other groups. These user-level parameters come from the distribution $\mathsf{Norm}(i1empOverall, \bar{\sigma_{i1}})$, whose parameters are to be estimated.

- Finally, prior predictive check was used to choose priors for the general coefficients.

For the impact of the cumulative number of interventions (we will only use lag 3 for reasons that will become clear), since the range of values of the predictor is wider, for computational feasibility we further needed to restrict coefficients to lie between -3 and 2, but these values are not plausible values of the parameters anyway ($exp(-3) \approx 0.04$ and $exp(2) \approx 7.38$). The model for the cumulative impact was tested with and without interaction with overall aggression in the before period, without the use of attacks on a given day (as already explained, to avoid the post-treatment bias). The two relevant options are:

$$\begin{gathered} log(\lambda_i) = l_{\mathsf{userID[i]}} + c_{\mathsf{userID[i]}} \times \mathsf{act} + ic3control_{\mathsf{userID[i]}} \times \mathsf{control} \times \mathsf{intCL3D} + \\ + ic3emp_{\mathsf{userID[i]}} \times \mathsf{emp} \times \mathsf{intCL3D} + \\ ic3norm_{\mathsf{userID[i]}} \times \mathsf{norm} \times \mathsf{intCL3D} \end{gathered}$$

$$\begin{gathered} log(\lambda_i) = l_{\mathsf{userID[i]}} + c_{\mathsf{userID[i]}} \times act + \\ + ic3control_{\mathsf{userID[i]}} \times \mathsf{control} \times \mathsf{intCL3D} + \\ + icabst3control_{\mathsf{userID[i]}} \times \mathsf{control} \times \mathsf{intCL3D} \times \mathsf{abst} + \end{gathered}$$

$$\begin{gathered} log(\lambda_i) = l_{\mathsf{userID[i]}} + c_{\mathsf{userID[i]}} \times \mathsf{act} + ic3control_{\mathsf{userID[i]}} \times \mathsf{control} \times \mathsf{intCL3D} + \\ + ic3emp_{\mathsf{userID[i]}} \times \mathsf{emp} \times \mathsf{intCL3D} + \\ + icabst3emp_{\mathsf{userID[i]}} \times \mathsf{emp} \times \mathsf{intCL3D} \times \mathsf{abst} + \\ + ic3norm_{\mathsf{userID[i]}}] \times \mathsf{norm} \times \mathsf{intCL3D} + \\ + icabst3norm_{\mathsf{userID[i]}} \times \mathsf{norm} \times \mathsf{intCL3D} \times \mathsf{abst} \end{gathered}$$

The models employing the second formula were superior in performance. It is not surprising that once attacks on a given day were removed from predictor, the overall aggression levels in the before period became predictive. The price to pay, however, is that now to obtain a user-specific multiplicative interpretation of the impact of cumulative interventions, we need to put the two elements together while multiplying one by the user's overall aggression and only then exponentiate, that is we need to inspect, for instance,

$exp(ic3emp_{\mathsf{userID[i]}} + icabst3emp_{\mathsf{userID[i]}} \times \mathsf{abst}[i])$, instead of simply looking at $exp(ic3emp_{\mathsf{userID[i]}})$.

Finally, in the before-and-after analysis, we put aside the time series element, and look at aggregated counts before and after the treatment period, thus obtaining a more of a long-term effect analysis. Moreover, this time we standardize counts, obtaining continuous variables and employing normal distribution in the likelihoods, thus also making sure the overall results are robust under a spectrum of modeling choices. We build and compared multiple additive models where the outcome variable is normally distributed around the predicted mean, which is a linear function of predictors (possibly with interactions). Our general criteria led to the model whose specification is as follows (we also selected regularizing prior parameters using prior predictive checks to avoid unreasonably narrow overall prior distributions, see the appendix for a longer explanation):

$$\begin{gathered}
\mathsf{AdiffS} \sim \mathsf{Norm}(\mu, \sigma) \\
\mu_i = \alpha + \beta_{\mathsf{ADS}}[\mathsf{group}_i] \times \mathsf{ADS} + \beta_{\mathsf{group}_i} + \beta_{\mathsf{IC}}[\mathsf{group}_i] \times \mathsf{IC} + \\
+ \beta_{\mathsf{ADSIC}} \times \mathsf{ADS} \times \mathsf{IC} + \beta_{\mathsf{CBS}}[\mathsf{group}_i] \times \mathsf{CBS} \\
\alpha \sim \mathsf{Norm}(0, .3) \\
\beta_{\mathsf{ADS}}[\mathsf{group}_i] \sim \mathsf{Norm}(0, .3) \\
\beta_{\mathsf{group}_i} \sim \mathsf{Norm}(0, .3) \\
\beta_{\mathsf{IC}}[\mathsf{group}_i] \sim \mathsf{Norm}(0, .3) \\
\beta_{\mathsf{ADSIC}} \sim \mathsf{Norm}(0, .3) \\
\beta_{\mathsf{CBS}}[\mathsf{group}_i] \sim \mathsf{Norm}(0, .3)
\end{gathered}$$

That is, we take the resulting mean to be the result of the general average (α) and the impact of the following coefficients: group-specific coefficient for ADS, group coefficient, group-specific coefficient for IC, interaction coefficient for ADS and IC, and group-specific coefficient for CBS. This is plausible *prima facie*, as which group a user belongs to might have an impact on how the number of attacks during the treatment is related to the number of attacks after, the role of the intervention count, and the role of comments before. Moreover, the levels of aggressive behavior displayed by the user during treatment might have an impact on the role played by the intervention count.

7.3. Results

7.3.1. Interventions on a given day

We built seven separate models for the impact of interventions k days ago, $1 \leq k \leq 7$. In Figure 8 we visualize the results for the three groups, with jitter based on user aggression in the before period.

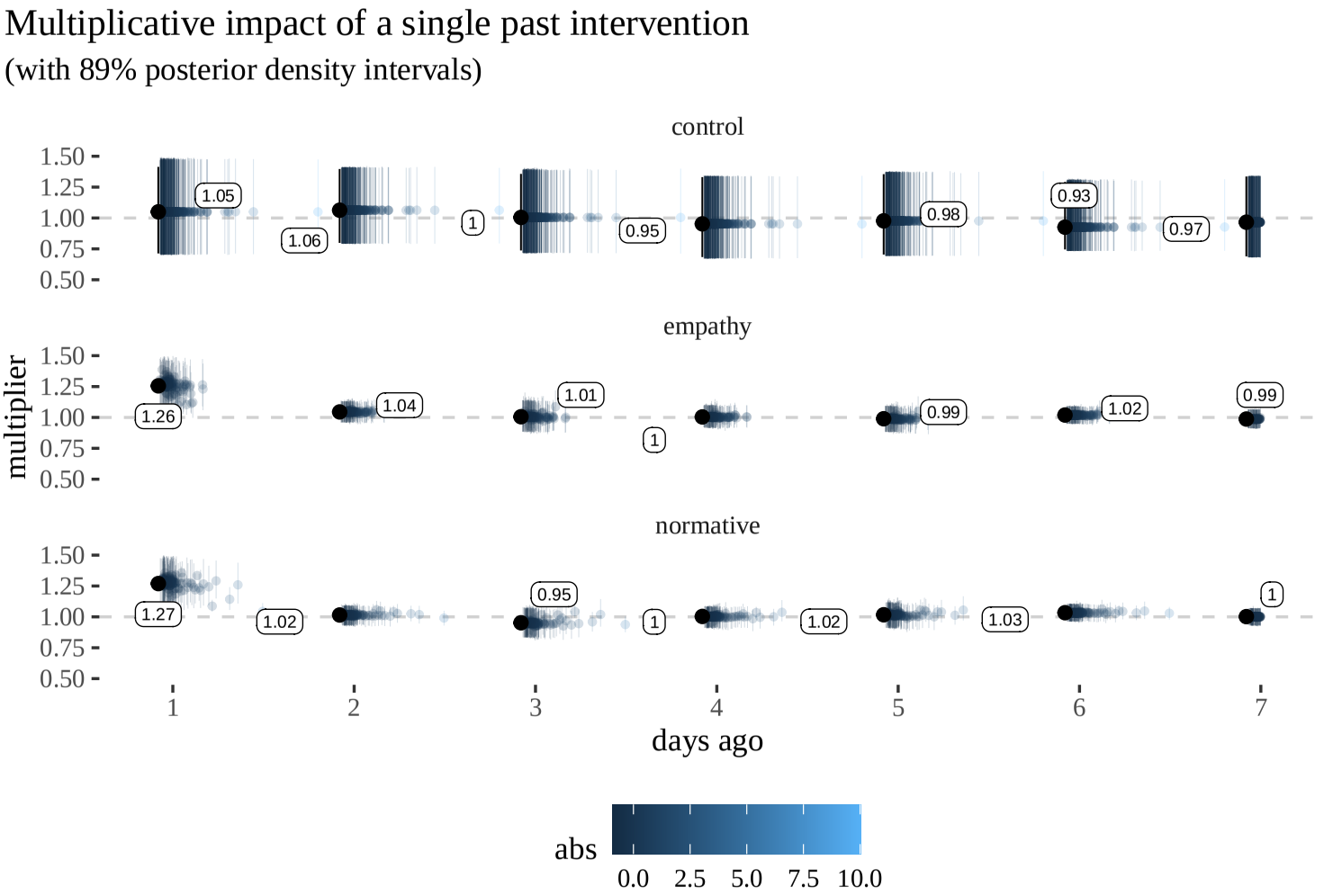

Figure 8: Impact of interventions received lag $1 \leq k \leq 7$ on attacks on a given day. High-level coefficients are pictured in black.

Notice that in short term, interventions actually increase aggression the next day (even taking the user's yesterday's aggression and today's activity in consideration). The effect, however, quickly wears off.

7.3.2. Cumulative sum of interventions

In our analysis of the effect of the cumulative number of interventions received so far, however, we intend separate this short-term effect from the long-term effect. To achieve this, we lag the cumulative interventions variable by 3, so that we're giving the user the minimal number of days needed for the short-term effect to wane. The individual users' multiplicative impact coefficients are visualized in Figure 9.

The effectiveness of normative interventions seems overall higher, except for low-aggression users, for which empathetic interventions might be equally or more useful. Importantly, linear extrapolation to extreme values might be misleading, so let us inspect on what happens with the general level multiplicative coefficients at the levels of aggression which are actually quite common, that is, at the 1st, 2nd and 3rd quartile (with respect to abst). This indicates that for the bulk of the sample the impact of cumulative interventions has been negative, slightly more so on users with lower aggression levels.

7.3.3. Long term before/after analysis

The general problem with interpreting models of this complexity involving interaction is that coefficients are not directly interpretable. For this reason, it is better to plot predicted effects for various combinations of predictors. In the construction of the plots we rely on the following:

- The values ADS range from -.67 to 10, with approximately 30% below -.5, around 80% below .3, and around 95% below 1.7, so we use these three settings of this variable in our visualizations.
- The values CBS range from -.82 to 18.3, with approximately 30% below -.4, around 80% below .3, and around 95% below 1.3, so we use these three settings of this variable in our visualizations.

Grouped before-and-after predicted change of attacks by the levels we just listed are visualized in Figure 11. For more clarity, let's inspect predicted contrasts, here understood as distances from the control group mean, by

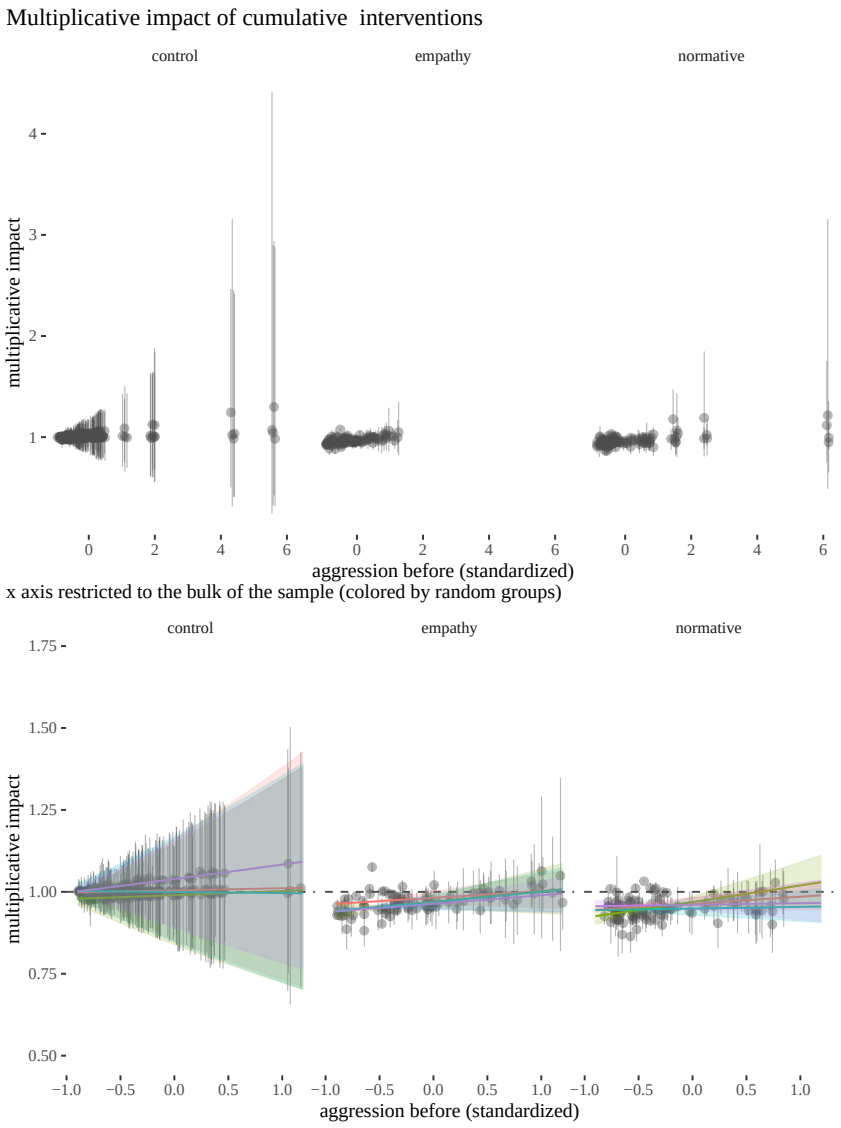

Figure 9: Multiplicative impact of cumulative interventions lag 3 on attacks. Individual users' coefficents only, full range (top), and with attention restricted to the bulk of the sample. Sub-sample coefficients depend on aggression and are represented as lines, colored by sub-sample. Note low number and high uncertainty for highly aggressive users, which motivate the restriction of the x axis for inspection.

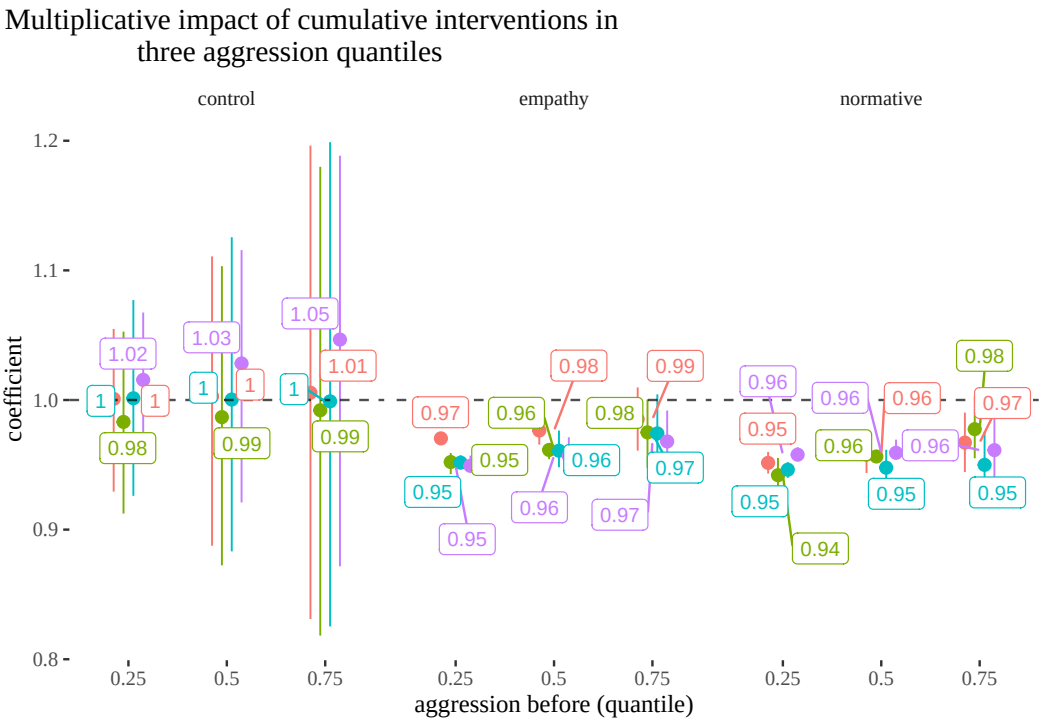

Figure 10: Multiplicative impact of cumulative interventions lag 3 on attacks. General level coefficents only, in three quantiles (.25, .5, .75). Colored by sub-sample.

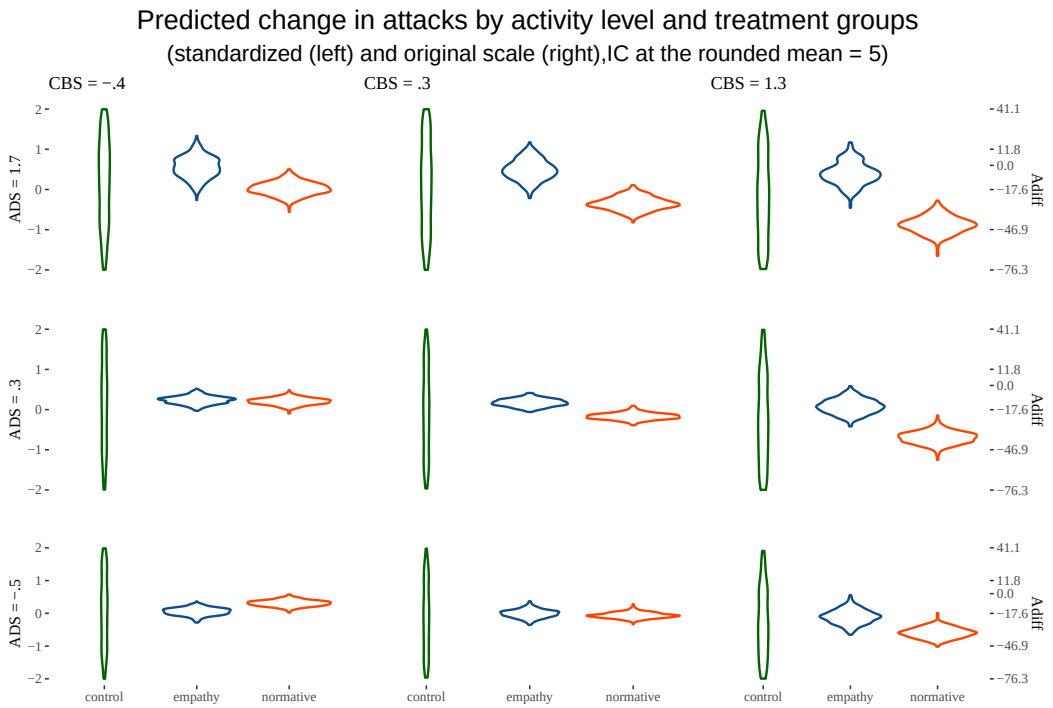

Figure 11: Predicted change in attacks, depending on user's activity level (CBS: comments before, standardized) and how aggressive overall they were (ADS: attacks during, standardized). The more aggressive and active the users, the higher the attacks drop in the normative group, slight drop correlated with emotive interventions for not too active users.

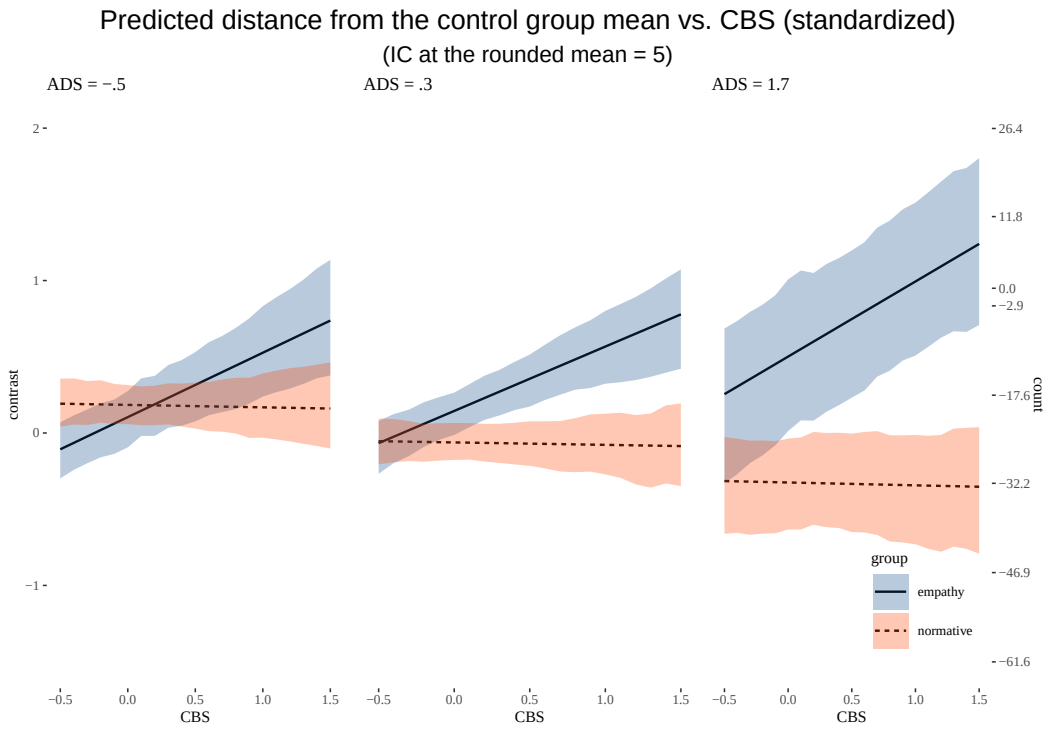

Figure 12: Predicted contrasts (difference in attacks as compared to the control group) for the two treatment groups vs activity before the treatment. Notice that empathetic interventions correlated with decreased attacks for less active users, but performed worse than normative interventions for more active users. Normative interventions, in contrast, seem to have better impact on more active users.

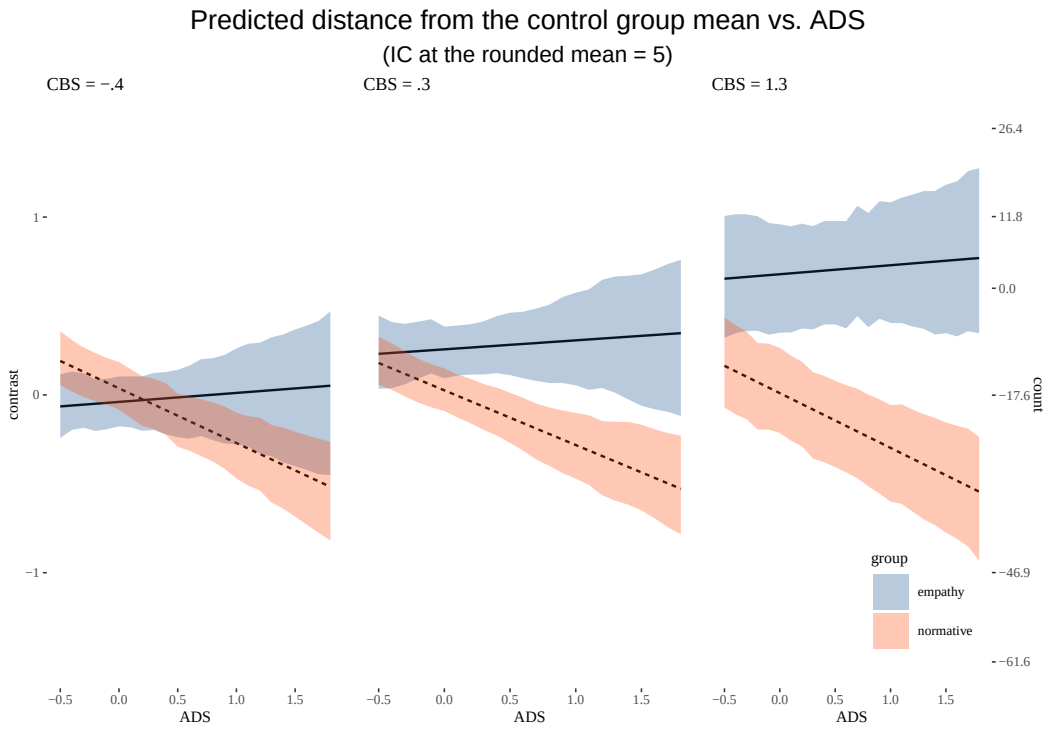

Figure 13: Predicted contrasts (difference in attacks as compared to the control group) for the two treatment groups vs aggression during the treatment period. Notice that empathetic interventions correlated with decreased attacks for less aggressive users, but performed worse than normative interventions for more aggressive users. Normative interventions, in contrast, seem to have better impact on more aggressive users.

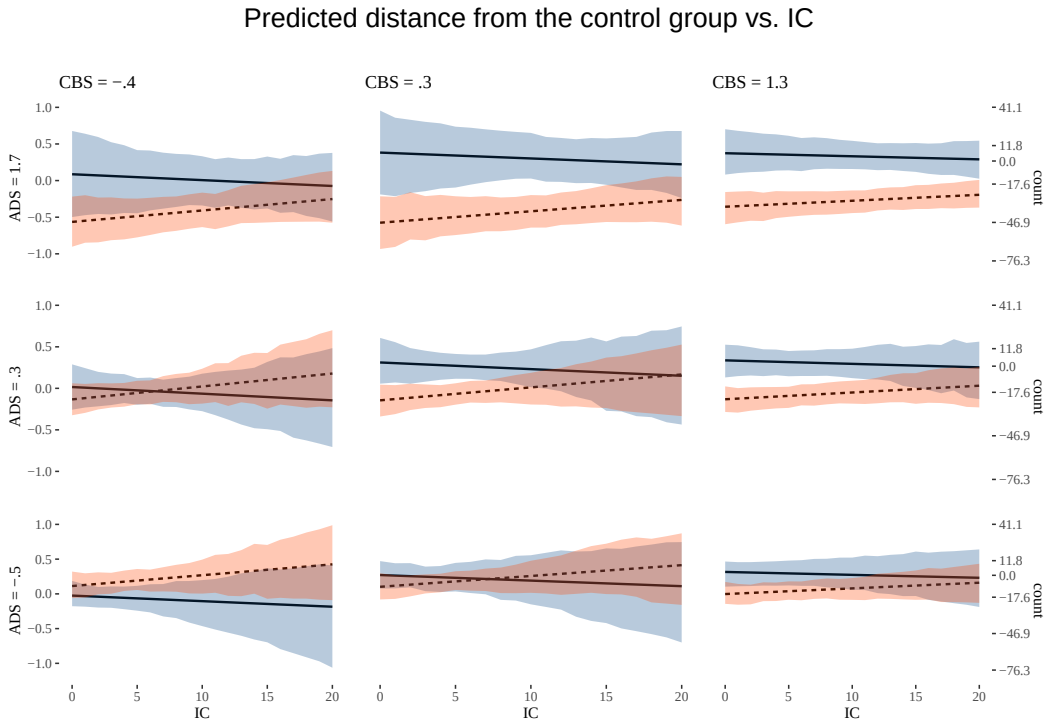

Figure 14: Contrasts (change in attacks as compared to the control group) vs the number of interventions received. Note that repeating empathetic interventions correlates with decreased attacks, while repeating normative interventions is counterproductive.

activity level, first versus CBS (comments before, standardized, Figure 12), then versus ADS (attacks during, standardized, Figure 13).

Now, let's inspect the impact of intervention counts by treatment type by looking at contrasts (distances from the control group mean) with 89% HPDIs by IC (intervention count). Notice the predicted effect of IC is weaker than group membership, so for visibility the y-axis has a smaller range. Also, not enough data was available to reliably estimate uncertainty for IC above 20, hence the restriction on the x-axis (already at lower values, lack of estimates is visible for the more extreme settings).

8. Managing volunteers

For two months of the treatment period, interventions were conducted by 19 volunteers, recruited online on an ongoing basis. The call for volunteers

was promoted on social media channels, and groups related to volunteering and sent out to 30+ NGOs and organizations with social impact, 10 out of which shared it via their social networks or distributed it among their employees and partners.

We expected that each volunteer would conduct around 10 interventions a day. Unfortunately, only a few people got engaged to such an extent at the beginning. Thus, we have implemented various incentives and gamification to increase volunteer engagement. Those included leaderboards, motivational statistics, and week-long contests with a prize. As you can see in Fig 15, the drop in volunteer engagement is noticeable within the first 2 weeks from the beginning of the treatment period. The introduction of the first competition with an economic incentive (Amazon Gift Card for the most active volunteers) increased the engagement only temporarily. It then dropped dramatically outside of the contest. This forced us to keep repeating contests regularly until the end of the treatment period. Obviously, this is not an optimal and sustainable long-term strategy.

Figure 15 16 visualizes the result of a Bayesian model of volunteer engagement based on the resulting observational data, with daily baselines and multipliers for enthusiasm and impact of competitions. Four distinct groups emerge from this picture. Those, who are active, and even more so during competitions. Those who are evenly active throughout the treatment period. Those, who are active only during competitions, and volunteers whose initial enthusiasm died completely very soon. The multiplier of the volunteer engagement during the competition is only 1.31 which means that their engagement increases by around 30% during that time. But the enthusiasm of some of the volunteers had its drop multiplier even below .9 **per day of experiment**. It accumulates as time progresses, which means that competitions might not be effective in volunteer activity resuscitation, especially in the long run.

If we were to conduct a similar experiment with volunteers again, we would test non-economic motivational schemes—for instance, creating an obligation by signing an agreement with volunteers on conducting at least 10 interventions daily, or creating more engaging content and training during the treatment period.

Besides the issue connected to low volunteer engagement, we have experienced cases in which people reported negative effects of interventions or even started to get aggressive while responding to attackers. Although we've taken extra measures to support volunteers (a psychologist available to all

participants, and distribution of and guide to their well-being), this aspect cannot be eliminated completely.

Additionally, many volunteers started to intervene in a repetitive manner, which resulted in accusations of behaving like an automated account (bot). During the process, we prepared an additional guide for our volunteers "How to write Powerful Interventions" and rewarded creativity during the contests. Both led to a noticeable increase in creativity, especially when it comes to the most engaged volunteers.

We need to mention at this point that in this experiment, the negative side of the collective intelligence approach outweighs the positive one. In our future endeavors, we will be further exploring the opportunity to employ bots to conduct counter-speech. Even though AI is not endowed with empathy, it can be a powerful ally in evoking empathetic responses and maintaining civil behavior online. An ally, who under the condition of being precise enough to act autonomously, can take the burden off of the shoulders of moderators and the community, not requiring them to screen through a multitude of potentially aggressive comments.

Another scenario worth considering is a collective intelligence approach in which 90% of interventions are conducted by an AI while 10% of the most difficult cases are taken care of by humans, trained professionals, such as moderators. During the study, we had one extraordinary volunteer who was engaged and creative to such an extent that no machine could compete. Collaborating with people like them would considerably optimize the process. Such a trained, engaged, and resilient person could supervise the interventions run by AI and deal with extreme cases.

9. Discussion

10. Volunteer engagement and impact of competitions

10.1. The challenge of keeping volunteers engaged

10.2. Volunteer activity data analysis

The winning model, given our model selection method, is specified as follows:

$$
\begin{gathered}
\mathsf{interventions} \sim \mathsf{NegativeBinomial}(\lambda, \phi) \\
log(\lambda) = l_{\mathsf{volunteerID[i]}} + enth_{\mathsf{volunteerID[i]}} \times \mathsf{daysOfProject} + comp_{\mathsf{volunteerID[i]}} \times \mathsf{competition} \\
l_{\mathsf{volunteerID[i]}} \sim \mathsf{Norm}(lbar, lsigmabar) \\
lbar \sim \mathsf{Norm}(2, .9) \\
lsigmabar, enthsigmabar, compsigmabar \sim \mathsf{Exp}(.5) \\
enth_{\mathsf{volunteerID[i]}} \sim \mathsf{Norm}(enthbar, enthsigmabar) \\
comp_{\mathsf{volunteerID[i]}} \sim \mathsf{Norm}(compbar, compsigmabar) \\
enthbar, compbar \sim \mathsf{Norm}(0, .3) \\
\phi = puser_{\mathsf{volunteerID[i]}} \\
puser_{\mathsf{volunteerID[i]}} \sim \mathsf{Exp}(1)
\end{gathered}
$$

Intuitively, volunteer interventions are assumed to have negative binomial distribution around their own expected value λ and individualized dispersion parameters ϕ. On each day each a user has their own daily expected value, which is determined by the following factors:

- First, there's user's individual baseline activity for the whole treatment period, $l_{\mathsf{volunteerID[i]}}$.
- next, each user has their own dispersion parameter, $puser_{\mathsf{volunteerID[i]}}$.
- then, there is (usually dwindling) enthusiasm: the impact of time on that user, $enth_{\mathsf{volunteerID[i]}}$ to be (after exponentiation) multiplied by the number of days that have passed since the experiment started,
- finally, we have the impact that the presence of competitions made on a user, $comp_{\mathsf{volunteerID[i]}}$, which (after exponentiation) becomes the activity multiplier to be applied during competitions only.

Moreover, the model is hierarchical: the individual level parameters are drawn from distributions whose parameters are in turn to be estimated as well. Thus, *lbar* is the overall baseline for the whole group, *enthbar* is the overall estimated group enthusiasm coefficient, and *compbar* is the overall estimated competition impact coefficient (all of them come with their own nuisance sigma parameters).
All of these parameters are given priors in a manner analogous to the introduction of priors for the other time series models, as explained in the appendix.[11]

[11]Interestingly, if we are interested in the causal effect of competitions, we should not

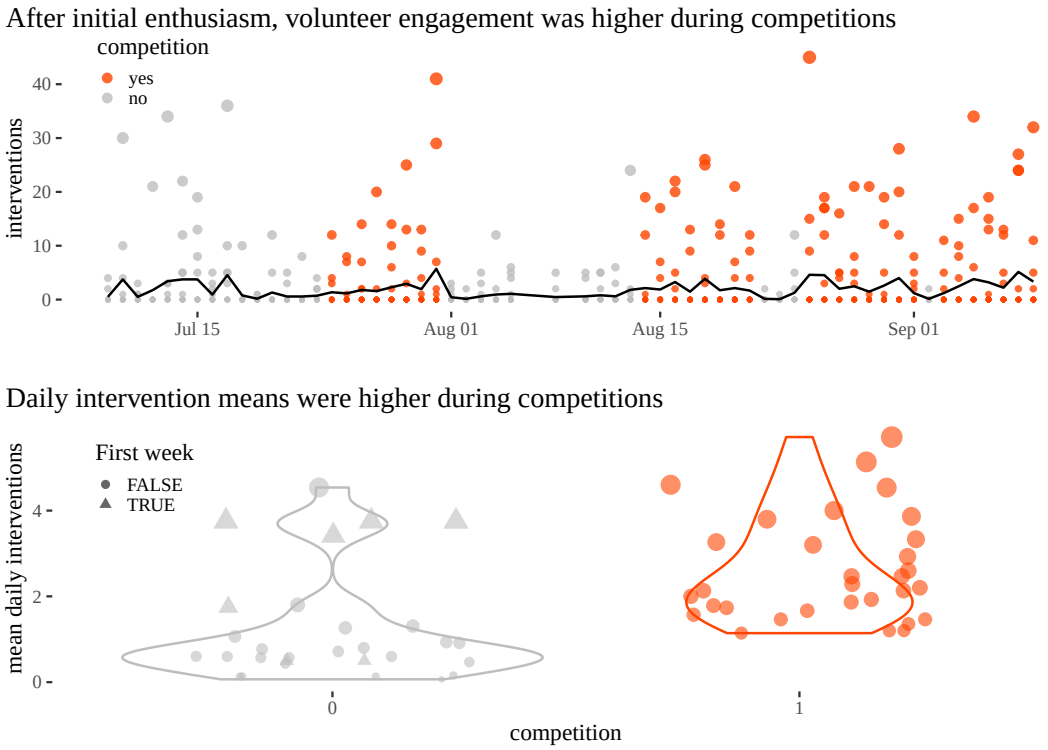

Figure 15: Daily individual voilunteer intervention counts accross time with competition periods marked (top) and daily group intervention means grouped by whether a competition was ongoing (bottom). Note most of high means in the non-competition period are in the first week.

Raw data and daily means are illustrated in Figure 15, and the individualized totals with the key coefficients based on the trained model are illustrated in Figure 16.

use an auto-regressive predictor. If we auto-regress on a lag in the $[1, 7]$ range, for some days we will be conditioning on interventions conducted during the same competition, which will already contain some information about the impact of that competition. In other words, auto-regression with short lags would lead to post-treatment bias. On the other hand, auto-regression with longer lags would either lead to dropping a lot of data in the beginning (where lagged information is not available), or degenerate the analysis by using 0s for missing lagged values in a long initial period. All this without much gain, as we have already inspected null models with auto-regression with large lags and they do not lead to performance improvement.

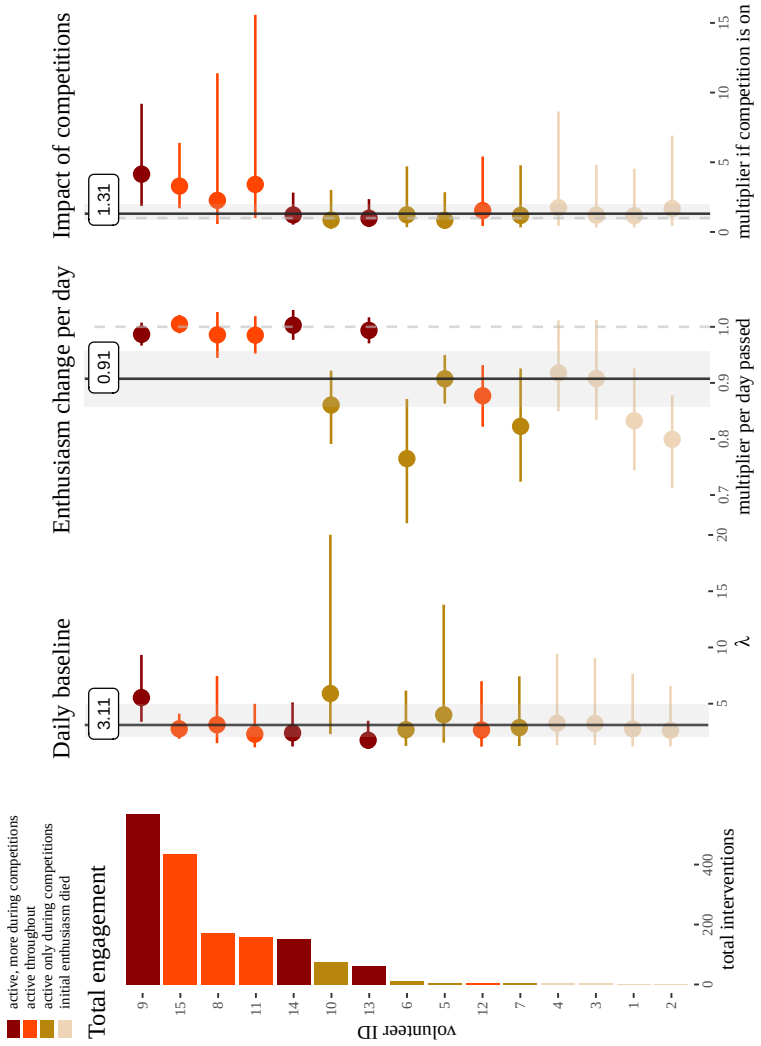

Figure 16: Volunteer total engagement with their daily baseline and multipliers for enthusiasm and impact of competition. Pointranges represent individual level coefficients, group coefficients are represented by black lines with shaded 89% HPDI areas.

Appendix A. Explanation of WAIC

Let y be the observations and Θ a posterior distribution. First, log-pointwise-predictive-density is defined by:

$$\mathsf{lppd}(y, \Theta) = \sum_i log \frac{1}{S} \sum_s p(y_i|\Theta_s)$$

where S is the number of samples in the posterior, and Θ_s is the s-th combination of sampled parameter values in the posterior distribution. That is, for each observation and each combination of parameters in the posterior we first compute its density, then we take the average density of that observation over all combinations of parameters in the posterior, and then take the logarithm. Finally, we sum these values up for all the observations. Crucially, when comparing posterior distributions with respect to the same dataset, lppds are proportional to unbiased estimates of their divergence from the real distribution (note that it is *only* proportional, and for this reason can be used for comparison of distributions only and makes no intuitive sense on its own). However, lppd always improves as the model gets more complex, so for model comparison it makes more sense to use the Widely Applicable Information Criterion (WAIC), which is an approximation of the out-of-sample deviance that converges to the cross-validation approximation in a large sample. It is defined as the log-posterior-predictive-density with an additional penalty proportional to the variance in the posterior predictions:

$$\mathsf{WAIC}(\mathsf{y}, \Theta) = -2(\mathsf{lppd} - \overbrace{\sum_i var_\theta \mathsf{log} p(y_i|\theta)}^{\mathsf{penalty}})$$

Thus to construct the penalty, we calculate the variance in log-probabilities for each observation and sum them up. Because of the analogy to Akaike's criterion, the penalty is sometimes called the effective number of parameters, p_{WAIC}. How does WAIC compare to other information criteria? AIC uses MAP estimates instead of the posterior and requires that priors be flat or overwhelmed by the likelihood, and assumes that the posterior distribution is approximately multivariate Gaussian and the sample size is much greater than the number of parameters used in the model. Bayesian Information Criterion (BIC) also requires flat priors and uses MAP estimates. WAIC does not make these assumptions, and provides almost exactly the same results as AIC, when AIC's assumptions are met.

Appendix B. Time series model selection

Suppose we are interested in the impact of interventions received n days ago. We started with a simple null model that uses the Poisson distribution, with either uses a single λ for all the users, or user-specific λs. The first Bayesian model has the following structure:

$$\begin{gathered}\mathsf{attacks_i} \sim \mathsf{Poisson}(\lambda) \\ log(\lambda) = l \\ l \sim \mathsf{Norm}(.05, 2.8)\end{gathered}$$

and the user-specific coefficient model had the following structure:

$$\begin{gathered}\mathsf{attacks}_i \sim \mathsf{Poisson}(\lambda_i) \\ log(\lambda_i) = l_{\mathsf{userID[i]}} \\ l_{\mathsf{userID[i]}} \sim \mathsf{Norm}(.05, 2.8)\end{gathered}$$

The priors were chosen using prior predictive check, so that the 89% density intervals reached between 0 and 34, with median around 1. Given our prior experience with similar user datasets this is a fairly wide informative prior. The comparison, unsurprisingly, preferred the user-specific λs.

Next, we introduced the auto-regressive element, conditioning on yesterday's attacks. The choice of priors for the auto-regression coefficient is guided by the visualization (intuitive direct understanding of the values is made difficult by the fact that the predictors work on the logarithmic scale) and the fact that larger values would result in a unreasonably extreme impact of yesterday's attacks.

$$\begin{gathered}\mathsf{attacks}_i \sim \mathsf{Poisson}(\lambda_i) \\ log(\lambda_i) = l_{\mathsf{userID[i]}} + a_{\mathsf{userID[i]}} \times \mathsf{attacksL1} \\ l_{\mathsf{userID[i]}} \sim \mathsf{Norm}(.05, 2.8) \\ a_{\mathsf{userID[i]}} \sim \mathsf{Norm}(0, .2)\end{gathered}$$

Next, we added today's activity level as a predictor, with user-specific coefficients. Adding activity levels helps. Note also that our priors taken separately were made more narrow, to preserve the overall width of the prior predictive distribution (this will be the usual strategy as we progress).

$$\begin{gathered}\mathsf{attacks}_i \sim \mathsf{Poisson}(\lambda_i) \\ log(\lambda_i) = l_{\mathsf{userID[i]}} + a_{\mathsf{userID[i]}} \times \mathsf{attacksL1} + c \times \mathsf{act} \\ l_{\mathsf{userID[i]}} \sim \mathsf{Norm}(.05, 2.3) \\ a_{\mathsf{userID[i]}} \sim \mathsf{Norm}(0, .1) \\ c \sim \mathsf{Norm}(0, .1)\end{gathered}$$

Unsurprisingly, it helps even more if the coefficients are user-specific:

$$\begin{gathered}
\mathsf{attacks}_i \sim \mathsf{Poisson}(\lambda_i) \\
log(\lambda_i) = l_{\mathsf{userID[i]}} + a_{\mathsf{userID[i]}} \times \mathsf{attacksL1} + c_{\mathsf{userID[i]}} \times \mathsf{act} \\
l_{\mathsf{userID[i]}} \sim \mathsf{Norm}(.05, 2.3) \\
a_{\mathsf{userID[i]}} \sim \mathsf{Norm}(0, .1) \\
c_{\mathsf{userID[i]}} \sim \mathsf{Norm}(0, .1)
\end{gathered}$$

A relatively large number of zeros suggests that moving to a zero-inflated Poisson distribution would be a good idea. It was not, so the following model structure was tested and abandoned:

$$\begin{gathered}
\mathsf{attacks}_i \sim \mathsf{ZiPoisson}(p, \lambda_i) \\
log(\lambda_i) = l_{\mathsf{userID[i]}} + a_{\mathsf{userID[i]}} \times \mathsf{attacksL1} + c_{\mathsf{userID[i]}} \times \mathsf{act} \\
l_{\mathsf{userID[i]}} \sim \mathsf{Norm}(.05, 2.3) \\
a_{\mathsf{userID[i]}} \sim \mathsf{Norm}(0, .1) \\
c_{\mathsf{userID[i]}} \sim \mathsf{Norm}(0, .1) \\
logit(p) = \pi \\
\pi \sim Norm(-1.5, 1)
\end{gathered}$$

Then we considered the negative binomial distribution, and the addition of week days as a predictor (both with general and user-level coefficients). While moving to the negative binomial distribution resulted in an improvement, adding week days did not improve the model performance, perhaps because we already conditioned on activity, and whatever the impact of weekdays was, has been already mediated through activity (in a sense, we committed a post-treatment bias with respect to weekdays; but that's fine, we did not really care about the impact of weekdays).

$$\begin{gathered}
\mathsf{attacks}_i \sim \mathsf{NegativeBinomial}(\lambda_i, \phi) \\
log(\lambda_i) = l_{\mathsf{userID[i]}} + a_{\mathsf{userID[i]}} \times \mathsf{attacksL1} + c_{\mathsf{userID[i]}} \times \mathsf{act} \\
l_{\mathsf{userID[i]}} \sim \mathsf{Norm}(.05, 2.3) \\
a_{\mathsf{userID[i]}} \sim \mathsf{Norm}(0, .1) \\
c_{\mathsf{userID[i]}} \sim \mathsf{Norm}(0, .1) \\
\phi \sim \mathsf{Exp}(1)
\end{gathered}$$

$$\begin{gathered}
\mathsf{attacks}_i \sim \mathsf{NegativeBinomial}(\lambda_i, \phi) \\
log(\lambda_i) = l_{\mathsf{userID[i]}} + a_{\mathsf{userID[i]}} \times \mathsf{attacksL1} + c_{\mathsf{userID[i]}} \times \mathsf{act} + w \times \mathsf{weekday} \\
l_{\mathsf{userID[i]}} \sim \mathsf{Norm}(.05, 2.3) \\
a_{\mathsf{userID[i]}} \sim \mathsf{Norm}(0, .1) \\
c_{\mathsf{userID[i]}} \sim \mathsf{Norm}(0, .1) \\
w \sim \mathsf{Norm}(0, .1) \\
\phi \sim \mathsf{Exp}(1)
\end{gathered}$$

$$\begin{gathered}
\mathsf{attacks}_i \sim \mathsf{NegativeBinomial}(\lambda_i, \phi) \\
log(\lambda_i) = l_{\mathsf{userID[i]}} + a_{\mathsf{userID[i]}} \times \mathsf{attacksL1} + c_{\mathsf{userID[i]}} \times \mathsf{act} + w_{\mathsf{userID[i]}} \times \mathsf{weekday} \\
l_{\mathsf{userID[i]}} \sim \mathsf{Norm}(.05, 2.3) \\
a_{\mathsf{userID[i]}} \sim \mathsf{Norm}(0, .1) \\
c_{\mathsf{userID[i]}} \sim \mathsf{Norm}(0, .1) \\
w_{\mathsf{userID[i]}} \sim \mathsf{Norm}(0, .1) \\
\phi \sim \mathsf{Exp}(1)
\end{gathered}$$

We also considered adding overall aggression in before period as a predictor, but the addition did not lead to improvement. One reason this is interesting is that interaction of interventions with overall aggression will turn out to be important for long-term effects.

$$\begin{gathered}
\mathsf{attacks}_i \sim \mathsf{NegativeBinomial}(\lambda_i, \phi) \\
log(\lambda_i) = l_{\mathsf{userID[i]}} + a_{\mathsf{userID[i]}} \times \mathsf{attacksL1} + c_{\mathsf{userID[i]}} \times \mathsf{act} + act + ab_{\mathsf{userID[i]}} \times \mathsf{ABS} \\
l_{\mathsf{userID[i]}} \sim \mathsf{Norm}(.05, 2.3) \\
a_{\mathsf{userID[i]}} \sim \mathsf{Norm}(0, .1) \\
c_{\mathsf{userID[i]}} \sim \mathsf{Norm}(0, .1) \\
ab_{\mathsf{userID[i]}} \sim \mathsf{Norm}(0, .1) \\
\phi \sim \mathsf{Exp}(1)
\end{gathered}$$

So, ultimately, the negative binomial model without week days or aggression before became our null model to which we considered adding intervention count and intervention types as predictors. For now, consider intervention type and interventions received with lag 1 (note that if, for instance, we are interested in the impact of interventions lag 2, we cannot condition on interventions lag 1, as this would lead to post-treatment bias). So what we will say about lag 1 will be exactly mirrored in the models for other lag values.

Adding intervention count, and adding intervention count with distinguishing intervention types resulted in improvements.

$$\begin{gathered}
\mathsf{attacks}_i \sim \mathsf{NegativeBinomial}(\lambda_i, \phi) \\
log(\lambda_i) = l_{\mathsf{userID[i]}} + a_{\mathsf{userID[i]}} \times \mathsf{attacksL1} + c_{\mathsf{userID[i]}} \times \mathsf{act} + i1 \times \mathsf{intL1D} \\
l_{\mathsf{userID[i]}} \sim \mathsf{Norm}(.05, 2.3) \\
a_{\mathsf{userID[i]}} \sim \mathsf{Norm}(0, .1) \\
c_{\mathsf{userID[i]}} \sim \mathsf{Norm}(0, .1) \\
i1 \sim \mathsf{Norm}(0, .1) \\
\phi \sim \mathsf{Exp}(1)
\end{gathered}$$

$$
\begin{gathered}
\mathsf{attacks}_i \sim \mathsf{NegativeBinomial}(\lambda_i, \phi) \\
log(\lambda_i) = l_{\mathsf{userID[i]}} + a_{\mathsf{userID[i]}} \times \mathsf{attacksL1} + c_{\mathsf{userID[i]}} \times \mathsf{act} + i1_{\mathsf{type[i]}} \times \mathsf{intL1D} \\
l_{\mathsf{userID[i]}} \sim \mathsf{Norm}(.05, 2.3) \\
a_{\mathsf{userID[i]}} \sim \mathsf{Norm}(0, .1) \\
c_{\mathsf{userID[i]}} \sim \mathsf{Norm}(0, .1) \\
i1_{\mathsf{type[i]}} \sim \mathsf{Norm}(0, .1) \\
\phi \sim \mathsf{Exp}(1)
\end{gathered}
$$

Taking ϕ parameters to be user-relative also resulted in improvement:

$$
\begin{gathered}
\mathsf{attacks}_i \sim \mathsf{NegativeBinomial}(\lambda_i, \phi_{\mathsf{userID[i]}}) \\
log(\lambda_i) = l_{\mathsf{userID[i]}} + a_{\mathsf{userID[i]}} \times \mathsf{attacksL1} + c_{\mathsf{userID[i]}} \times \mathsf{act} + i1 \times \mathsf{intL1D} \\
l_{\mathsf{userID[i]}} \sim \mathsf{Norm}(.05, 2.3) \\
a_{\mathsf{userID[i]}} \sim \mathsf{Norm}(0, .1) \\
c_{\mathsf{userID[i]}} \sim \mathsf{Norm}(0, .1) \\
i1_{\mathsf{type[i]}} \sim \mathsf{Norm}(0, .1) \\
\phi_{\mathsf{userID[i]}} \sim \mathsf{Exp}(1)
\end{gathered}
$$

Finally, we made a crucial move to deploy hierarchical modeling. The general idea is that while we do keep user-specific coefficients wherever we had them, we also do not assume that they are independent, but rather that they come from their respective distributions, and we estimate the general features of those distributions at the same time. Also, for convenience this time we used treatment type indicator variables.

$$
\begin{gathered}
\mathsf{attacks}_i \sim \mathsf{NegativeBinomial}(\lambda_i, \phi_{\mathsf{userID[i]}}) \\
log(\lambda_i) = l_{\mathsf{userID[i]}} + a_{\mathsf{userID[i]}} \times \mathsf{attacksL1} + c_{\mathsf{userID[i]}} \times \mathsf{act} + \\
+ i1control_{\mathsf{userID[i]}} \times \mathsf{control} \times \mathsf{intL1D} + \\
+ i1emp_{\mathsf{userID[i]}} \times \mathsf{emp} \times \mathsf{intL1D} + \\
+ i1norm_{\mathsf{userID[i]}} \times \mathsf{norm} \times \mathsf{intL1D} \\
l_{\mathsf{userID[i]}} \sim \mathsf{Norm}(\bar{l}, \bar{\sigma_l}) \\
a_{\mathsf{userID[i]}} \sim \mathsf{Norm}(\bar{a}, \bar{\sigma_a}) \\
c_{\mathsf{userID[i]}} \sim \mathsf{Norm}(\bar{c}, \bar{\sigma_c}) \\
i1control_{\mathsf{userID[i]}} \sim \mathsf{Norm}(i1controlOverall, \bar{\sigma_{i1}}) \\
i1emp_{\mathsf{userID[i]}} \sim \mathsf{Norm}(i1empOverall, \bar{\sigma_{i1}}) \\
i1norm_{\mathsf{userID[i]}} \sim \mathsf{Norm}(i1normOverall, \bar{\sigma_{i1}}) \\
i1controlOverall \sim \mathsf{Norm}(0, .2) \\
i1empOverall \sim \mathsf{Norm}(0, .2) \\
i1normOverall \sim \mathsf{Norm}(0, .2) \\
\bar{\lambda} \sim \mathsf{Norm}(.00001, 2.5) \\
\bar{\sigma_l} \sim \mathsf{Exp}(1.5) \\
\bar{a} \sim \mathsf{Norm}(0, .2) \\
\bar{\sigma_a} \sim \mathsf{Exp}(5) \\
\bar{c} \sim \mathsf{Norm}(0, .2) \\
\bar{\sigma_c} \sim \mathsf{Exp}(5) \\
\bar{\sigma_{i1}} \sim \mathsf{Exp}(5)
\end{gathered}
$$

Now, let us rethink the priors. The coefficients need to be exponentiated to be understood multiplicatively. For instance, the prior for $i1empOverall$ is $\mathsf{Norm}(0, .2)$. To understand what priors for the exponentiated individual coefficients this entails, we can simulate: (1) draw 1e4 values $i1bar$ of the mean from $\mathsf{Norm}(0, .2)$, (2) draw 1e4 values $i1sigmabar$ of the standard deviation parameter from $\mathsf{Exp}(5)$, and each time (3) draw 1e4 parameters from $\mathsf{Norm}(i1bar, i1sigmabar)$. The resulting distribution looks as in Figure B.17. This is still a very wide prior for the multiplicative impact of emphatetic interventions, centered around 1, allowing even extremely unlikely values close to 0 or 2 (upon reflection: you really should not expect a single intervention to reduce aggression to zero or to double it in everyone). In the cumulative model for computation reasons we will need to narrow down the distributions, but the general point hold: prior predictive check still ensures that they are centered around neutral values and that they allow for a very reasonable range of values.

Appendix C. Model choice for the long term analysis

Let us elaborate on how we decided to use the seemingly fairly complicated model we already described in the body of the paper. Once preliminary causal considerations guided our restrictions on variable selection, we proceed by building models of increasing complexity, and comparing them in terms of Widely Acceptable Information Criterion (which we have already discussed). The models differ mostly in the underlying linear formulae. For computational ease we will here use quadratic approximations, while in the final analysis we will deploy Hamiltionian Monte Carlo. The names are meant to decode the model structure: the predictors are listed before dashes, whereas interactions are listed after dashes. The comparison results are in Table C.2 and plotted in Figure C.18. Notice that there are ways of building a complicated models that do not result in improvement, as they rather lead to expected performance lower than that of the null model.

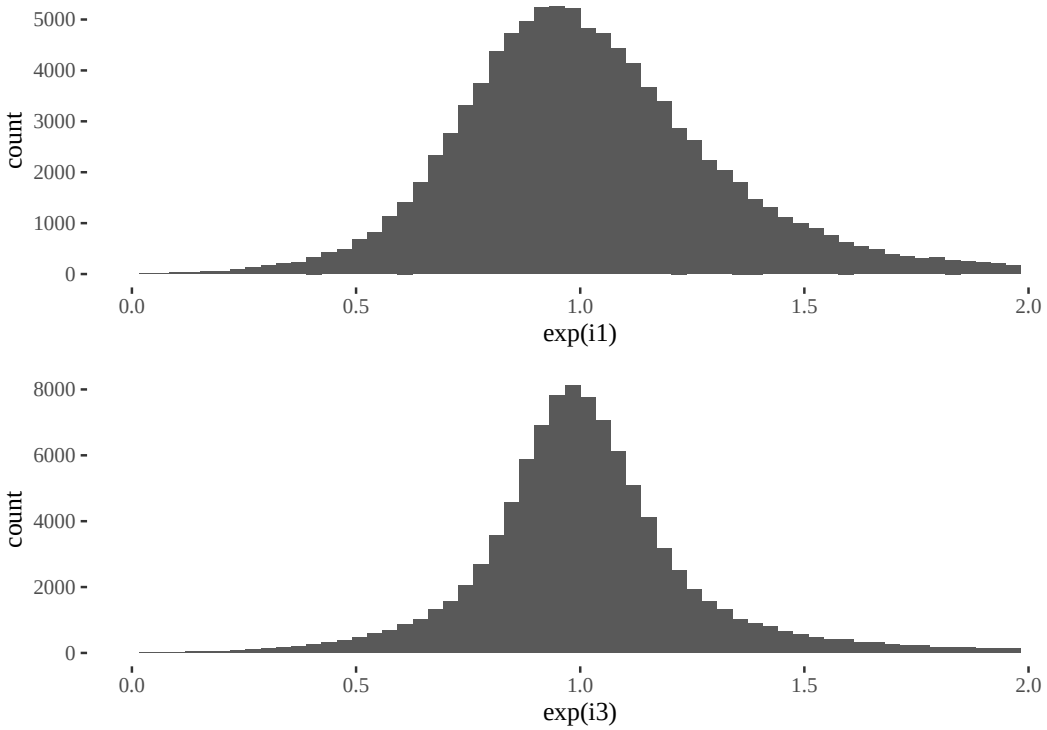

Figure B.17: Simulated priors for the individual i1 coefficients, and prior for the cumulative impact model with larger input variability (hence, the prior is more narrow to eliminate unrealistically huge impact).

	WAIC	SE	dWAIC	dSE	pWAIC	weight
Final	1184.829	89.779	0.000	NA	26.871	0.590
tooFar	1186.126	89.413	1.297	2.758	28.181	0.308
ADSITICCBS_ITIC_ADSIC	1188.337	87.058	3.508	6.184	24.822	0.102
IT	1345.087	144.443	160.259	132.802	18.104	0.000
null	1345.550	145.960	160.721	134.243	18.616	0.000
ADS	1348.696	143.821	163.867	132.558	22.718	0.000
ADSITIC_ADSIC	1351.556	152.861	166.728	139.154	29.070	0.000
ADSIT	1351.646	145.161	166.817	133.795	25.032	0.000
ADSITIC	1352.087	146.835	167.258	134.608	27.254	0.000
ADSIT_ADSIT	1352.672	155.862	167.844	142.092	31.855	0.000
ADSIC	1352.892	146.359	168.064	134.313	26.421	0.000
ADSITIC_ADSIC_ADSIT	1355.482	155.522	170.653	141.405	33.558	0.000
ADSITIC_ADSIT_ITIC_ADSIC	1355.783	155.273	170.954	141.128	33.771	0.000

Table C.2: Model comparison results.

Null	$\mu_i = \alpha$
ADS	$\mu_i = \alpha + \beta_{\mathsf{ADS}} \times \mathsf{ADS}$
ADSIC	$\mu_i = \alpha + \beta_{\mathsf{ADS}} \times \mathsf{ADS} + \beta_{\mathsf{IC}} \times \mathsf{IC}$
IT	$\mu_i = \beta_{\mathsf{group}[i]}$
ADSIT	$\mu_i = \alpha + \beta_{\mathsf{ADS}} \times \mathsf{ADS} + \beta_{\mathsf{group}[i]}$
ADSITIC	$\mu_i = \alpha + \beta_{\mathsf{ADS}} \times \mathsf{ADS} + \beta_{\mathsf{group}[i]} + \beta_{\mathsf{IC}} \times \mathsf{IC}$
ADSITIC-ADSIC	$\mu_i = \alpha + \beta_{\mathsf{ADS}} \times \mathsf{ADS} + \beta_{\mathsf{group}[i]} + \beta_{\mathsf{IC}} \times \mathsf{IC} + +\beta_{\mathsf{ADSIC}} \times \mathsf{ADS} \times \mathsf{IC}$
ADSITIC-ADSIC-ADSIT	$\mu_i = \alpha + \beta_{\mathsf{ADS}}[\mathsf{group}_i] \times \mathsf{ADS} + \beta_{\mathsf{group}[i]} + +\beta_{\mathsf{IC}} \times \mathsf{IC} + \beta_{\mathsf{ADSIC}} \times \mathsf{ADS} \times \mathsf{IC}$
ADSIT-ADSIT	$\mu_i = \alpha + \beta_{\mathsf{ADS}}[\mathsf{group}_i] \times \mathsf{ADS} + \beta_{\mathsf{group}[i]}$
ADSITIC-ADSIT-ITIC-ADSIC	$\mu_i = \alpha + \beta_{\mathsf{ADS}}[\mathsf{group}_i] \times \mathsf{ADS} + \beta_{\mathsf{group}[i]} + +\beta_{\mathsf{IC}}[\mathsf{group}_i] \times \mathsf{IC} + \beta_{\mathsf{ADSIC}} \times \mathsf{ADS} \times \mathsf{IC}$
ADSITICCBS-ITIC-ADSIC	$\mu_i = \alpha + \beta_{\mathsf{ADS}}[\mathsf{group}_i] \times \mathsf{ADS} + \beta_{\mathsf{group}[i]} + +\beta_{\mathsf{IC}}[\mathsf{group}_i] \times \mathsf{IC} + \beta_{\mathsf{CBS}} \times \mathsf{CBS} + \beta_{\mathsf{ADSIC}} \times \mathsf{ADS} \times \mathsf{IC}$
Final	$\mu_i = \alpha + \beta_{\mathsf{ADS}}[\mathsf{group}_i] \times \mathsf{ADS} + \beta_{\mathsf{group}_i} + +\beta_{\mathsf{IC}}[\mathsf{group}_i] \times \mathsf{IC} + +\beta_{\mathsf{ADSIC}} \times \mathsf{ADS} \times \mathsf{IC} + \beta_{\mathsf{CBS}}[\mathsf{group}_i] \times \mathsf{CBS}$
tooFar	$\mu_i = \alpha + \beta_{\mathsf{ADS}}[\mathsf{group}_i] \times \mathsf{ADS} + +\beta_{\mathsf{group}_i} + \beta_{\mathsf{IC}}[\mathsf{group}_i] \times \mathsf{IC} + +\beta_{\mathsf{ADSIC}} \times \mathsf{ADS} \times \mathsf{IC} + +\beta_{\mathsf{CBS}}[\mathsf{group}_i] \times \mathsf{CBS} + \beta_{\mathsf{CBSIC}} \times \mathsf{CBS} \times \mathsf{IC}$

The three models that stand out differ in including CBS as a predictor. Moreover the final model includes an interaction between treatment group and CBS. Adding a further interaction between CBS and IC takes us too far. We will employ the top model (Final) in further analyses.

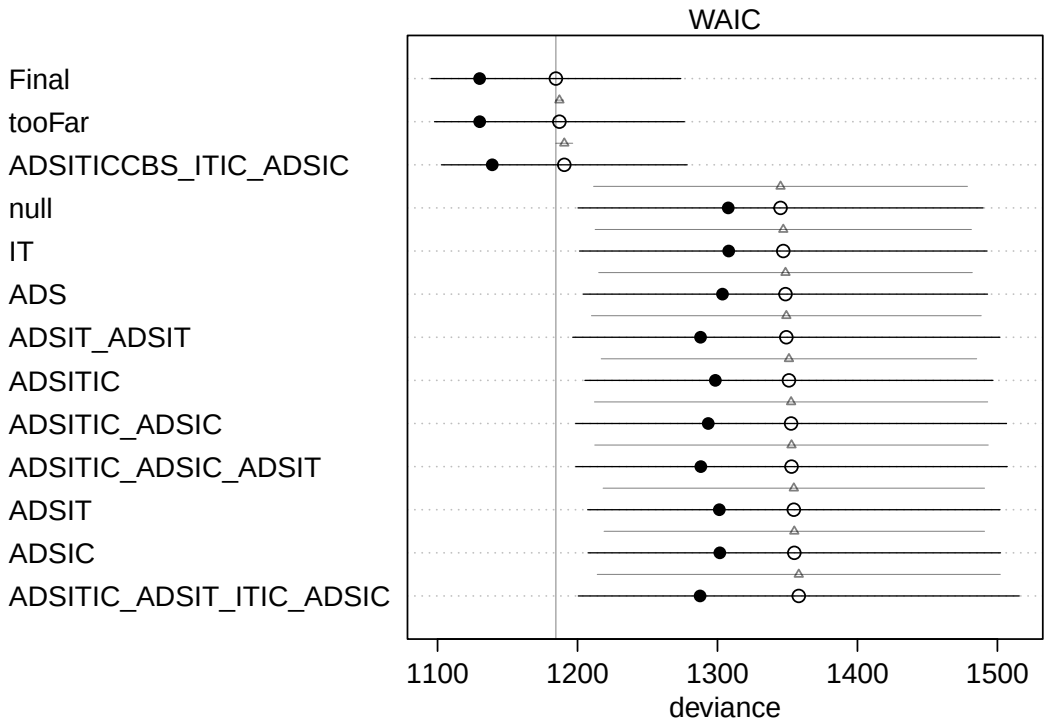

Figure C.18: Model comparison, WAIC scores. The filled points are the in-sample deviance values. The open points are the WAIC values. The line segments represent standard errors of the WAIC scores. really want however is the standard error of the difference in WAIC between the two models. The triangle is the difference to the top rated model, and the line segment going through it is the standard error of this difference.

Now, to sensibly set up our priors, let's build two models with the general structure reached. One with fairly wide priors that one might initially think are appropriate, one with regularizing priors. The key phenomenon to watch out for in such contexts (slightly complex models with interactions) is that it is hard to intuitively predict the impact of coefficient priors on prior predictions. For this reason, we run prior predictive checks for both models, and we select the priors that do not result in unrealistically wide prior predictions.

References

[1] Alhujailli, A., Karwowski, W., Wan, T.T., Hancock, P., 2020. Affective and stress consequences of cyberbullying. Symmetry 12, 1536.

[2] Álvarez-Benjumea, A., Winter, F., 2018. Normative change and culture of hate: An experiment in online environments. European Sociological Review 34, 223–237.

[3] Barlett, C.P., Simmers, M.M., Roth, B., Gentile, D., 2021. Comparing cyberbullying prevalence and process before and during the covid-19 pandemic. The Journal of Social Psychology , 1–11.

[4] 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.

[5] Benesch, S., Ruths, D., Dillon, K.P., Saleem, H.M., Wright, L., 2016. Considerations for successful counterspeech. Dangerous Speech Project .

[6] Bhattacharya, A., 2021. How Covid-19 lockdowns weakened Facebook's content moderation algorithms. URL: `https://qz.com/india/1976450/facebook-covid-19-lockdowns-hurt-content-moderation-algorithms/`.

[7] Bilewicz, M., 2016. Psychological antecedents of social distance. what leads to contact avoidance? NAUKA .

[8] Bilewicz, M., Tempska, P., Leliwa, G., Dowgiałło, M., Tańska, M., Urbaniak, R., Wroczyński, M., 2021. Artificial intelligence against hate:

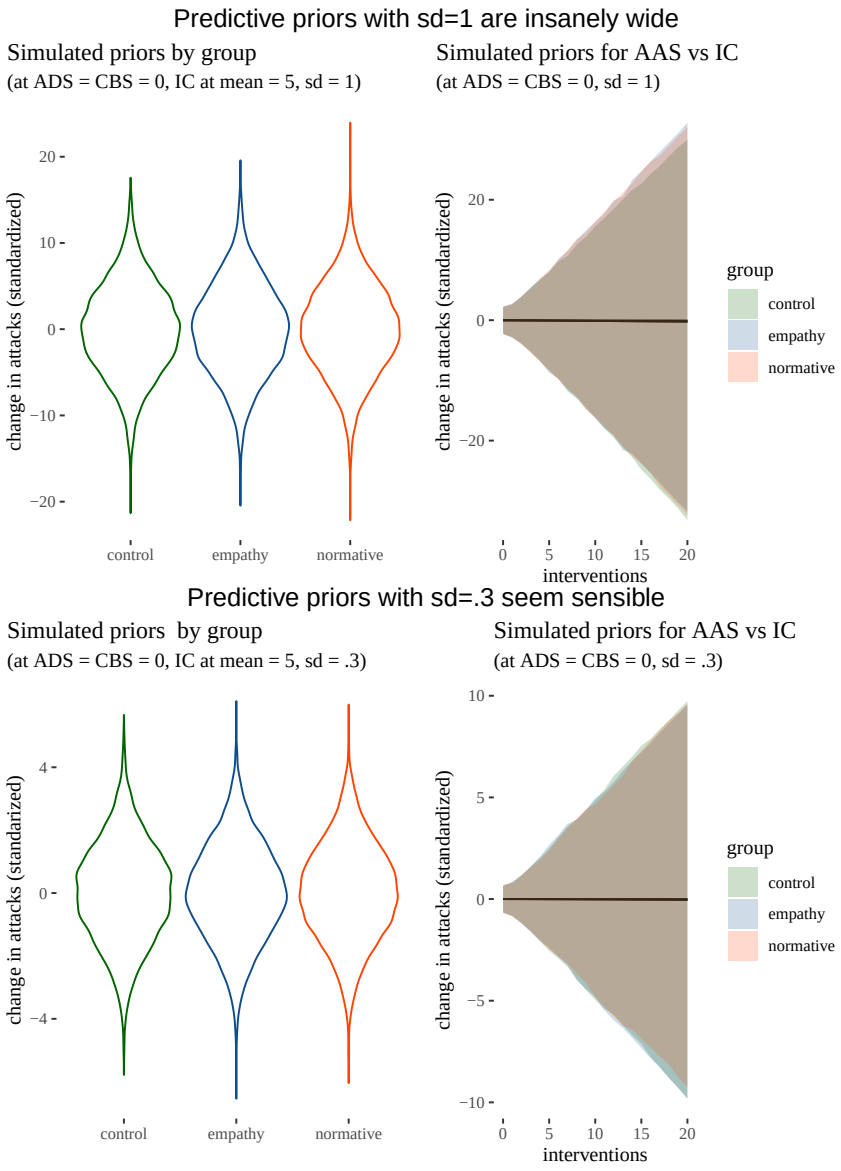

Figure C.19: Prior predictive check for two different sets of priors.

Intervention reducing verbal aggression in the social network environment. Aggressive behavior 47, 260–266.

[9] Binns, R., Veale, M., Van Kleek, M., Shadbolt, N., 2017. Like trainer, like bot? inheritance of bias in algorithmic content moderation, in: International conference on social informatics, Springer. pp. 405–415.

[10] Björkqvist, K., Österman, K., Kaukiainen, A., 2000. Social intelligence-empathy= aggression? Aggression and violent behavior 5, 191–200.

[11] Bloom, P., 2017. Against empathy: The case for rational compassion. Random House.

[12] 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.

[13] Caravita, S.C., Di Blasio, P., Salmivalli, C., 2009. Unique and interactive effects of empathy and social status on involvement in bullying. Social development 18, 140–163.

[14] Chandrasekharan, E., Pavalanathan, U., Srinivasan, A., Glynn, A., Eisenstein, J., Gilbert, E., 2017. You can't stay here: The efficacy of Reddit's 2015 ban examined through hate speech. Proceedings of the ACM on Human-Computer Interaction 1, 1–22.

[15] Cheng, J., Bernstein, M., Danescu-Niculescu-Mizil, C., Leskovec, J., 2017. Anyone can become a troll: Causes of trolling behavior in online discussions, in: Proceedings of the 2017 ACM conference on computer supported cooperative work and social computing, pp. 1217–1230.

[16] Chin, H., Molefi, L.W., Yi, M.Y., 2020. Empathy is all you need: How a conversational agent should respond to verbal abuse, in: Proceedings of the 2020 CHI conference on human factors in computing systems, pp. 1–13.

[17] Cialdini, R.B., Demaine, L.J., Sagarin, B.J., Barrett, D.W., Rhoads, K., Winter, P.L., 2006. Managing social norms for persuasive impact. Social influence 1, 3–15.

[18] Cialdini, R.B., Goldstein, N.J., 2004. Social influence: Compliance and conformity. Annual review of psychology 55, 591–621.

[19] 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.

[20] 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.

[21] Data, O.S., 2021. About the online safety data initiative. URL: https://onlinesafetydata.blog.gov.uk/about-us/.

[22] Dean, S., 2021. A teen who was bullied on snapchat died. his mom is suing to hold social media liable. URL: https://www.latimes.com/business/story/2021-05-10/lawsuit-snap-teen-suicide-yolo-lmk.

[23] Díaz-Narváez, V.P., Coronado, A.M.E., Bilbao, J.L., González, F., Padilla, M., Howard, M., Silva, M.G., Bullen, M., Gutierrez, F., de Villalba, T.V., et al., 2015. Empathy gender in dental students in latin america: an exploratory and cross-sectional study. Health 7, 1527.

[24] Dictionary, M.W., 2002. Merriam-webster. On-line at http://www.mw. com/home. htm 8, 2.

[25] Esteva, A., Kuprel, B., Novoa, R.A., Ko, J., Swetter, S.M., Blau, H.M., Thrun, S., 2017. Dermatologist-level classification of skin cancer with deep neural networks. Nature 542, 115–118. URL: http://www.nature.com/articles/nature21056, doi:10.1038/nature21056.

[26] Failory, 2022. Yik yak's shut down: Why did the location-based app failed? URL: https://www.failory.com/cemetery/yik-yak.

[27] Fink, C., 2018. Dangerous speech, anti-muslim violence, and facebook in myanmar. Journal of International Affairs 71, 43–52.

[28] Fischer, P., Krueger, J.I., Greitemeyer, T., Vogrincic, C., Kastenm¨uller, A., Frey, D., Heene, M., Wicher, M., Kainbacher, M., 2011. The bystander-effect: a meta-analytic review on bystander intervention

in dangerous and non-dangerous emergencies. Psychological bulletin 137, 517.

[29] French, C., 2021. As the pandemic forces us online, LGBTQ2S+ teens deal with cyberbullying. URL: https://www.ctvnews.ca/canada/as-the-pandemic-forces-us-online-lgbtq2s-teens-deal-with-cyberbullying-1.5430945.

[30] Garland, J., Ghazi-Zahedi, K., Young, J.G., Hébert-Dufresne, L., Galesic, M., 2020. Countering hate on social media: Large scale classification of hate and counter speech. arXiv preprint arXiv:2006.01974 .

[31] Gerrard, Y., 2020. ¡? covid19?¿ the covid-19 mental health content moderation conundrum. Social Media+ Society 6, 2056305120948186.

[32] Geva, M., Goldberg, Y., Berant, J., 2019. Are we modeling the task or the annotator? an investigation of annotator bias in natural language understanding datasets. arXiv preprint arXiv:1908.07898 .

[33] Goldstein, N.J., Cialdini, R.B., Griskevicius, V., 2008. A room with a viewpoint: Using social norms to motivate environmental conservation in hotels. Journal of consumer Research 35, 472–482.

[34] Grant, J., 2021. Australia's esafety commissioner targets abuse online as covid-19 supercharges cyberbullying — the strategist. URL: https://www.aspistrategist.org.au/australias-esafety-commissioner-targets-abuse-online-as-covid-19-supercharge

[35] Gröndahl, T., Pajola, L., Juuti, M., Conti, M., Asokan, N., 2018. All you need is" love" evading hate speech detection, in: Proceedings of the 11th ACM workshop on artificial intelligence and security, pp. 2–12.

[36] Hasson, Y., Tamir, M., Brahms, K.S., Cohrs, J.C., Halperin, E., 2018. Are liberals and conservatives equally motivated to feel empathy toward others? Personality and Social Psychology Bulletin 44, 1449–1459.

[37] Hinduja, S., Patchin, J.W., 2010. Bullying, cyberbullying, and suicide. Archives of suicide research 14, 206–221.

[38] Hoff, D.L., Mitchell, S.N., 2009. Cyberbullying: Causes, effects, and remedies. Journal of Educational Administration .

[39] Jordan, M.I., Mitchell, T.M., 2015. Machine learning: Trends, perspectives, and prospects. Science 349, 255–260.

[40] Kallgren, C.A., Reno, R.R., Cialdini, R.B., 2000. A focus theory of normative conduct: When norms do and do not affect behavior. Personality and social psychology bulletin 26, 1002–1012.

[41] Karunakaran, S., Ramakrishan, R., 2019. Testing stylistic interventions to reduce emotional impact of content moderation workers, in: Proceedings of the AAAI Conference on Human Computation and Crowdsourcing, pp. 50–58.

[42] Keipi, T., Näsi, M., Oksanen, A., Räsänen, P., 2016. Online hate and harmful content: Cross-national perspectives. Taylor & Francis.

[43] Koetsier, J., 2020. Report: Facebook Makes 300,000 Content Moderation Mistakes Every Day. URL: https://www.forbes.com/sites/johnkoetsier/2020/06/09/300000-facebook-content-moderation-mistakes-daily-report-says/.

[44] Kowalski, R.M., Limber, S.P., 2013. Psychological, physical, and academic correlates of cyberbullying and traditional bullying. Journal of adolescent health 53, S13–S20.

[45] L1ght, 2020. L1ght releases groundbreaking report on corona-related hate speech and online toxicity. URL: https://l1ght.com/l1ght-releases-groundbreaking-report-on-corona-related-hate-speech-and-onlin

[46] the Label, D., 2017. In:Game Abuse. URL: https://www.ditchthelabel.org/research-papers/ingame-abuse/.

[47] Laub, Z., 2019. Hate Speech on Social Media: Global Comparisons. URL: https://www.cfr.org/backgrounder/hate-speech-social-media-global-comparisons.

[48] Leader Maynard, J., Benesch, S., 2016. Dangerous speech and dangerous ideology: An integrated model for monitoring and prevention. Genocide Studies and Prevention 9.

[49] League, A.D., 2020. Free to Play? Hate, Harassment and Positive Social Experience in Online Games 2020. URL: https://www.adl.org/free-to-play-2020.

[50] LeCun, Y., Bengio, Y., Hinton, G., 2015. Deep learning. nature 521, 436–444.

[51] Lee, E., Leets, L., 2002. Persuasive storytelling by hate groups online: Examining its effects on adolescents. American behavioral scientist 45, 927–957.

[52] Legault, L., Gutsell, J.N., Inzlicht, M., 2011. Ironic effects of antiprejudice messages: How motivational interventions can reduce (but also increase) prejudice. Psychological Science 22, 1472–1477.

[53] Lin, Y., 2021. 10 Twitter Statistics Every Marketer Should Know in 2021 [Infographic]. URL: https://www.oberlo.com/blog/twitter-statistics.

[54] Lipton, Z.C., Steinhardt, J., 2019. Troubling trends in machine learning scholarship: Some ml papers suffer from flaws that could mislead the public and stymie future research. Queue 17, 45–77.

[55] MacAvaney, S., Yao, H.R., Yang, E., Russell, K., Goharian, N., Frieder, O., 2019. Hate speech detection: Challenges and solutions. PloS one 14, e0221152.

[56] Machackova, H., Dedkova, L., Mezulanikova, K., 2015. Brief report: The bystander effect in cyberbullying incidents. Journal of adolescence 43, 96–99.

[57] McGee, M., 2012. Facebook: 3.2 Billion Likes & Comments Every Day. URL: https://martech.org/facebook-3-2-billion-likes-comments-every-day/.

[58] Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., Galstyan, A., 2021. A survey on bias and fairness in machine learning. ACM Computing Surveys (CSUR) 54, 1–35.

[59] Melis, G., Dyer, C., Blunsom, P., 2017. On the state of the art of evaluation in neural language models. arXiv preprint arXiv:1707.05589 .

[60] Morris, A., Cushman, F., 2018. A common framework for theories of norm compliance. Social Philosophy and Policy 35, 101–127.

[61] Munger, K., 2017. Tweetment effects on the tweeted: Experimentally reducing racist harassment. Political Behavior 39, 629–649.

[62] Musgrave, K., Belongie, S., Lim, S.N., 2020. A metric learning reality check, in: European Conference on Computer Vision, Springer. pp. 681–699.

[63] Newton, C., 2013. Killer app: Why do anonymous q&a networks keep leading to suicides? URL: https://www.theverge.com/2013/9/17/4740902/no-good-answers-why-didnt-ask-fm-learn-from-the-formspring-suicides.

[64] Newton, C., 2020. Facebook will pay $52 million in settlement with moderators who developed PTSD on the job. URL: https://www.theverge.com/2020/5/12/21255870/facebook-content-moderator-settlement-scola-ptsd-mental-health.

[65] Nickerson, A.B., Mele, D., Princiotta, D., 2008. Attachment and empathy as predictors of roles as defenders or outsiders in bullying interactions. Journal of school psychology 46, 687–703.

[66] Ouvrein, G., De Backer, C.J., Vandebosch, H., 2018. Online celebrity aggression: A combination of low empathy and high moral disengagement? the relationship between empathy and moral disengagement and adolescents' online celebrity aggression. Computers in human behavior 89, 61–69.

[67] Parks, L., 2019. Dirty data: content moderation, regulatory outsourcing, and the cleaners. Film Quarterly 73, 11–18.

[68] Price, V., Nir, L., Cappella, J.N., 2006. Normative and informational influences in online political discussions. Communication Theory 16, 47–74.

[69] Ptaszyński, M., Leliwa, G., Piech, M., Smywiński-Pohl, A., 2018. Cyberbullying detection–technical report 2/2018, Department of Computer Science AGH, University of Science and Technology. arXiv preprint arXiv:1808.00926 .

[70] Ptaszynski, M., Zasko-Zielinska, M., Marcinczuk, M., Leliwa, G., Fortuna, M., Soliwoda, K., Dziublewska, I., Hubert, O., Skrzek, P., Piesiewicz, J., Karbowska, P., Dowgiallo, M., Eronen, J., Tempska, P., Brochocki, M., Godny, M., Wroczynski, M., 2021. Looking for razors and needles in a haystack: Multifaceted analysis of suicidal declarations on social media—a pragmalinguistic approach. International Journal of Environmental Research and Public Health 18. URL: https://www.mdpi.com/1660-4601/18/22/11759, doi:10.3390/ijerph182211759.

[71] Ptaszynski, M.E., Masui, F., 2018. Automatic Cyberbullying Detection: Emerging Research and Opportunities: Emerging Research and Opportunities. IGI Global.

[72] reddit, 2020. Reddit in 2020. URL: https://www.reddit.com/r/blog/comments/k967mm/reddit_in_2020/.

[73] Reuters, 2020. Facebook frustrates advertisers as boycott over hate speech kicks off. URL: https://www.cnbc.com/2020/07/01/facebook-frustrates-advertisers-as-boycott-over-hate-speech-kicks-off.html.

[74] Rieger, D., Schmitt, J.B., Frischlich, L., 2018. Hate and counter-voices in the internet: Introduction to the special isssue. Studies in Communication and Media , 459–472.

[75] Roberts, S.T., 2014. Behind the screen: The hidden digital labor of commercial content moderation. University of Illinois at Urbana-Champaign.

[76] Roberts, S.T., 2016. Commercial content moderation: Digital laborers' dirty work. Noble and Tynes .

[77] Rosa, H., Pereira, N., Ribeiro, R., Ferreira, P.C., Carvalho, J.P., Oliveira, S., Coheur, L., Paulino, P., Simão, A.V., Trancoso, I., 2019. Automatic cyberbullying detection: A systematic review. Computers in Human Behavior 93, 333–345.

[78] Röttger, P., Vidgen, B., Nguyen, D., Waseem, Z., Margetts, H., Pierrehumbert, J., 2020. Hatecheck: Functional tests for hate speech detection models. arXiv preprint arXiv:2012.15606 .

[79] Sap, M., Card, D., Gabriel, S., Choi, Y., Smith, N.A., 2019. The Risk of Racial Bias in Hate Speech Detection, in: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, Florence, Italy. pp. 1668–1678. URL: https://aclanthology.org/P19-1163, doi:10.18653/v1/P19-1163.

[80] Schieb, C., Preuss, M., 2016. Governing hate speech by means of counterspeech on facebook, in: 66th ica annual conference, at fukuoka, japan, pp. 1–23.

[81] Schmidt, A., Wiegand, M., 2017. A survey on hate speech detection using natural language processing, in: Proceedings of the fifth international workshop on natural language processing for social media, pp. 1–10.

[82] Seifert, D., 2013. Social question and answer site Formspring to shut down on March 31st. URL: https://www.theverge.com/2013/3/15/4110196/social-question-answer-site-formspring-shut-down-march-31st.

[83] Sejnowski, T.J., 2020. The unreasonable effectiveness of deep learning in artificial intelligence. Proceedings of the National Academy of Sciences 117, 30033–30038.

[84] Soral, W., Bilewicz, M., Winiewski, M., 2018. Exposure to hate speech increases prejudice through desensitization. Aggressive behavior 44, 136–146. doi:10.1002/ab.21737.

[85] Soral, W., Malinowska, K., Bilewicz, M., 2022. The role of empathy in reducing hate speech proliferation. two contact-based interventions in online and off-line settings. Peace and Conflict: Journal of Peace Psychology .

[86] Sorrentino, A., Baldry, A.C., Farrington, D.P., Blaya, C., 2019. Epidemiology of cyberbullying across europe: Differences between countries and genders. Educational Sciences: Theory & Practice 19.

[87] Stangor, C., Sechrist, G.B., Jost, J.T., 2001. Changing racial beliefs by providing consensus information. Personality and Social Psychology Bulletin 27, 486–496.

[88] Steffgen, G., König, A., Pfetsch, J., Melzer, A., 2011. Are cyberbullies less empathic? adolescents' cyberbullying behavior and empathic responsiveness. Cyberpsychology, Behavior, and Social Networking 14, 643–648.

[89] Steiger, M., Bharucha, T.J., Venkatagiri, S., Riedl, M.J., Lease, M., 2021. The psychological well-being of content moderators, in: Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems, CHI, pp. 1–14.

[90] Steindl, C., Jonas, E., Sittenthaler, S., Traut-Mattausch, E., Greenberg, J., 2015. Understanding psychological reactance. Zeitschrift für Psychologie .

[91] Steinhardt, J., Koh, P.W., Liang, P., 2017. Certified defenses for data poisoning attacks, in: Proceedings of the 31st International Conference on Neural Information Processing Systems, pp. 3520–3532.

[92] Steinhardt, J., Liang, P., 2015. Learning fast-mixing models for structured prediction, in: International Conference on Machine Learning, PMLR. pp. 1063–1072.

[93] Stepanikova, I., 2012. Racial-ethnic biases, time pressure, and medical decisions. Journal of health and social behavior 53, 329–343.

[94] Stephan, W.G., Finlay, K., 1999. The role of empathy in improving intergroup relations. Journal of Social issues 55, 729–743.

[95] Trengove, M., Kazim, E., Almeida, D., Hilliard, A., Zannone, S., Lomas, E., 2022. A critical review of the online safety bill. Patterns , 100544.

[96] Tworek, H.J., 2021. Fighting hate with speech law: Media and german visions of democracy. The Journal of Holocaust Research 35, 106–122.

[97] Urbaniak, R., Ptaszyński, M., Tempska, P., Leliwa, G., Brochocki, M., Wroczyński, M., 2022a. Personal attacks decrease user activity in social networking platforms. Computers in Human Behavior 126, 106972.

[98] Urbaniak, R., Tempska, P., Dowgiałło, M., Ptaszyński, M., Fortuna, M., Marcińczuk, M., Piesiewicz, J., Leliwa, G., Soliwoda, K., Dziublewska, I., Sulzhytskaya, N., Karnicka, A., Skrzek, P., Karbowska, P., Brochocki, M., Wroczyński, M., 2022b. Namespotting: Username toxicity and actual toxic behavior on reddit. Computers in Human Behavior 136, 107371. URL: https://www.sciencedirect.com/science/article/pii/S0747563222001935, doi:https://doi.org/10.1016/j.chb.2022.107371.

[99] 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.

[100] Vanian, J., 2019. Google's Hate Speech Detection A.I. Has a Racial Bias Problem. URL: https://fortune.com/2019/08/16/google-jigsaw-perspective-racial-bias/.

[101] Vogels, E.A., 2021. The State of Online Harassment. URL: https://www.pewresearch.org/internet/2021/01/13/the-state-of-online-harassment/.

[102] Wachs, S., Wright, M.F., Sittichai, R., Singh, R., Biswal, R., Kim, E.m., Yang, S., Gámez-Guadix, M., Almendros, C., Flora, K., et al., 2019. Associations between witnessing and perpetrating online hate in eight countries: The buffering effects of problem-focused coping. International journal of environmental research and public health 16, 3992.

[103] Williams, M., 2019. Hatred behind the screens: A report on the rise of online hate speech. Mishcon de Reya .

[104] Williams, M.L., Burnap, P., Javed, A., Liu, H., Ozalp, S., 2020. Hate in the machine: anti-black and anti-muslim social media posts as predictors of offline racially and religiously aggravated crime. The British Journal of Criminology 60, 93–117. doi:10.1093/bjc/azz049.

[105] Woolley, A.W., Aggarwal, I., Malone, T.W., 2015. Collective intelligence and group performance. Current Directions in Psychological Science 24, 420–424.

[106] Wright, M.F., Wachs, S., 2021. Does empathy and toxic online disinhibition moderate the longitudinal association between witnessing and perpetrating homophobic cyberbullying? International journal of bullying prevention 3, 66–74.

[107] Wu, T., Ribeiro, M.T., Heer, J., Weld, D.S., 2019. Errudite: Scalable, reproducible, and testable error analysis, in: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pp. 747–763.

[108] Xie, J., Sreenivasan, S., Korniss, G., Zhang, W., Lim, C., Szymanski, B.K., 2011. Social consensus through the influence of committed minorities. Physical Review E 84, 011130.

[109] Yang, B., Wang, B., Sun, N., Xu, F., Wang, L., Chen, J., Yu, S., Zhang, Y., Zhu, Y., Dai, T., et al., 2021. The consequences of cyberbullying and traditional bullying victimization among adolescents: gender differences in psychological symptoms, self-harm and suicidality. Psychiatry research 306, 114219.

[110] Yanovitzky, I., Rimal, R., 2006a. Communication and normative influence: An introduction to the special issue. Communication Theory 16, 1–6.

[111] Yanovitzky, I., Rimal, R.N., 2006b. Communication and normative influence: An introduction to the special issue. Communication Theory 16, 1–6.

[112] Yin, W., Zubiaga, A., 2021. Towards generalisable hate speech detection: a review on obstacles and solutions. PeerJ Computer Science 7, e598.

[113] Zaki, J., 2014. Empathy: a motivated account. Psychological bulletin 140, 1608.

[114] Ziegele, M., Jost, P., Bormann, M., Heinbach, D., 2018. Journalistic counter-voices in comment sections: Patterns, determinants, and potential consequences of interactive moderation of uncivil user comments. SCM Studies in Communication and Media 7, 525–554.

[115] Ziems, C., He, B., Soni, S., Kumar, S., 2020. Racism is a virus: Anti-asian hate and counterhate in social media during the covid-19 crisis. arXiv preprint arXiv:2005.12423 .