

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

Abstract goes here.

Keywords: keyword, keyword

PACS: PACScode, PACScode

2000 MSC: MSCcode, MSCcode

Contents

1	Automated versus human-based moderation	2
2	Different approaches to content moderation	6
3	Counter-speech interventions	7
4	Technology applied for personal attack detection	12
5	Data analysis	14
5.1	Data collection and exploration	14
5.2	Causal thinking, choice of variables and models	23
5.3	Results	29
5.3.1	Interventions on a given day	29
5.3.2	Cumulative sum of interventions	29
5.3.3	Long term before/after analysis	29
6	Discussion	36
7	Volunteer engagement and impact of competitions	36
7.1	The challenge of keeping volunteers engaged	36
7.2	Volunteer activity data analysis	36

Appendix A	Explanation of WAIC	40
Appendix B	Time series model selection	41
Appendix C	Model choice for the long term analysis	45

1. 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 [27, 28]. State-of-the-art solutions are mostly governed by statistical methods, including deep learning and machine learning [29, 30, 31]. 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 [32, 33, 34]. For instance, Sap et al. [35] 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,¹ 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 [36]. 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

¹<https://perspectiveapi.com>

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. [37], 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 sentences—a word which is negatively correlated with hate speech—resulted in the F1 score² performance 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 [38, 39]. As shown again by Gröndahl et al. [37], 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. [40], Lipton and Steinhardt [41], Musgrave et al. [42] 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 [41] 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. [43] 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

²This is an overall measure of a model’s accuracy that combines precision and recall.

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. [44], 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 [41], there is a tremendous value coming from the thorough understanding of a particular method, and a variety of techniques are vital in the 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 [45] 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. [46] 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. [47], or the Online Safety Data Initiative (OSDI) [48], 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 [49], 500M tweets are sent daily on Twitter [50], and over 2B comments made by users of Reddit in 2020 [51], 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 [52, 53]. 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 [52]. 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 [54], workplace wellness programs, clinical support, or psychological training [55], 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 [56]. Also, as described by Parks [57], 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 [58], high time

pressure can amplify human biases. Considering that Facebook employs 15K moderators [59] 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.

2. 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 [60, 61, 62]. 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 [63, 64, 65].

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.

3. 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 with the purpose of stopping it and reducing its consequences [66, 67]. Hate speech is a type of speech used with the intention to insult, offend, or intimidate a person based on their identity (such as race, religion, sexual orientation, national origin, or disability) [68]. 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 [69].

Benesch et al. [70] 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. [71] proposed a three-layered classification of counter-speech: primary prevention, where counter-voices are being disseminated to the general public in order 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 classification studies. 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 [67].

Another study by Ziegele et al. [72] 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, the work by Ziems et al. [73] 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, they 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 solution for effects of hate on social media platforms.

[74] 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, [69] 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 content to be a more effective strategy of toxicity reduction [17], it also collides with civil liberties and diffuses rather than reduces hate [75]. Therefore counter-speech seems to be a golden means for reducing antisocial behavior while preserving civil liberties.

In our study, we have employed counter-speaking strategies to see whether they can be applied 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. This can include name-calling, insults, or other types of derogatory language that is directed at an individual or group [76]. 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 [77] 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 presence of other people in critical situations reduces the probability of assistance from bystanders [78].

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 [72, 69, 79].

The issue of low bystander reactivity has been addressed in our study by engaging volunteers personally committed to lowering online violence, and creating an opportunity structure for quick response, by real-time notifica-

tions about the occurrences of personal attacks that they should respond to (see however our section on volunteer engagement for a discussion of difficulties with this strategy).

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. Let us elaborate.

Social norms are agreed to be powerful determinants of human behavior. According to theories of norm compliance people tend to comply with the norms they are exposed to [80, 81].

Multiple studies have found that perceptions of what others are doing cause us to behave similarly. Norms can be applied to exert persuasive impact, for example, to solve a social problem such as littering or wasting resources [82, 81, 83].

Some scholars propose that perceived social norms serve as shortcuts helping us quickly adapt to new environments [81]. 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 [80].

The Focus Theory of Normative Conduct [84] is a theoretical psychological framework used for understanding best conditions and practices for exerting normative influence. Its main assumption is that norms that are focal in attention are likely to influence behavior. One important distinction made within this approach is between two types of norms: a descriptive norm refers to what is typically done, by showing an example of normal behavior in a given situation, and an injunctive or prescriptive norm refers to what is commonly approved, such as the rules of proper conduct and moral beliefs [81].

A descriptive norm motivates compliance by providing evidence about what is likely to be an effective and adaptive action, while an injunctive norm motivates by informing the agent which actions will be approved and which will result in social sanctions [85, 84].

The positive influence of injunctive and descriptive norms on unwanted social behavior was first presented in a study on littering in public spaces conducted by Cialdini et al. [81]. The 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.

Injunctive messages directly asking not to litter had significantly reduced littering. The subjects littered more in a highly littered environment and littered less in a clean environment, although these effects were not statistically significant.³

A descriptive norm can be also evoked by providing written information about what others usually do. Goldstein et al. [83] distributed flyers with normative messages that contained descriptive information about reuse of hotel towels (“the majority of guests reuse their towels”). This strategy has been proven superior to traditional appeals focusing on environmental 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 think and behave 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 [86], and public expression of prejudice, and reactions to hostile jokes were highly correlated with social approval of these expressions [87].

When it comes to the online realm, group norm promotion and social sanctioning of online racist harassment resulted in the reduction of racist tweets (although only in the condition in which a white male with high followers was approaching another white male). In other words, sending direct, counter-speech messages to users with similar identities, resulted in a decrease in their use of hate speech on Twitter [88]. Moreover, in a study conducted by Bilewicz et al. [76], a bot disguised as a Reddit user, equipped with normative and empathetic interventions, significantly decreased the number of personal attacks generated in this forum. In addition, types of statements made by people in online political discussions influenced the way others expressed their opinions [89], and the expression of online hate-speech in [17] was strongly dependant on the amount of hate-speech already present in a given environment.

The context of online interactions can also contribute to the spread of

³It is also worth noticing that seeing litter in an otherwise clean environment could have also engaged an injunctive norm, as it might have reminded subjects about societal objections to littering.

counter-normative behaviors. For example, negative mood and prior exposition to troll content was the strongest predictor for trolling behavior in [90].

In line with the above discussion, for the purpose of this study we have distinguished two types of responses to personal attacks: "empathetic" and "normative". The empathetic approach is aimed at countering the so-called desensitization to violence. It is a very dangerous process, finally leading to the dehumanization of stigmatized groups. Being frequently exposed to online hate speech builds tolerance to the phenomenon and as a result, we no longer see such statements as offensive, which lowers our sympathy for the victims and increases prejudice. By sending empathetic interventions we hoped to evoke compassion and self-reflection, and to humanize the person on the other side of the screen [76].

The normative approach stems from the already discussed 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. Taking from these studies, we refer to two types of norms: injunctive norms—describing what the members of this community are supposed to do and what is approved or disapproved by others (laws, rules, code of conduct, *savoir-vivre*, etc.) and descriptive norms—what the members of this community actually do [91, 85, 82].

4. 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 1, 2 and ??, 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 [? 19] and in

the mass media^{4,5} has lead to an increase in the number of automatic tools being developed and deployed to support content moderators in their work.⁶

In our research we applied Samurai detection software, a proprietary technology developed Samurai Labs.⁷ 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 [?], detection of toxic usernames [?], and detection of personal attacks on Reddit [19]. 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 making nasty insinuations without providing evidence. Three examples of typical personal attacks are as follows:

- *You are legit mentally retarded homie.*
- *Eat a bag of dicks, fuckstick.*

⁴<https://www.japantimes.co.jp/news/2022/12/03/business/tech/twitter-harmful-content/>

⁵<https://www.forbes.com/sites/forbestechcouncil/2022/06/14/the-growing-role-of-ai-in-content-moderation/>

⁶<https://influencemarketinghub.com/content-moderation-tools/>.

⁷<https://www.samurailabs.ai/>, described in (Ptaszyński, Leliwa, Piech, & Smywiński-Pohl, 2018; Wroczynski & Leliwa, 2019).

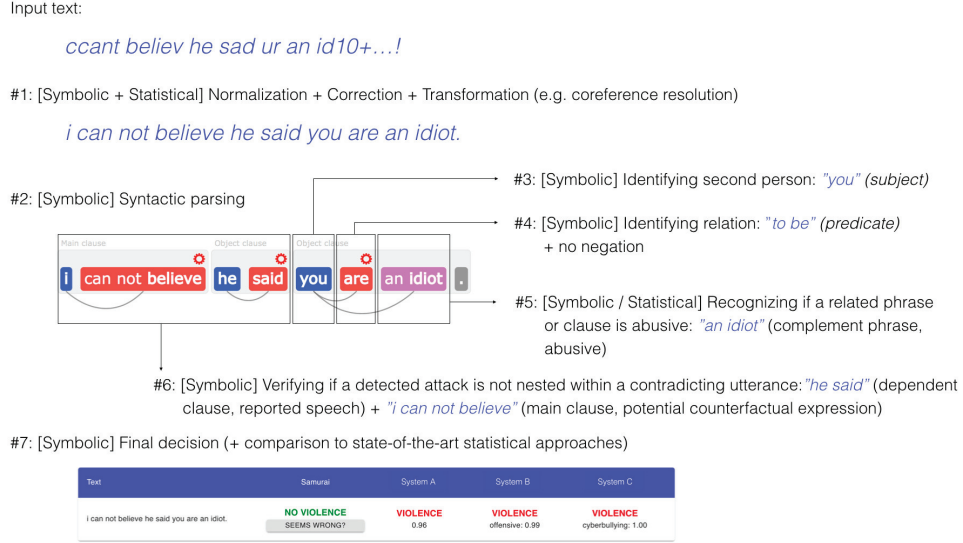


Figure 1: Example of processing of one sentence by the applied Samurai technology.

- *Fuck off with your sensitivity you douche.*

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 methods. The performance of Samurai has been measured on a number of occasions and evaluated as sufficiently reliable for practical use [? ? 19?].

5. Data analysis

5.1. Data collection and exploration

We conducted a 6-months field 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 interesting 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 at 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 unmoderated 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

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.

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.

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 1.⁸ 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.

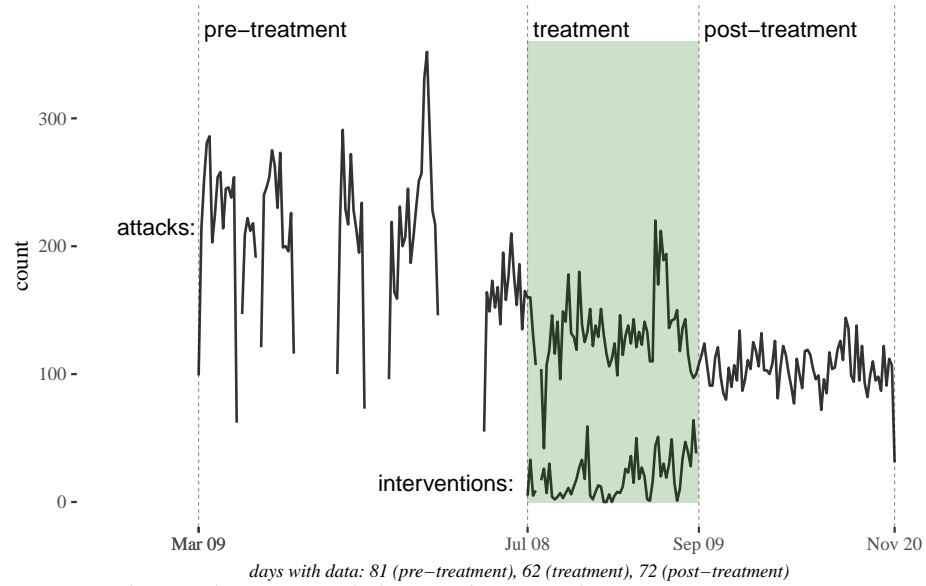
Second, in the distribution of standardized difference in attacks, the peaks of distributions are shifted a bit between the groups, with lowest median for the normative group, but the differences seem minor (Figure 4). This might suggest no impact of the interventions. This conclusion would be too hasty,

⁸Further 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.

add
ref

Personal attacks and interventions time series

no line at data gaps



Personal attacks vs weekdays (six months)

no weekly patterns

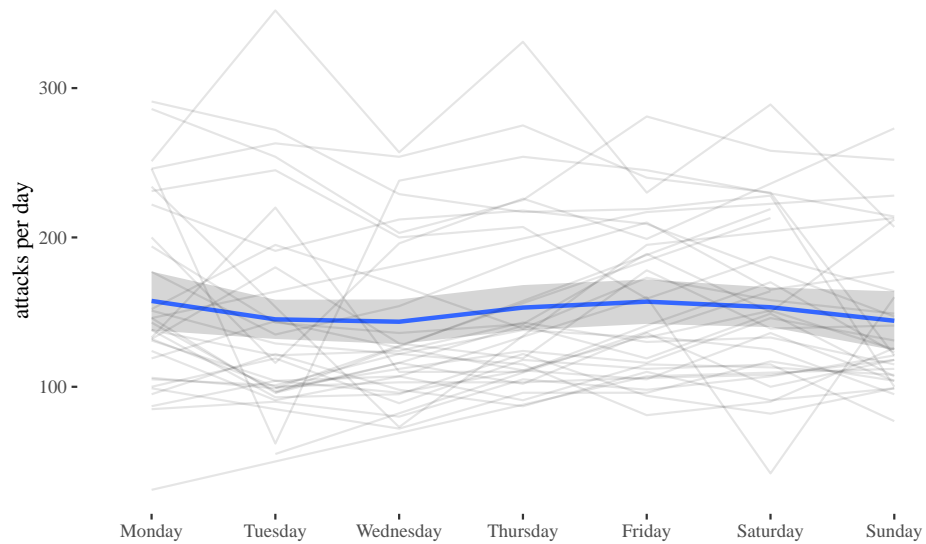


Figure 2: Daily sums of attacks and interventions throughout the three experimental periods, with GAM smoothing (left) and daily attack sums from all weeks in the experimental period plotted against week days (right)—no pattern seems to arise.

Total intervention counts by users and treatment groups

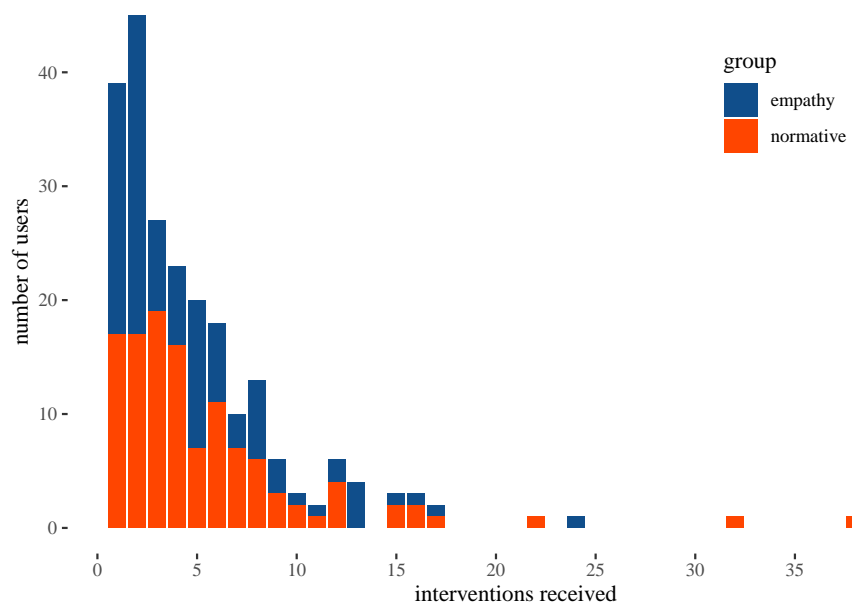


Figure 3: Distribution of daily interventions (by treatment group).

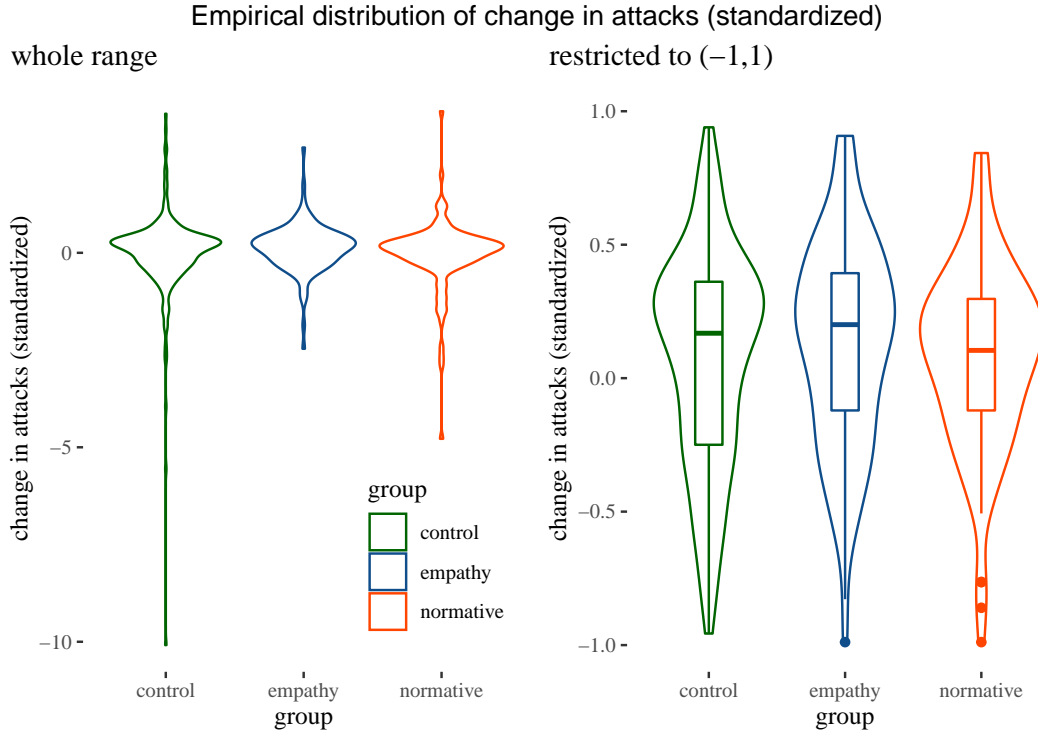


Figure 4: Empirical distribution of change in attacks (by treatment group).

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 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

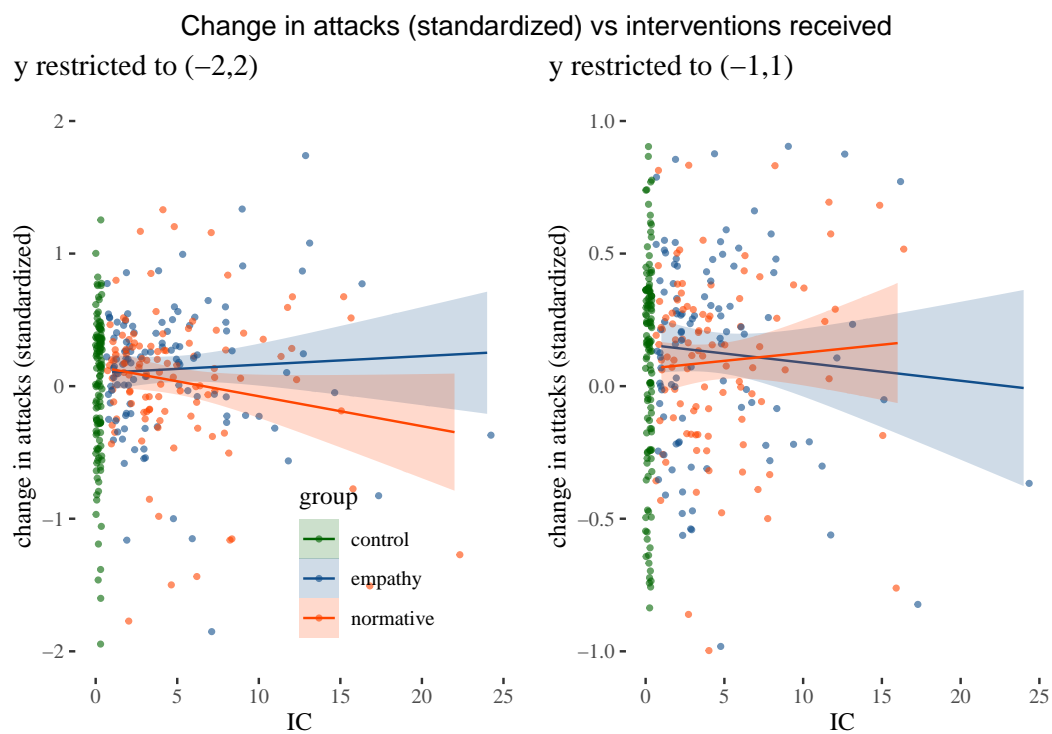


Figure 5: Change in attacks vs the number of interventions received by treatment group, jittered with linear smoothing.

Individual time series by group and activity (examples)

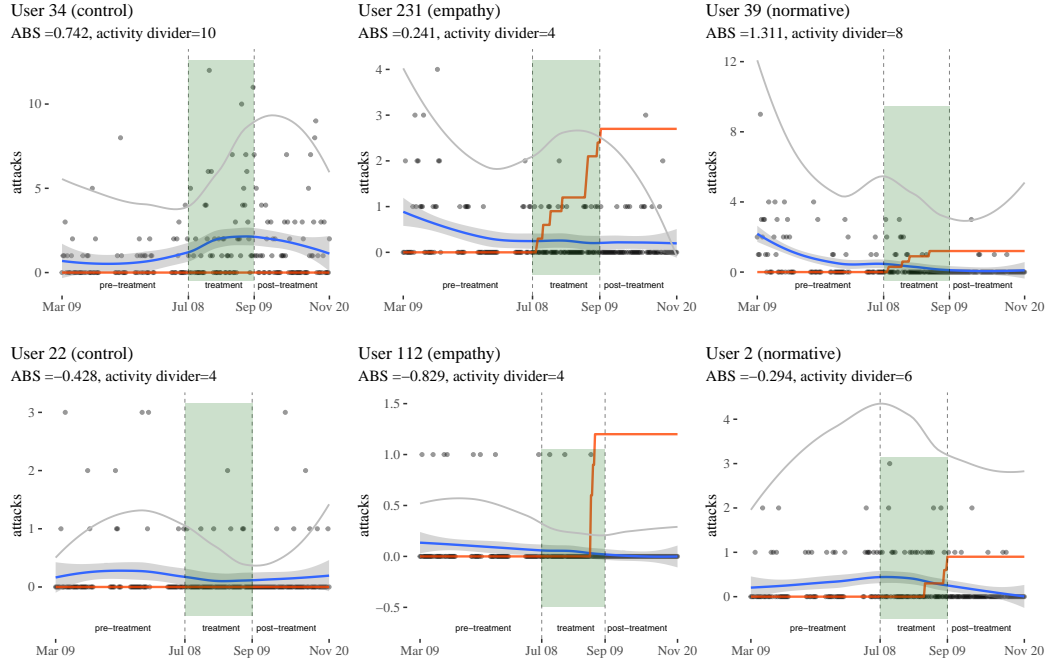


Figure 6: Examples of individual time series. Black points are attacks (smoothed in blue), red lines represent the cumulative number of interventions received (lag 3) divided by 10, gray lines represent overall activity level divided by a variable divider listed in the subtitle. Divisions introduced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

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, to which we will now move.

5.2. Causal thinking, choice of variables and models

First, we inspect correlations between predictors to avoid multicollinearity, 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 multicollinearity 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

Only activity levels through time are strongly correlated

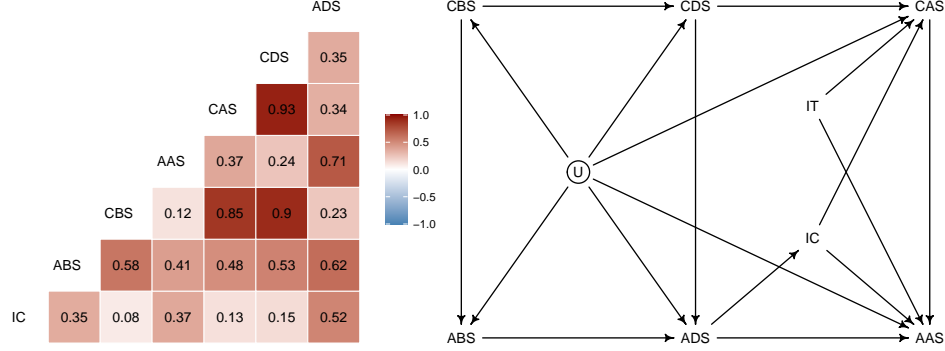


Figure 7: Correlations between predictors (left) and a plausible causal model (right) used in the before-and-after analysis.

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, 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, $k \in \{1, \dots, 7\}$). 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.⁹

⁹In 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

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 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:

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

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

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.

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 subgroup—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):

$$\begin{aligned}
& \text{attacks}_i \sim \text{NegativeBinomial}(\lambda_i, \phi_{\text{userID}[i]}) \\
& \log(\lambda_i) = l_{\text{userID}[i]} + a_{\text{userID}[i]} \times \text{attacksL1} + c_{\text{userID}[i]} \times \text{act} + \\
& \quad + i1control_{\text{userID}[i]} \times \text{control} \times \text{intL1D} + \\
& \quad + i1emp_{\text{userID}[i]} \times \text{emp} \times \text{intL1D} + \\
& \quad + i1norm_{\text{userID}[i]} \times \text{norm} \times \text{intL1D} \\
& \quad l_{\text{userID}[i]} \sim \text{Norm}(\bar{l}, \bar{\sigma}_l) \\
& \quad a_{\text{userID}[i]} \sim \text{Norm}(\bar{a}, \bar{\sigma}_a) \\
& \quad c_{\text{userID}[i]} \sim \text{Norm}(\bar{c}, \bar{\sigma}_c) \\
& \quad i1control_{\text{userID}[i]} \sim \text{Norm}(i1controlOverall, \bar{\sigma}_{i1}) \\
& \quad i1emp_{\text{userID}[i]} \sim \text{Norm}(i1empOverall, \bar{\sigma}_{i1}) \\
& \quad i1norm_{\text{userID}[i]} \sim \text{Norm}(i1normOverall, \bar{\sigma}_{i1}) \\
& \quad i1controlOverall \sim \text{Norm}(0, .2) \\
& \quad i1empOverall \sim \text{Norm}(0, .2) \\
& \quad i1normOverall \sim \text{Norm}(0, .2) \\
& \quad \bar{\lambda} \sim \text{Norm}(.00001, 2.5) \\
& \quad \bar{\sigma}_l \sim \text{Exp}(1.5) \\
& \quad \bar{a} \sim \text{Norm}(0, .2) \\
& \quad \bar{\sigma}_a \sim \text{Exp}(5) \\
& \quad \bar{c} \sim \text{Norm}(0, .2) \\
& \quad \bar{\sigma}_c \sim \text{Exp}(5) \\
& \quad \bar{\sigma}_{i1} \sim \text{Exp}(5)
\end{aligned}$$

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

- Each user has their own baseline aggression level, $l_{\text{userID}[i]}$.
- However, these individual aggression levels are not disconnected, they come from a distribution themselves, $\text{Norm}(\bar{l}, \bar{\sigma}_l)$. \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 $a_{\text{userID}[i]}$, which capture the correlation between yesterday's attacks with today's at-

tacks, so to speak. These also come from a general distribution $\text{Norm}(\bar{a}, \bar{\sigma}_a)$, 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, $c_{\text{userID}[i]}$, all coming from a general distribution $\text{Norm}(\bar{c}, \bar{\sigma}_c)$ whose parameters are also to be estimated.
- For any particular treatment group, say, empathy, we have a user level coefficient $i1emp_{\text{userID}[i]}$, 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 $\text{Norm}(i1emp_{\text{Overall}}, \bar{\sigma}_{i1})$, 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:

$$\begin{aligned} \log(\lambda_i) = & l_{\text{userID}[i]} + c_{\text{userID}[i]} \times \text{act} + ic3control_{\text{userID}[i]} \times \text{control} \times \text{intCL3D} + \\ & + ic3emp_{\text{userID}[i]} \times \text{emp} \times \text{intCL3D} + \\ & + ic3norm_{\text{userID}[i]} \times \text{norm} \times \text{intCL3D} \\ \log(\lambda_i) = & l_{\text{userID}[i]} + c_{\text{userID}[i]} \times \text{act} + \\ & + ic3control_{\text{userID}[i]} \times \text{control} \times \text{intCL3D} + \\ & + icabst3control_{\text{userID}[i]} \times \text{control} \times \text{intCL3D} \times \text{abst} + \end{aligned}$$

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

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

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):

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

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.

5.3. Results

5.3.1. 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.

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.

5.3.2. 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 efficiency 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.

5.3.3. 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.

Multiplicative impact of a single past intervention (with 89% posterior density intervals)

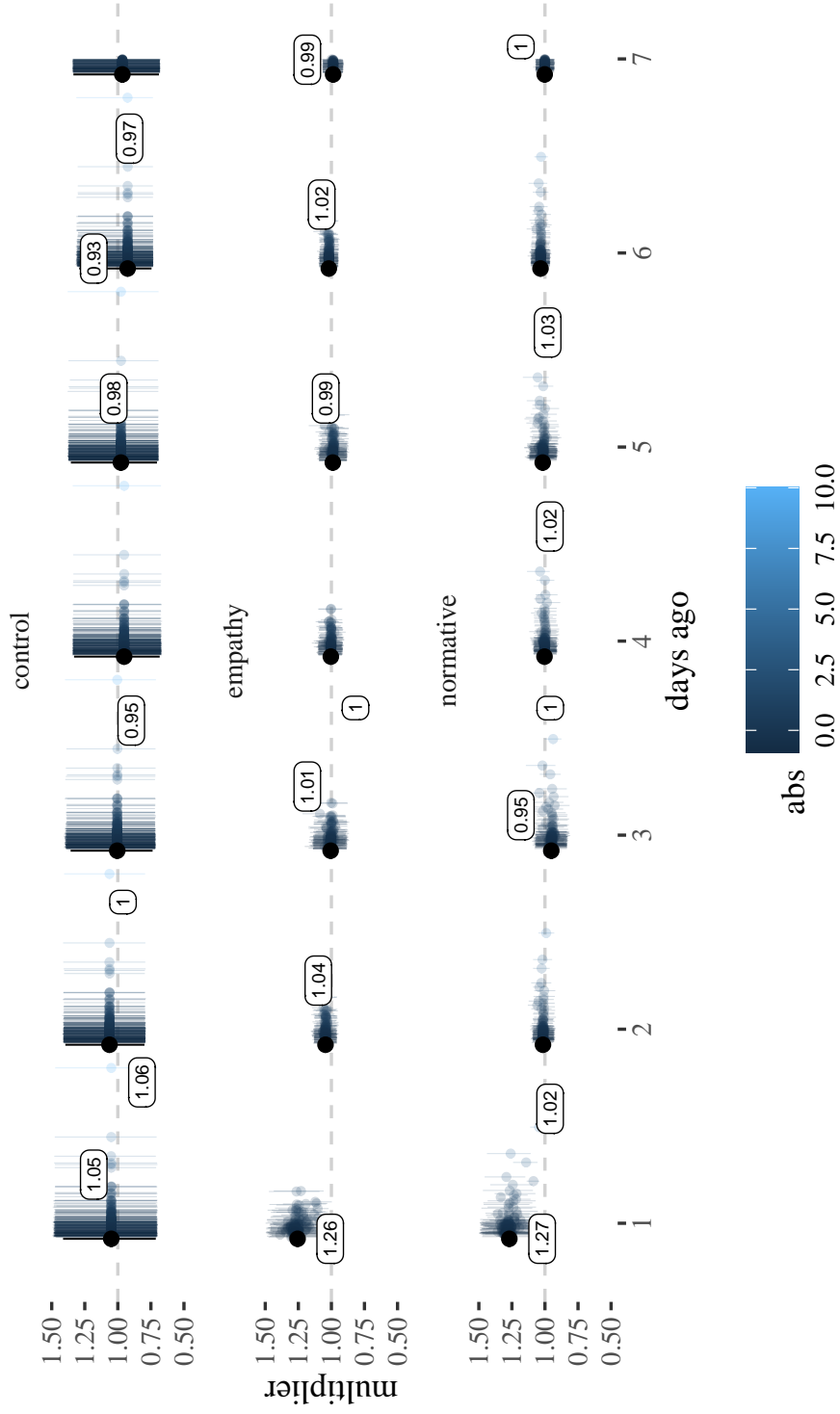


Figure 8: Impact of interventions received lag $1 \leq k \leq 7$ on attacks on a given day. High-level coefficients are pictured in black.

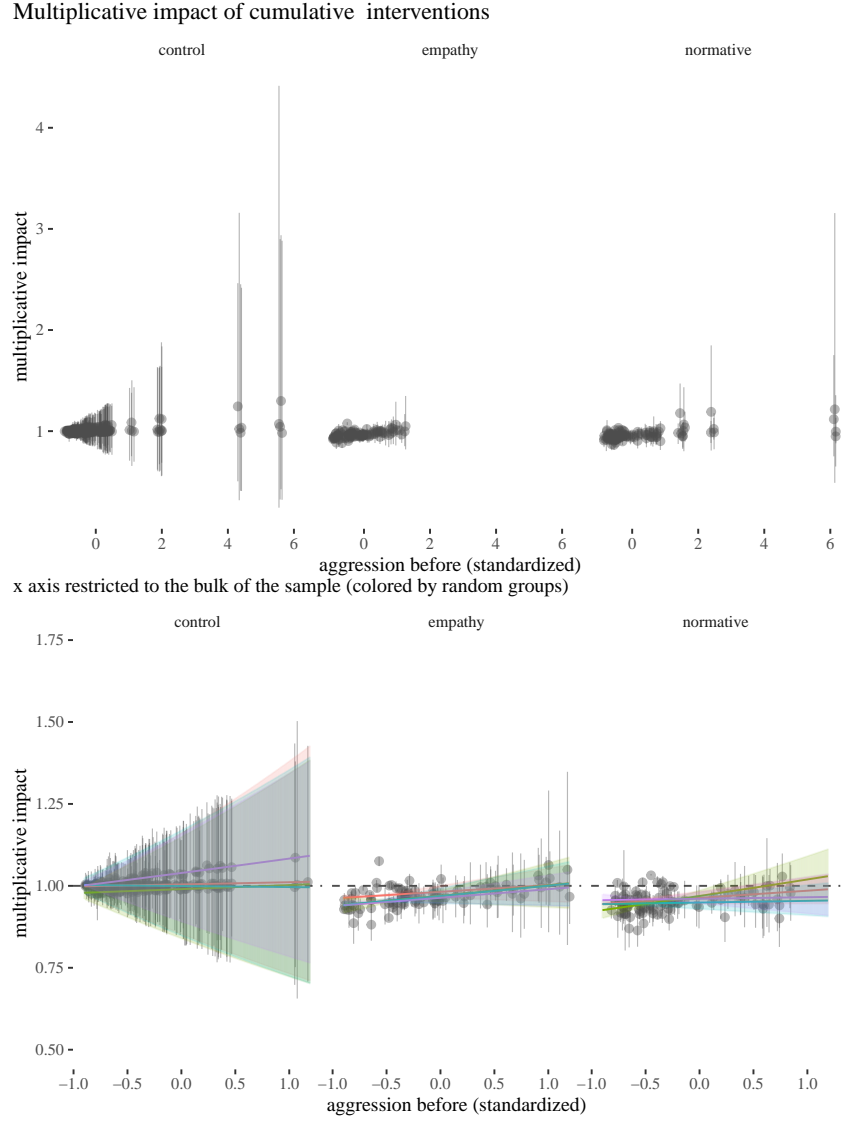


Figure 9: Multiplicative impact of cumulative interventions lag 3 on attacks. Individual users' coefficients only, full range (top), and with attention restricted to the bulk of the sample. Sub-sample coefficients depend on aggression and are represented as lines, colored by sub-sample. Note low number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

Multiplicative impact of cumulative interventions in three aggression quantiles

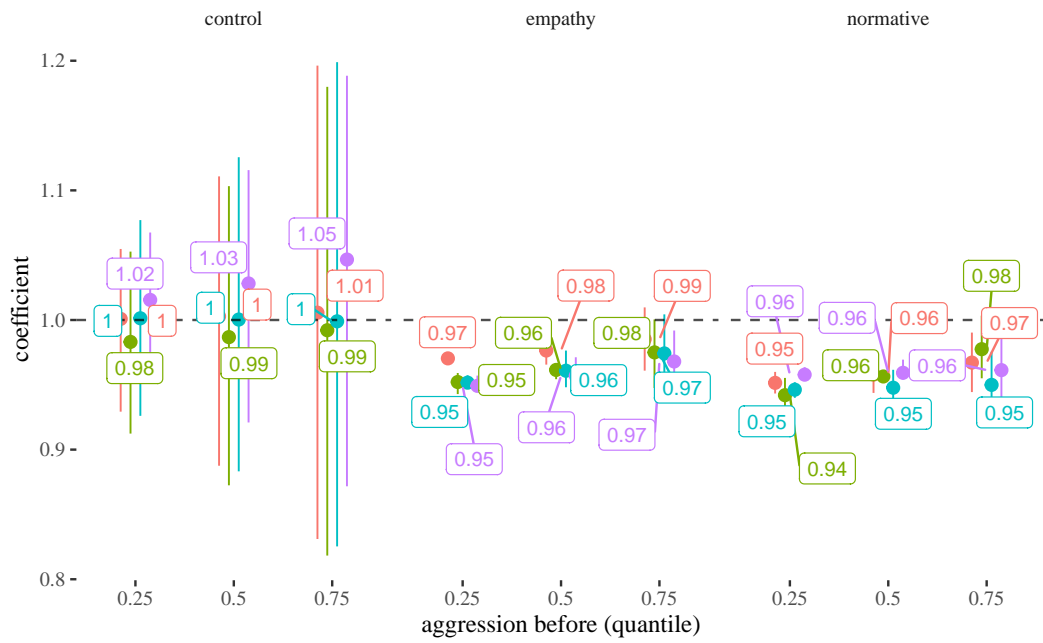


Figure 10: Multiplicative impact of cumulative interventions lag 3 on attacks. General level coefficients only, in three quantiles (.25, .5, .75). Colored by sub-sample.

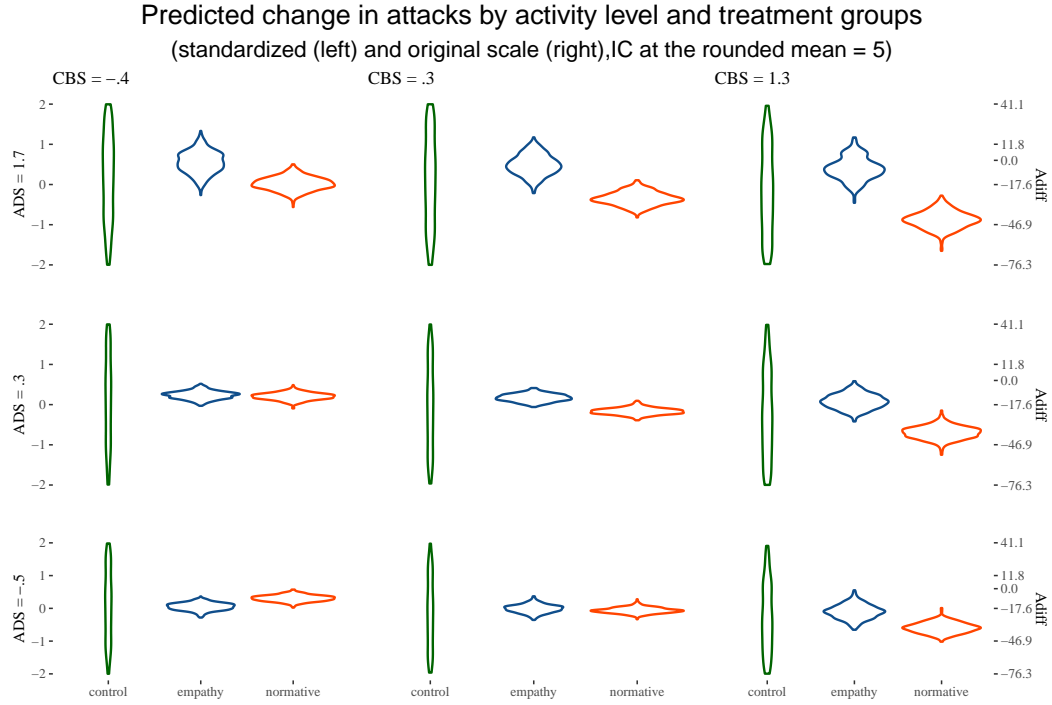


Figure 11: Predicted change in attacks, depending on user’s activity level (CBS: comments before, standardized) and how aggressive overall they were (ADS: attacks during, standardized). The more aggressive and active the users, the higher the attacks drop in the normative group, slight drop correlated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

- The values CBS range from -0.82 to 18.3, with approximately 30% below -0.4, around 80% below 0.3, 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 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

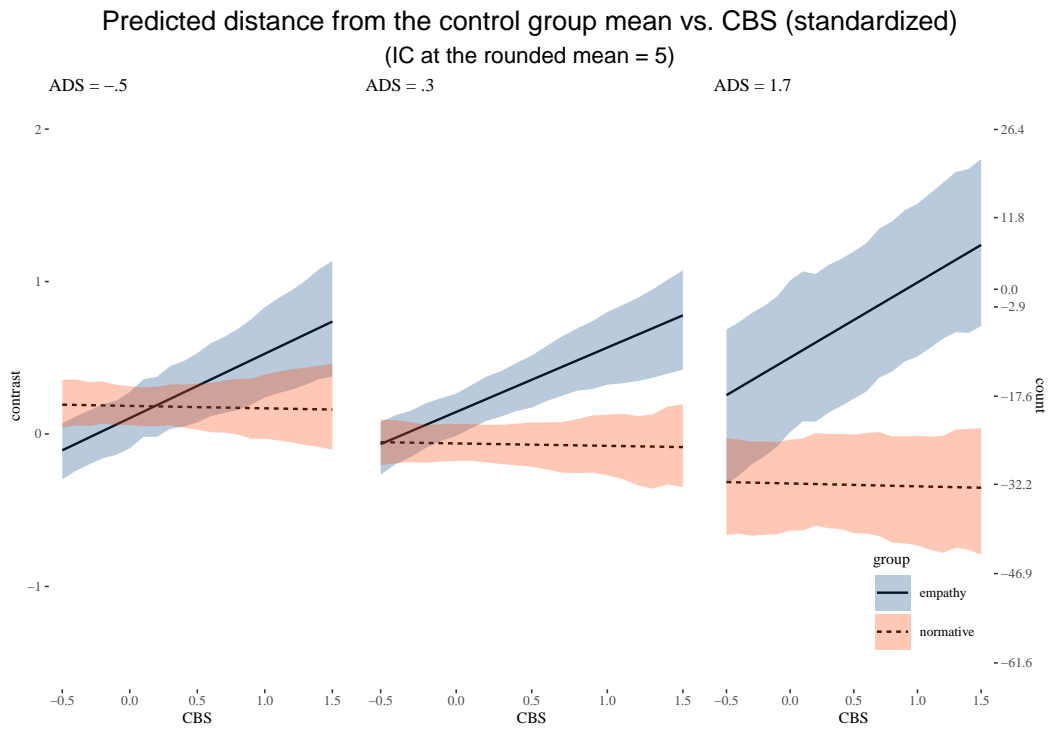


Figure 12: Predicted contrasts (difference in attacks as compared to the control group) for the two treatment groups vs activity before the treatment. Notice that empathetic interventions correlated with decreased attacks for less active users, but performed worse than normative interventions for more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

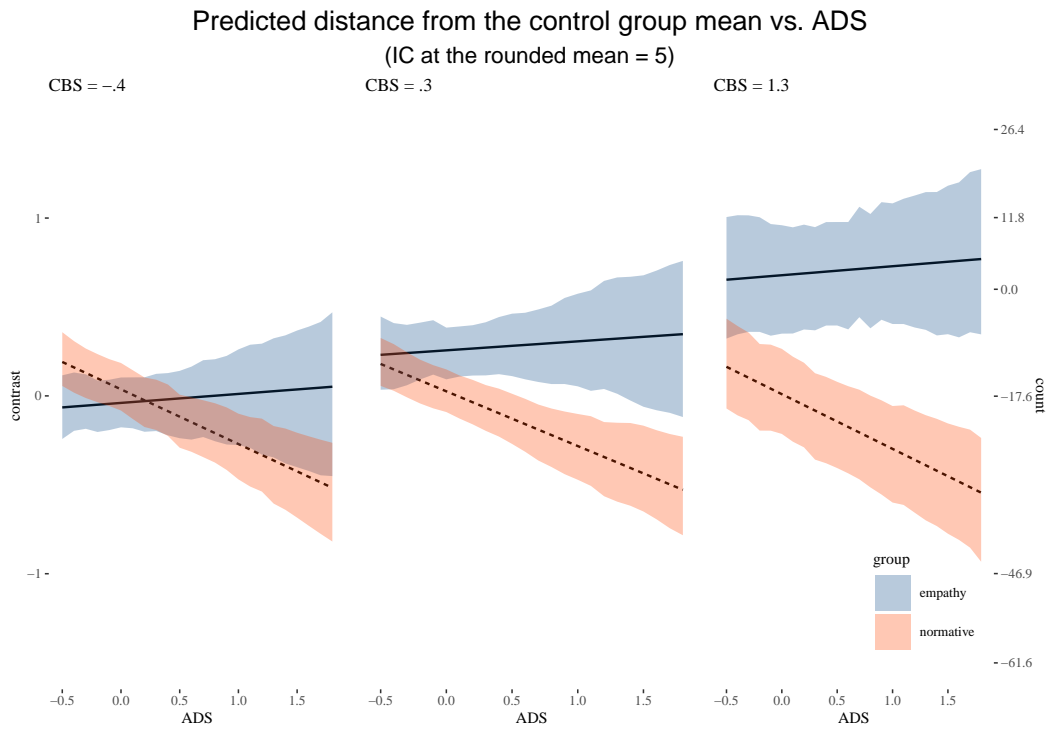


Figure 13: Predicted contrasts (difference in attacks as compared to the control group) for the two treatment groups vs aggression during the treatment period. Notice that empathetic interventions correlated with decreased attacks for less aggressive users, but performed worse than normative interventions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

Predicted distance from the control group vs. IC

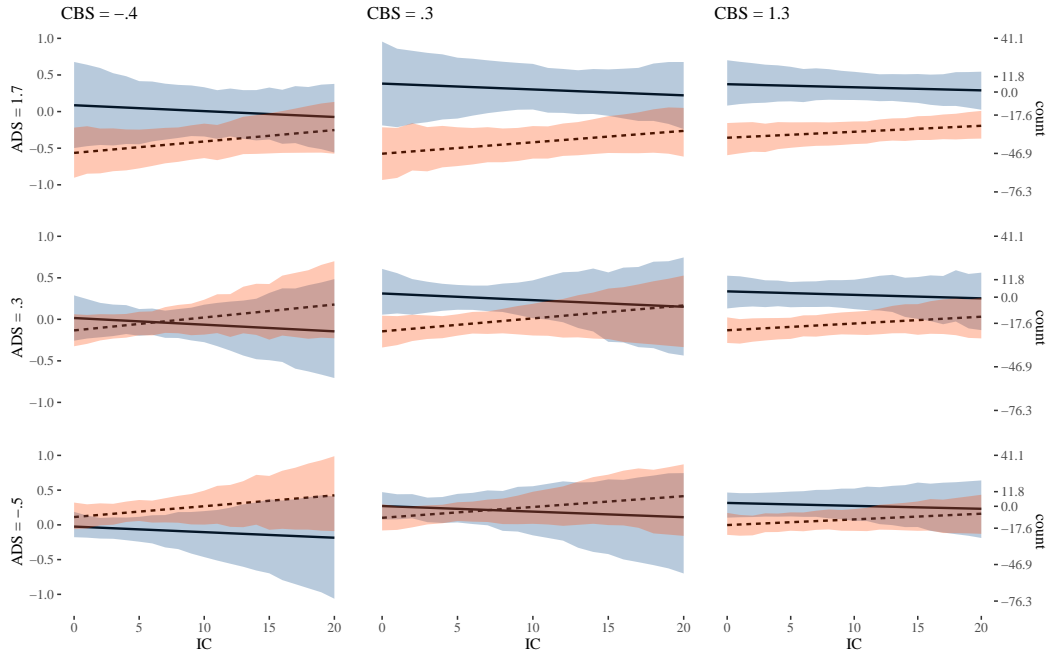


Figure 14: Contrasts (change in attacks as compared to the control group) vs the number of interventions received. Note that repeating empathetic interventions correlates with decreased attacks, while repeating normative interventions is counterproductive.

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).

6. 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.

7. Discussion

8. Volunteer engagement and impact of competitions

8.1. The challenge of keeping volunteers engaged

8.2. Volunteer activity data analysis

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

$$\begin{aligned}
& \text{interventions} \sim \text{NegativeBinomial}(\lambda, \phi) \\
& \log(\lambda) = l_{\text{volunteerID}[i]} + \text{enth}_{\text{volunteerID}[i]} \times \text{daysOfProject} + \text{comp}_{\text{volunteerID}[i]} \times \text{competition} \\
& l_{\text{volunteerID}[i]} \sim \text{Norm}(lbar, l\text{sigmabar}) \\
& lbar \sim \text{Norm}(2, .9) \\
& l\text{sigmabar}, \text{enth}\text{sigmabar}, \text{comp}\text{sigmabar} \sim \text{Exp}(.5) \\
& \text{enth}_{\text{volunteerID}[i]} \sim \text{Norm}(\text{enthbar}, \text{enth}\text{sigmabar}) \\
& \text{comp}_{\text{volunteerID}[i]} \sim \text{Norm}(\text{compbar}, \text{comp}\text{sigmabar}) \\
& \text{enthbar}, \text{compbar} \sim \text{Norm}(0, .3) \\
& \phi = puser_{\text{volunteerID}[i]} \\
& puser_{\text{volunteerID}[i]} \sim \text{Exp}(1)
\end{aligned}$$

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, $l_{\text{volunteerID}[i]}$.
- next, each user has their own dispersion parameter, $p_{\text{user}_{\text{volunteerID}[i]}}$.
- then, there is (usually dwindling) enthusiasm: the impact of time on that user, $enth_{\text{volunteerID}[i]}$ 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, $comp_{\text{volunteerID}[i]}$, 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.¹⁰

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.

¹⁰Interestingly, if we are interested in the causal effect of competitions, we should not 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.

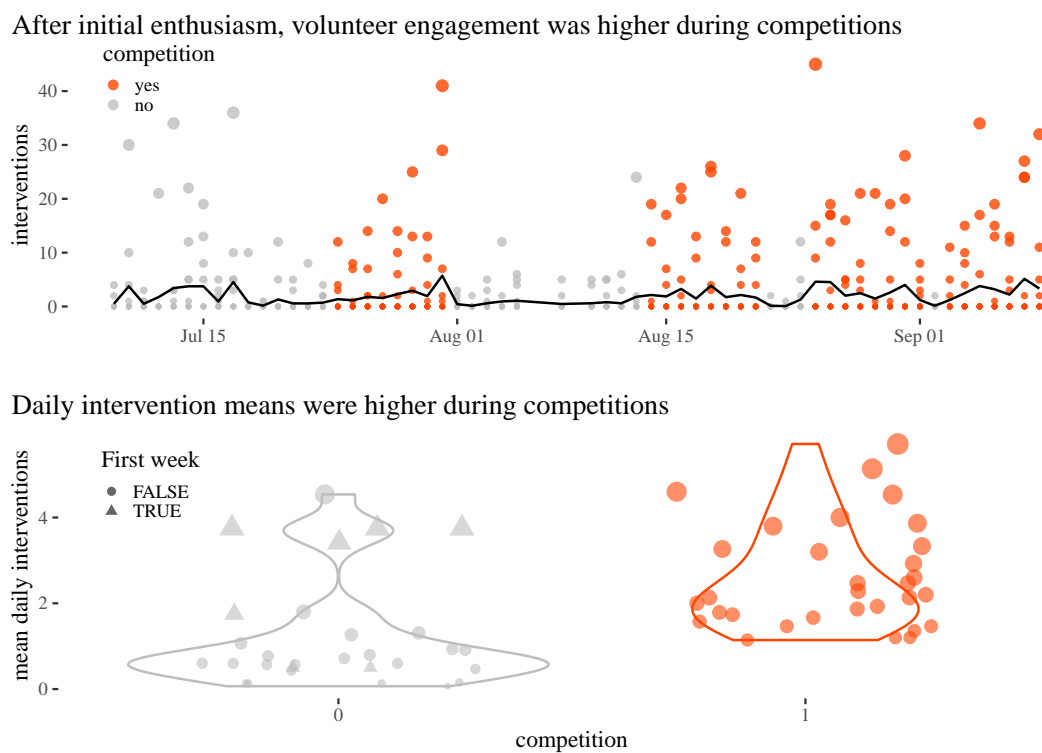


Figure 15: Daily individual volunteer intervention counts across time with competition periods marked (top) and daily group intervention means grouped by whether a competition was ongoing (bottom). Note most of high means in the non-competition period are in the first week.

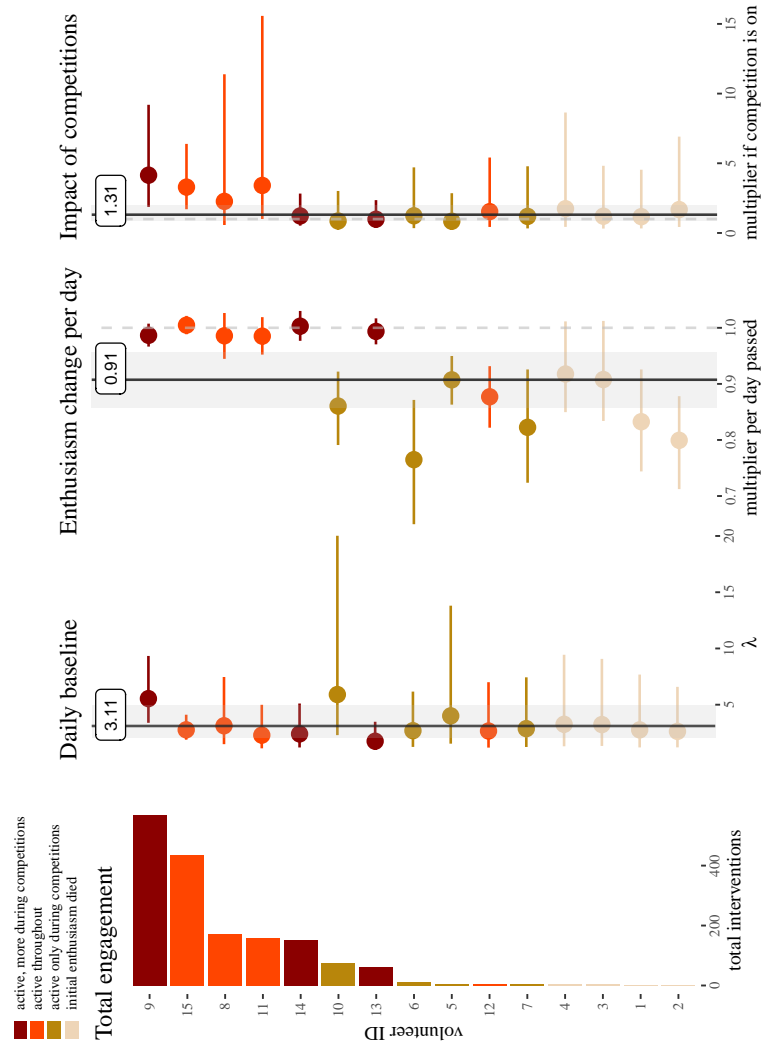


Figure 16: Volunteer total engagement with their daily baseline and multipliers for enthusiasm and impact of competition. Pointranges represent individual level coefficients, group coefficients are represented by black lines with shaded 89% HPDI areas.

Appendix A. Explanation of WAIC

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

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

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:

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

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 λ s. The first Bayesian model has the following structure:

$$\begin{aligned}\text{attacks}_i &\sim \text{Poisson}(\lambda) \\ \log(\lambda) &= l \\ l &\sim \text{Norm}(.05, 2.8)\end{aligned}$$

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

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

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 λ s.

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.

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

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).

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

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

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

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:

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

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).

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

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

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

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.

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

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.

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

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

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

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

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.

$$\begin{aligned}
\text{attacks}_i &\sim \text{NegativeBinomial}(\lambda_i, \phi_{\text{userID}[i]}) \\
\log(\lambda_i) &= l_{\text{userID}[i]} + a_{\text{userID}[i]} \times \text{attacksL1} + c_{\text{userID}[i]} \times \text{act} + \\
&\quad + i1_{\text{control}}_{\text{userID}[i]} \times \text{control} \times \text{intL1D} + \\
&\quad + i1_{\text{emp}}_{\text{userID}[i]} \times \text{emp} \times \text{intL1D} + \\
&\quad + i1_{\text{norm}}_{\text{userID}[i]} \times \text{norm} \times \text{intL1D} \\
l_{\text{userID}[i]} &\sim \text{Norm}(\bar{l}, \bar{\sigma}_l) \\
a_{\text{userID}[i]} &\sim \text{Norm}(\bar{a}, \bar{\sigma}_a) \\
c_{\text{userID}[i]} &\sim \text{Norm}(\bar{c}, \bar{\sigma}_c) \\
i1_{\text{control}}_{\text{userID}[i]} &\sim \text{Norm}(i1_{\text{control}}_{\text{Overall}}, \bar{\sigma}_{i1}) \\
i1_{\text{emp}}_{\text{userID}[i]} &\sim \text{Norm}(i1_{\text{emp}}_{\text{Overall}}, \bar{\sigma}_{i1}) \\
i1_{\text{norm}}_{\text{userID}[i]} &\sim \text{Norm}(i1_{\text{norm}}_{\text{Overall}}, \bar{\sigma}_{i1}) \\
i1_{\text{control}}_{\text{Overall}} &\sim \text{Norm}(0, .2) \\
i1_{\text{emp}}_{\text{Overall}} &\sim \text{Norm}(0, .2) \\
i1_{\text{norm}}_{\text{Overall}} &\sim \text{Norm}(0, .2) \\
\bar{\lambda} &\sim \text{Norm}(.00001, 2.5) \\
\bar{\sigma}_l &\sim \text{Exp}(1.5) \\
\bar{a} &\sim \text{Norm}(0, .2) \\
\bar{\sigma}_a &\sim \text{Exp}(5) \\
\bar{c} &\sim \text{Norm}(0, .2) \\
\bar{\sigma}_c &\sim \text{Exp}(5) \\
\bar{\sigma}_{i1} &\sim \text{Exp}(5)
\end{aligned}$$

Now, let us rethink the priors. The coefficients need to be exponentiated to be understood multiplicatively. For instance, the prior for *ilempOverall* is $\text{Norm}(0, .2)$. To understand what priors for the exponentiated individual coefficients this entails, we can simulate: (1) draw 1e4 values *ilbar* of the mean from $\text{Norm}(0, .2)$, (2) draw 1e4 values *isigmabar* of the standard deviation parameter from $\text{Exp}(5)$, and each time (3) draw 1e4 parameters from $\text{Norm}(\text{ilbar}, \text{isigmabar})$. The resulting distribution looks as in Figure B.17. This is still a very wide prior for the multiplicative impact of empathetic 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 Hamiltonian 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.

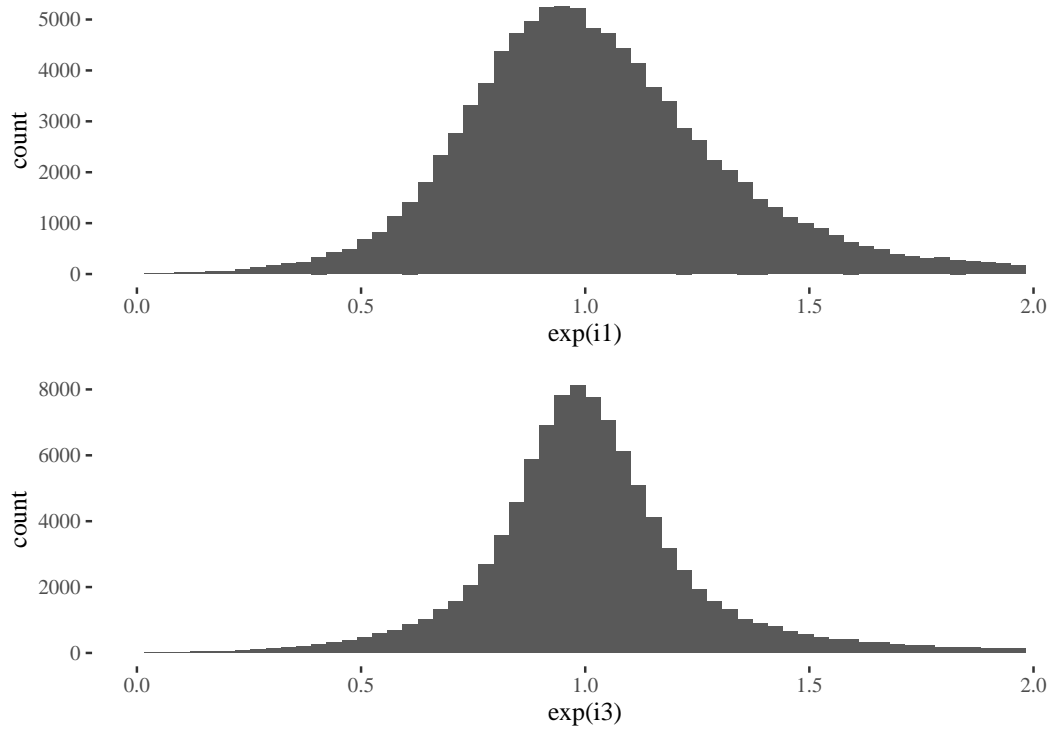


Figure B.17: Simulated priors for the individual $i1$ coefficients, and prior for the cumulative impact model with larger input variability (hence, the prior is more narrow to eliminate unrealistically huge impact).

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

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.

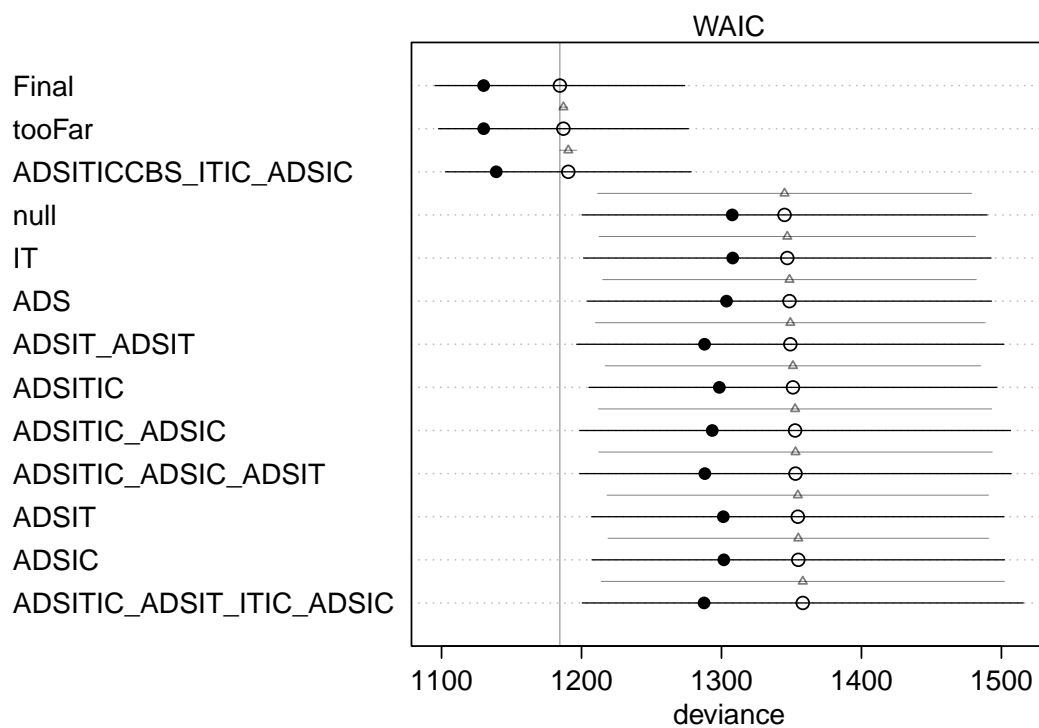


Figure C.18: Model comparison, WAIC scores. The filled points are the in-sample deviance values. The open points are the WAIC values. The line segments represent standard errors of the WAIC scores. really want however is the standard error of the difference in WAIC between the two models. The triangle 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] Z. Laub, Hate Speech on Social Media: Global Comparisons, 2019. URL: <https://www.cfr.org/backgrounder/hate-speech-social-media-global-comparisons>.
- [2] A. Sorrentino, A. C. Baldry, D. P. Farrington, C. Blaya, Epidemiology of cyberbullying across europe: Differences between countries and genders, Educational Sciences: Theory & Practice 19 (2019).
- [3] E. A. Vogels, The State of Online Harassment, 2021. URL: <https://www.pewresearch.org/internet/2021/01/13/the-state-of-online-harassment/>.
- [4] C. P. Barlett, M. M. Simmers, B. Roth, D. Gentile, Comparing cyberbullying prevalence and process before and during the covid-19 pandemic, The Journal of Social Psychology (2021) 1–11.
- [5] J. Grant, Australia's esafety commissioner targets abuse online as covid-19 supercharges cyberbullying — the strategist, 2021. URL: <https://www.aspistrategist.org.au/australias-esafety-commissioner-targets-abuse-online-as-covid-19-supercharges>
- [6] L1ght, L1ght releases groundbreaking report on corona-related hate speech and online toxicity, 2020. URL: <https://l1ght.com/l1ght-releases-groundbreaking-report-on-corona-related-hate-speech-and-online>
- [7] A. Bhattacharya, How Covid-19 lockdowns weakened Facebook's content moderation algorithms, 2021. URL: <https://qz.com/india/1976450/facebook-covid-19-lockdowns-hurt-content-moderation-algorithms/>.

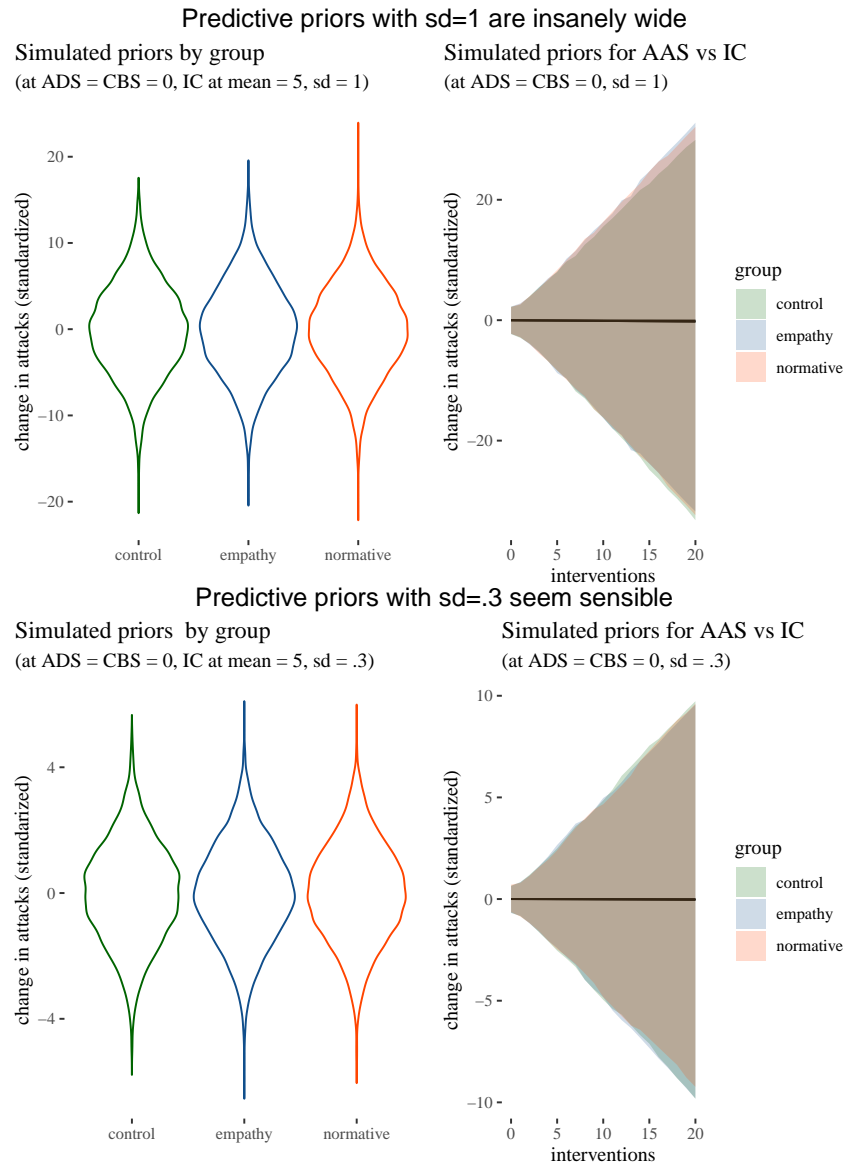


Figure C.19: Prior predictive check for two different sets of priors.

- [8] Y. Gerrard, ¿? covid19?¿ the covid-19 mental health content moderation conundrum, *Social Media+ Society* 6 (2020) 2056305120948186.
- [9] B. Yang, B. Wang, N. Sun, F. Xu, L. Wang, J. Chen, S. Yu, Y. Zhang, Y. Zhu, T. Dai, et al., The consequences of cyberbullying and traditional bullying victimization among adolescents: gender differences in psychological symptoms, self-harm and suicidality, *Psychiatry research* 306 (2021) 114219.
- [10] S. Hinduja, J. W. Patchin, Bullying, cyberbullying, and suicide, *Archives of suicide research* 14 (2010) 206–221.
- [11] A. Alhujailli, W. Karwowski, T. T. Wan, P. Hancock, Affective and stress consequences of cyberbullying, *Symmetry* 12 (2020) 1536.
- [12] R. M. Kowalski, S. P. Limber, Psychological, physical, and academic correlates of cyberbullying and traditional bullying, *Journal of adolescent health* 53 (2013) S13–S20.
- [13] E. Lee, L. Leets, Persuasive storytelling by hate groups online: Examining its effects on adolescents, *American behavioral scientist* 45 (2002) 927–957.
- [14] M. Williams, Hatred behind the screens: A report on the rise of online hate speech (2019).
- [15] C. Fink, Dangerous speech, anti-muslim violence, and facebook in myanmar, *Journal of International Affairs* 71 (2018) 43–52.
- [16] W. Soral, M. Bilewicz, M. Winiewski, Exposure to hate speech increases prejudice through desensitization, *Aggressive behavior* 44 (2018) 136–146. doi:10.1002/ab.21737.
- [17] A. Álvarez-Benjumea, F. Winter, Normative change and culture of hate: An experiment in online environments, *European Sociological Review* 34 (2018) 223–237. doi:10.2139/ssrn.3126089.
- [18] M. L. Williams, P. Burnap, A. Javed, H. Liu, S. Ozalp, 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 (2020) 93–117. doi:10.1093/bjc/azz049.

- [19] R. Urbaniak, M. Ptasiński, P. Tempka, G. Leliwa, M. Brochocki, M. Wroczyński, Personal attacks decrease user activity in social networking platforms, *Computers in Human Behavior* 126 (2022) 106972.
- [20] H. J. Tworek, Fighting hate with speech law: Media and german visions of democracy, *The Journal of Holocaust Research* 35 (2021) 106–122.
- [21] M. Trengove, E. Kazim, D. Almeida, A. Hilliard, S. Zannone, E. Lomas, A critical review of the online safety bill, *Patterns* (2022) 100544.
- [22] Facebook frustrates advertisers as boycott over hate speech kicks off, 2020. URL: <https://www.cnbc.com/2020/07/01/facebook-frustrates-advertisers-as-boycott-over-hate-speech-kicks-off.html>.
- [23] C. Newton, Killer app: Why do anonymous networks keep leading to suicides?, 2013. URL: <https://www.theverge.com/2013/9/17/4740902/no-good-answers-why-didnt-ask-fm-learn-from-the-formspring-suicides>.
- [24] Failory, Yik yak’s shut down: Why did the location-based app failed?, 2022. URL: <https://www.failory.com/cemetery/yik-yak>.
- [25] A teen who was bullied on snapchat died. his mom is suing to hold social media liable, 2021. URL: <https://www.latimes.com/business/story/2021-05-10/lawsuit-snap-teen-suicide-yolo-lmk>.
- [26] D. Seifert, Social question and answer site formspring to shut down on March 31st, 2013. URL: <https://www.theverge.com/2013/3/15/4110196/social-question-answer-site-formspring-shut-down-march-31st>.
- [27] S. MacAvaney, H.-R. Yao, E. Yang, K. Russell, N. Goharian, O. Frieder, Hate speech detection: Challenges and solutions, *PloS one* 14 (2019) e0221152.
- [28] A. Schmidt, M. Wiegand, A survey on hate speech detection using natural language processing, in: *Proceedings of the fifth international workshop on natural language processing for social media*, 2017, pp. 1–10.

- [29] M. I. Jordan, T. M. Mitchell, Machine learning: Trends, perspectives, and prospects, *Science* 349 (2015) 255–260.
- [30] Y. LeCun, Y. Bengio, G. Hinton, Deep learning, *nature* 521 (2015) 436–444.
- [31] T. J. Sejnowski, The unreasonable effectiveness of deep learning in artificial intelligence, *Proceedings of the National Academy of Sciences* 117 (2020) 30033–30038.
- [32] R. Binns, M. Veale, M. Van Kleek, N. Shadbolt, Like trainer, like bot? inheritance of bias in algorithmic content moderation, in: *International conference on social informatics*, Springer, 2017, pp. 405–415.
- [33] M. Geva, Y. Goldberg, J. Berant, Are we modeling the task or the annotator? an investigation of annotator bias in natural language understanding datasets, *arXiv preprint arXiv:1908.07898* (2019).
- [34] N. Mehrabi, F. Morstatter, N. Saxena, K. Lerman, A. Galstyan, A survey on bias and fairness in machine learning, *ACM Computing Surveys (CSUR)* 54 (2021) 1–35.
- [35] M. Sap, D. Card, S. Gabriel, Y. Choi, N. A. Smith, 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, 2019, pp. 1668–1678. URL: <https://aclanthology.org/P19-1163>. doi:10.18653/v1/P19-1163.
- [36] J. Vanian, Google’s Hate Speech Detection A.I. Has a Racial Bias Problem, 2019. URL: <https://fortune.com/2019/08/16/google-jigsaw-perspective-racial-bias/>.
- [37] T. Gröndahl, L. Pajola, M. Juuti, M. Conti, N. Asokan, All you need is” love” evading hate speech detection, in: *Proceedings of the 11th ACM workshop on artificial intelligence and security*, 2018, pp. 2–12.
- [38] H. Rosa, N. Pereira, R. Ribeiro, P. C. Ferreira, J. P. Carvalho, S. Oliveira, L. Coheur, P. Paulino, A. V. Simão, I. Trancoso, Automatic cyberbullying detection: A systematic review, *Computers in Human Behavior* 93 (2019) 333–345.

- [39] W. Yin, A. Zubiaga, Towards generalisable hate speech detection: a review on obstacles and solutions, *PeerJ Computer Science* 7 (2021) e598.
- [40] T. Wu, M. T. Ribeiro, J. Heer, D. S. Weld, Errudite: Scalable, reproducible, and testable error analysis, in: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 2019, pp. 747–763.
- [41] Z. C. Lipton, J. Steinhardt, Troubling trends in machine learning scholarship: Some ml papers suffer from flaws that could mislead the public and stymie future research., *Queue* 17 (2019) 45–77.
- [42] K. Musgrave, S. Belongie, S.-N. Lim, A metric learning reality check, in: *European Conference on Computer Vision*, Springer, 2020, pp. 681–699.
- [43] J. Steinhardt, P. W. Koh, P. Liang, Certified defenses for data poisoning attacks, in: *Proceedings of the 31st International Conference on Neural Information Processing Systems*, 2017, pp. 3520–3532.
- [44] G. Melis, C. Dyer, P. Blunsom, On the state of the art of evaluation in neural language models, *arXiv preprint arXiv:1707.05589* (2017).
- [45] J. Steinhardt, P. Liang, Learning fast-mixing models for structured prediction, in: *International Conference on Machine Learning*, PMLR, 2015, pp. 1063–1072.
- [46] A. Esteva, B. Kuprel, R. A. Novoa, J. Ko, S. M. Swetter, H. M. Blau, S. Thrun, Dermatologist-level classification of skin cancer with deep neural networks, *Nature* 542 (2017) 115–118. URL: <http://www.nature.com/articles/nature21056>. doi:10.1038/nature21056.
- [47] P. Röttger, B. Vidgen, D. Nguyen, Z. Waseem, H. Margetts, J. Pierrehumbert, Hatecheck: Functional tests for hate speech detection models, *arXiv preprint arXiv:2012.15606* (2020).
- [48] About the online safety data initiative, ????
- [49] Facebook: 3.2 Billion Likes & Comments Every Day, 2012. URL: <https://martech.org/facebook-3-2-billion-likes-comments-every-day/>.

- [50] 10 Twitter Statistics Every Marketer Should Know in 2021 [Infographic], 2021. URL: <https://www.oberlo.com/blog/twitter-statistics>.
- [51] Reddit in 2020, 2020. URL: https://www.reddit.com/r/blog/comments/k967mm/reddit_in_2020/.
- [52] S. T. Roberts, Behind the screen: The hidden digital labor of commercial content moderation, University of Illinois at Urbana-Champaign, 2014.
- [53] S. T. Roberts, Commercial content moderation: Digital laborers' dirty work (2016).
- [54] S. Karunakaran, R. Ramakrishnan, Testing stylistic interventions to reduce emotional impact of content moderation workers, in: Proceedings of the AAAI Conference on Human Computation and Crowdsourcing, volume 7, 2019, pp. 50–58.
- [55] M. Steiger, T. J. Bharucha, S. Venkatagiri, M. J. Riedl, M. Lease, The psychological well-being of content moderators, in: Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems, CHI, volume 21, 2021.
- [56] C. Newton, Facebook will pay \$52 million in settlement with moderators who developed PTSD on the job, 2020. URL: <https://www.theverge.com/2020/5/12/21255870/facebook-content-moderator-settlement-scola-ptsd-mental-health>.
- [57] L. Parks, Dirty data: content moderation, regulatory outsourcing, and the cleaners, *Film Quarterly* 73 (2019) 11–18.
- [58] I. Stepanikova, Racial-ethnic biases, time pressure, and medical decisions, *Journal of health and social behavior* 53 (2012) 329–343.
- [59] J. Koetsier, Report: Facebook Makes 300,000 Content Moderation Mistakes Every Day, 2020. URL: <https://www.forbes.com/sites/johnkoetsier/2020/06/09/300000-facebook-content-moderation-mistakes-daily-report-says/>.
- [60] D. L. Hoff, S. N. Mitchell, Cyberbullying: Causes, effects, and remedies, *Journal of Educational Administration* (2009).

- [61] T. Keipi, M. Näsi, A. Oksanen, P. Räsänen, Online hate and harmful content: Cross-national perspectives, Taylor & Francis, 2016.
- [62] S. Wachs, M. F. Wright, R. Sittichai, R. Singh, R. Biswal, E.-m. Kim, S. Yang, M. Gámez-Guadix, C. Almendros, K. Flora, et al., 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 (2019) 3992.
- [63] C. French, As the pandemic forces us online, LGBTQ2S+ teens deal with cyberbullying, 2021. URL: <https://www.ctvnews.ca/canada/as-the-pandemic-forces-us-online-lgbtq2s-teens-deal-with-cyberbullying-1.5430945>.
- [64] A.-D. League, Free to Play? Hate, Harassment and Positive Social Experience in Online Games 2020, 2020. URL: <https://www.adl.org/free-to-play-2020>.
- [65] In:Game Abuse, 2017. URL: <https://www.ditchthelabel.org/research-papers/ingame-abuse/>.
- [66] J. Leader Maynard, S. Benesch, Dangerous speech and dangerous ideology: An integrated model for monitoring and prevention, *Genocide Studies and Prevention* 9 (2016).
- [67] J. Garland, K. Ghazi-Zahedi, J.-G. Young, L. Hébert-Dufresne, M. Galesic, Countering hate on social media: Large scale classification of hate and counter speech, *arXiv preprint arXiv:2006.01974* (2020).
- [68] M.-W. Dictionary, Merriam-webster, On-line at [http://www. mw. com/home. htm](http://www.mw.com/home.htm) 8 (2002) 2.
- [69] C. Schieb, M. Preuss, Governing hate speech by means of counterspeech on facebook, in: 66th ica annual conference, at fukuoka, japan, 2016, pp. 1–23.
- [70] S. Benesch, D. Ruths, K. P. Dillon, H. M. Saleem, L. Wright, Considerations for successful counterspeech, *Dangerous Speech Project* (2016).
- [71] D. Rieger, J. B. Schmitt, L. Frischlich, Hate and counter-voices in the internet: Introduction to the special issue, *Studies in Communication and Media* (2018) 459–472.

- [72] M. Ziegele, P. Jost, M. Bormann, D. Heinbach, 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 (2018) 525–554.
- [73] C. Ziems, B. He, S. Soni, S. Kumar, Racism is a virus: Anti-asian hate and counterhate in social media during the covid-19 crisis, *arXiv preprint arXiv:2005.12423* (2020).
- [74] J. Xie, S. Sreenivasan, G. Korniss, W. Zhang, C. Lim, B. K. Szymanski, Social consensus through the influence of committed minorities, *Physical Review E* 84 (2011) 011130.
- [75] E. Chandrasekharan, U. Pavalanathan, A. Srinivasan, A. Glynn, J. Eisenstein, E. Gilbert, 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 (2017) 1–22.
- [76] M. Bilewicz, P. Tempuska, G. Leliwa, M. Dowgiałło, M. Tańska, R. Urbaniak, M. Wroczyński, Artificial intelligence against hate: Intervention reducing verbal aggression in the social network environment, *Aggressive Behavior* 47 (2021) 260–266. URL: <https://onlinelibrary.wiley.com/doi/abs/10.1002/ab.21948>. doi:<https://doi.org/10.1002/ab.21948>. arXiv:<https://onlinelibrary.wiley.com/doi/pdf/10.1002/ab.21948>.
- [77] M. Brauer, N. Chaurand, Descriptive norms, prescriptive norms, and social control: An intercultural comparison of people’s reactions to uncivil behaviors, *European Journal of Social Psychology* 40 (2010) 490–499.
- [78] P. Fischer, J. I. Krueger, T. Greitemeyer, C. Vogrincic, A. Kastenmüller, D. Frey, M. Heene, M. Wicher, M. Kainbacher, The bystander-effect: a meta-analytic review on bystander intervention in dangerous and non-dangerous emergencies., *Psychological bulletin* 137 (2011) 517.
- [79] L. Legault, J. N. Gutsell, M. Inzlicht, Ironic effects of antiprejudice messages: How motivational interventions can reduce (but also increase) prejudice, *Psychological Science* 22 (2011) 1472–1477.

- [80] A. Morris, F. Cushman, A common framework for theories of norm compliance, *Social Philosophy and Policy* 35 (2018) 101–127.
- [81] R. B. Cialdini, R. R. Reno, C. A. Kallgren, A focus theory of normative conduct: Recycling the concept of norms to reduce littering in public places., *Journal of personality and social psychology* 58 (1990) 1015.
- [82] R. B. Cialdini, L. J. Demaine, B. J. Sagarin, D. W. Barrett, K. Rhoads, P. L. Winter, Managing social norms for persuasive impact, *Social influence* 1 (2006) 3–15.
- [83] N. J. Goldstein, R. B. Cialdini, V. Griskevicius, A room with a viewpoint: Using social norms to motivate environmental conservation in hotels, *Journal of consumer Research* 35 (2008) 472–482.
- [84] C. A. Kallgren, R. R. Reno, R. B. Cialdini, A focus theory of normative conduct: When norms do and do not affect behavior, *Personality and social psychology bulletin* 26 (2000) 1002–1012.
- [85] R. B. Cialdini, N. J. Goldstein, Social influence: Compliance and conformity, *Annual review of psychology* 55 (2004) 591–621.
- [86] C. Stangor, G. B. Sechrist, J. T. Jost, Changing racial beliefs by providing consensus information, *Personality and Social Psychology Bulletin* 27 (2001) 486–496.
- [87] C. S. Crandall, A. Eshleman, L. O’Brien, Social norms and the expression and suppression of prejudice: the struggle for internalization., *Journal of personality and social psychology* 82 (2002) 359.
- [88] K. Munger, Tweetment effects on the tweeted: Experimentally reducing racist harassment, *Political Behavior* 39 (2017) 629–649. doi:10.1007/s11109-016-9373-5.
- [89] V. Price, L. Nir, J. N. Cappella, Normative and informational influences in online political discussions, *Communication Theory* 16 (2006) 47–74.
- [90] J. Cheng, M. Bernstein, C. Danescu-Niculescu-Mizil, J. Leskovec, 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*, 2017, pp. 1217–1230.

- [91] R. B. Cialdini, C. A. Kallgren, R. R. Reno, A Focus Theory of Normative Conduct: A Theoretical Refinement and Reevaluation of the Role of Norms in Human Behavior, in: *Advances in Experimental Social Psychology*, volume 24, Elsevier, 1991, pp. 201–234. URL: <https://linkinghub.elsevier.com/retrieve/pii/S0065260108603305>. doi:10.1016/S0065-2601(08)60330-5.