# 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. Data analysis

### *1.1. Exploration and data*

For the duration of the project we selected 486 Reddit users and tracked their activity (with some breaks resulting from API restrictions and technical issues, which were mostly sorted out in the observation period), starting on 2020-03-09, beginning the intervention period on 2020-07-08, leading to a further observation period starting on 2020-09-09 and ending on 2020-11-20. The time series of attacks observed and of interventions conducted can be inspected in Figure 1, along with a quick search for weekly patterns. Some of the users turned out to be bots, a few ceased to be active during the experiment (with no strong reason to think this happened due to them receiving an intervention) and a few received treatment of two different types by accident (we relied on multiple volunteers and such mistakes were likely to happen). Ultimately, in the time series data, we ended up with data on 440 users.

how?

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 total 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.[1] The distribution of IC in the treatment groups is visualized in Figure 2. Note that the distributions are somewhat different, even though the total intervention counts are similar. The issue is discussed in Section XXXXX.

add ref

Second, in the distribution of standardized difference in attacks, the peaks of distributions are shifted a bit between the groups, with lowest median for the normative group, but the differences seem minor (Figure 3). This might suggest no impact of the interventions. This conclusion would be too hasty, as the impact of other predictor variables and interactions involved can mask actual associations.

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

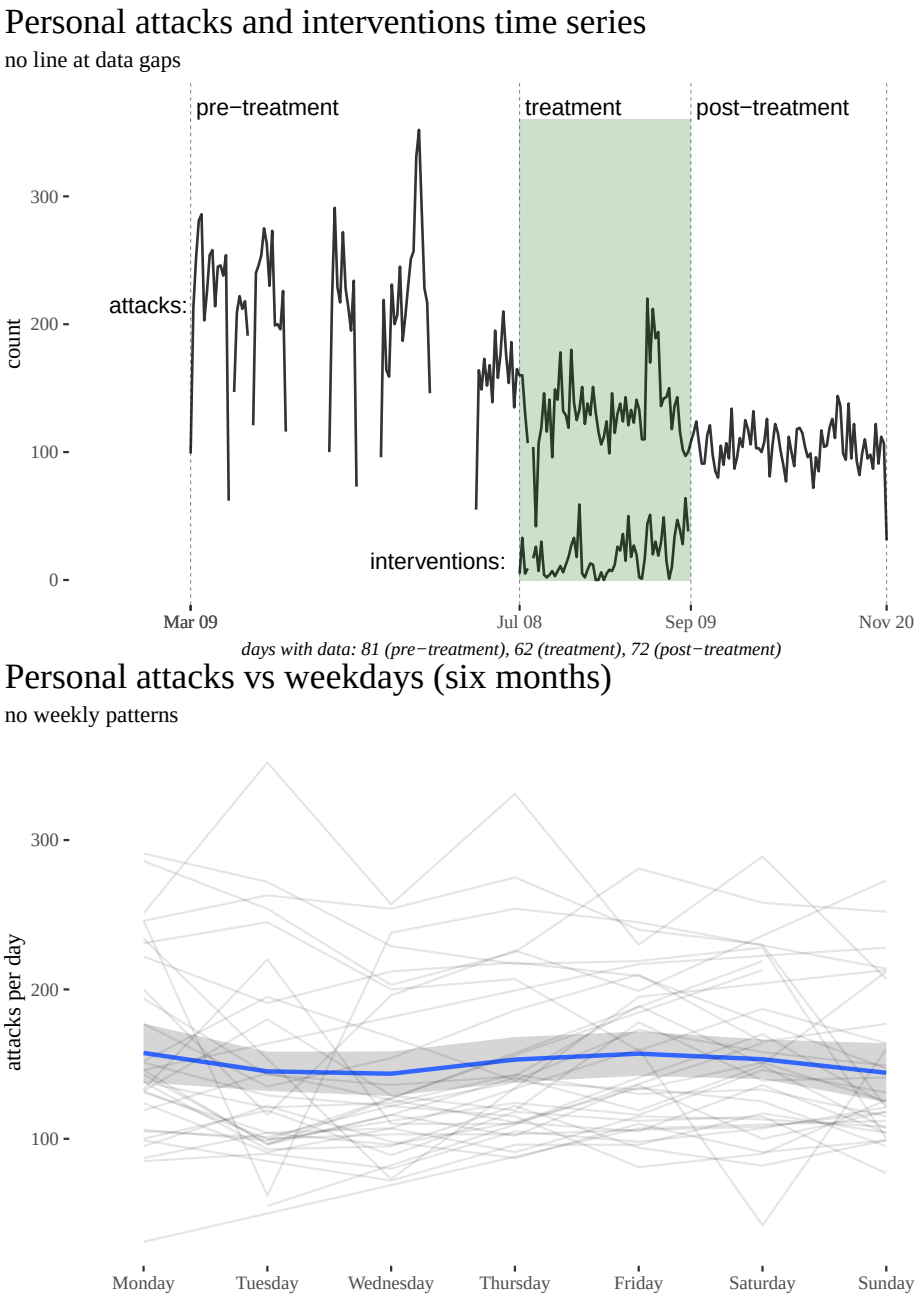


Figure 1: Daily sums of attacks and interventions throughout the three experimental periods, with GAM smoothing (left) and daily attack sums from all weeks in the experimental period plotted against week days (right)—no pattern seems to arise.

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

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

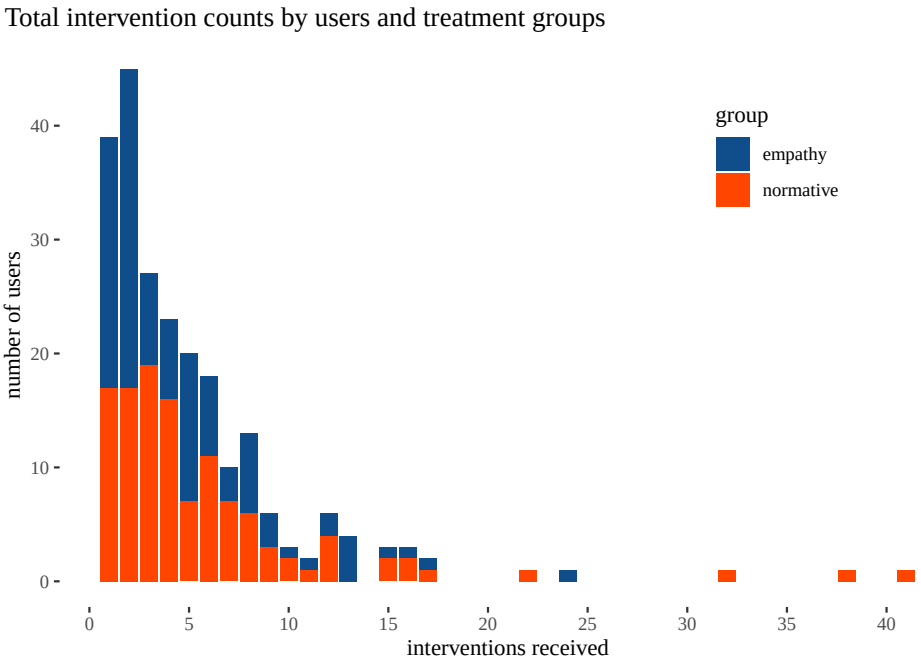


Figure 2: Distribution of daily interventions (by treatment group).

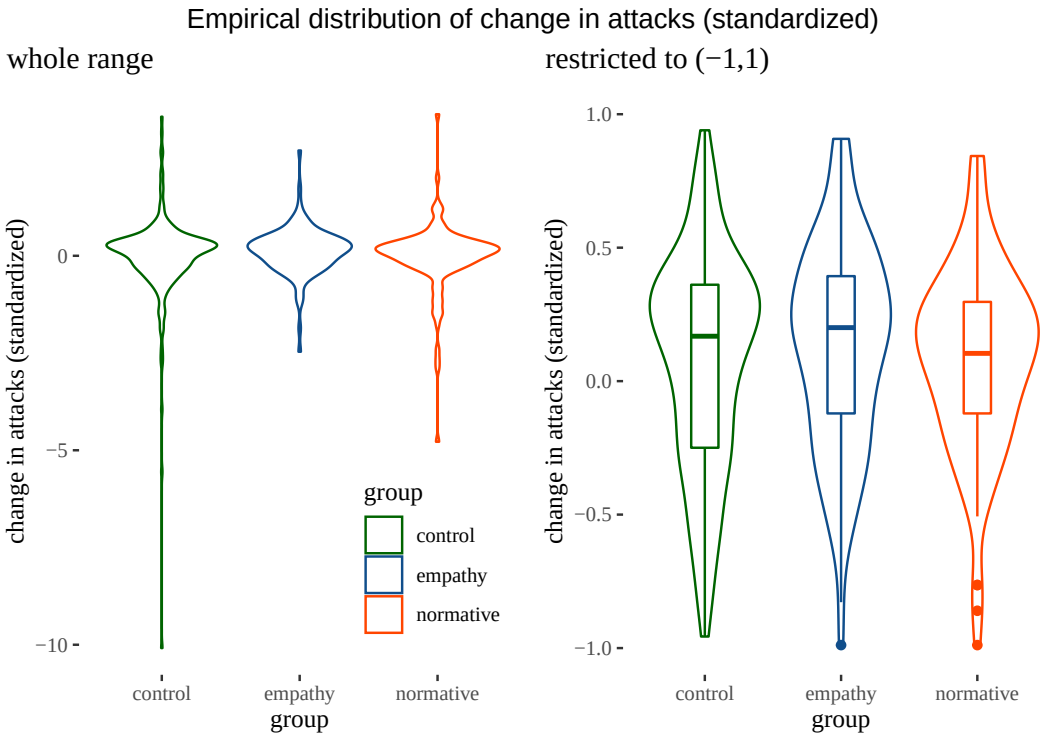


Figure 3: Empirical distribution of change in attacks (by treatment group).

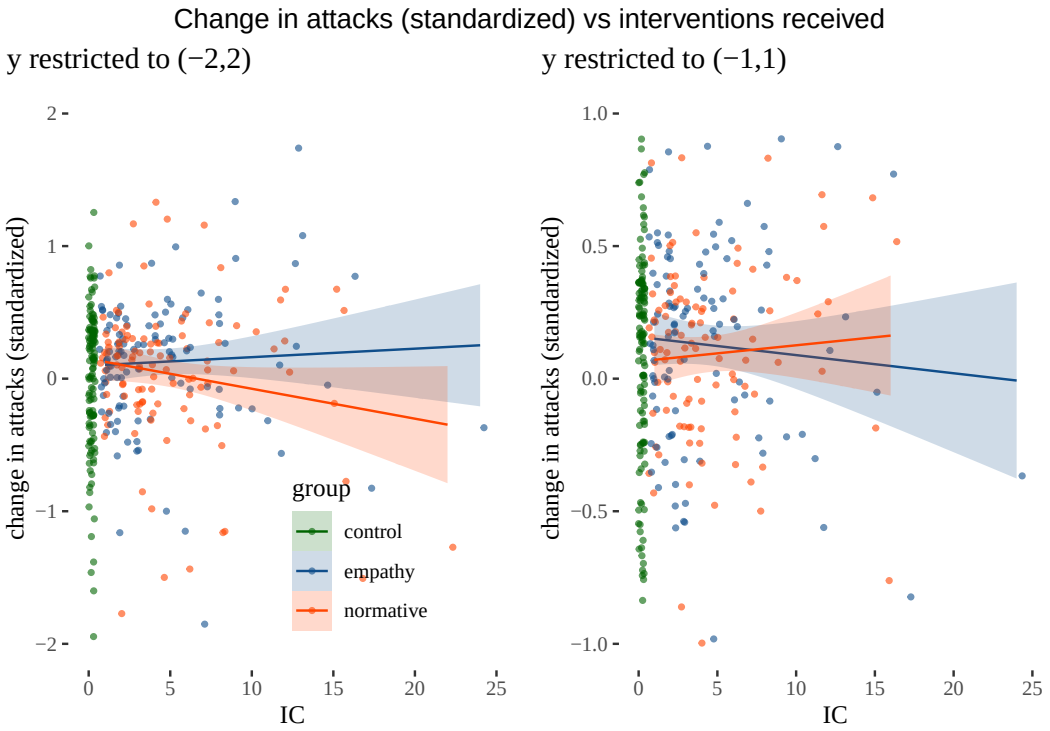


Figure 4: Change in attacks vs the number of interventions received by treatment group, jittered with linear smoothing.

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 4). 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 3, can be obtained by visualizing individual time series. Figure 5 contains six fairly typical examples.

The general phenomenon is that while in the control group attacks tend not to diminish, unless activity itself diminishes, they tend to diminish in the normative group (although the more aggressive the user is, the less of

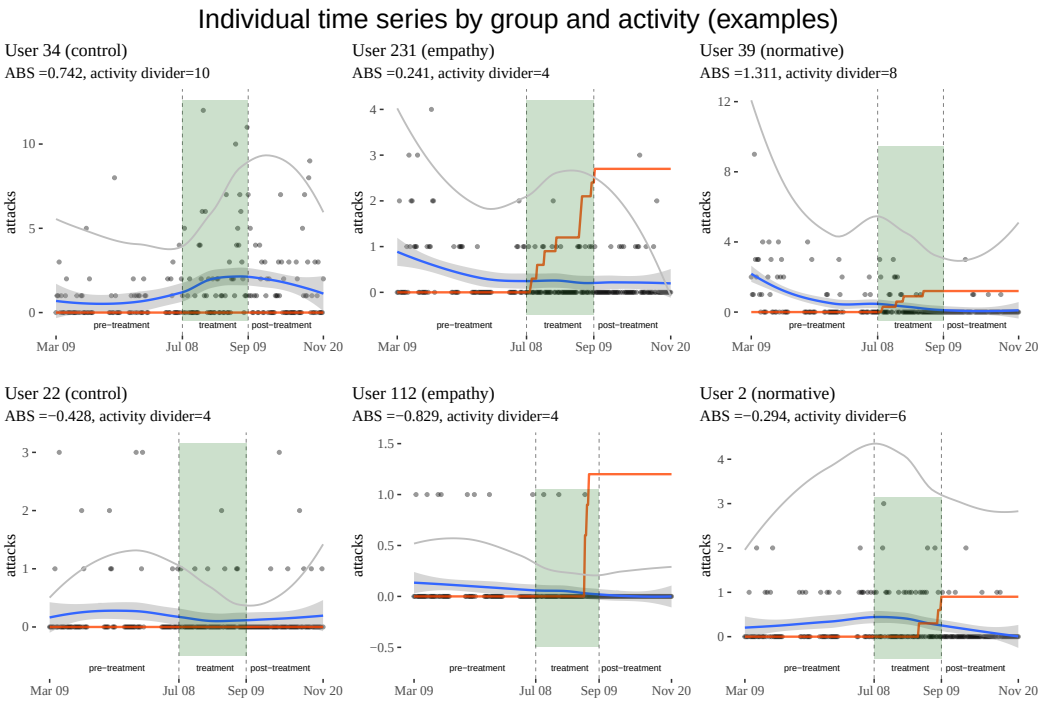


Figure 5: Examples of individual time series. Black points are attacks (smoothed in blue), red lines represent the cumulative number of interventions received (lag 3) divided by 10, gray lines represent overall activity level divided by a variable divider listed in the subtitle. Divisions introduced for visual comparability of general trends.

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.

*1.2. Causal thinking, choice of variables and models*

First, we inspect correlations between predictors to avoid multicolinearity, as highly correlated predictors do not improve predictive performance and artificially inflate uncertainty in their corresponding coefficients in the models. We then develop a plausible causal model of the situation (Figure 6). It turns out that to avoid multicolinearity we cannot condition on CDS if we condition on CAS or CBS. 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 and attacks after directly. Intervention count impacts attacks after and comments after. The same directions of impact are included for intervention type. Finally, comments through time are connected causally, and so are attacks. The structure for the time series data is analogous, except now instead of before and after, we have multiple daily indices.

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

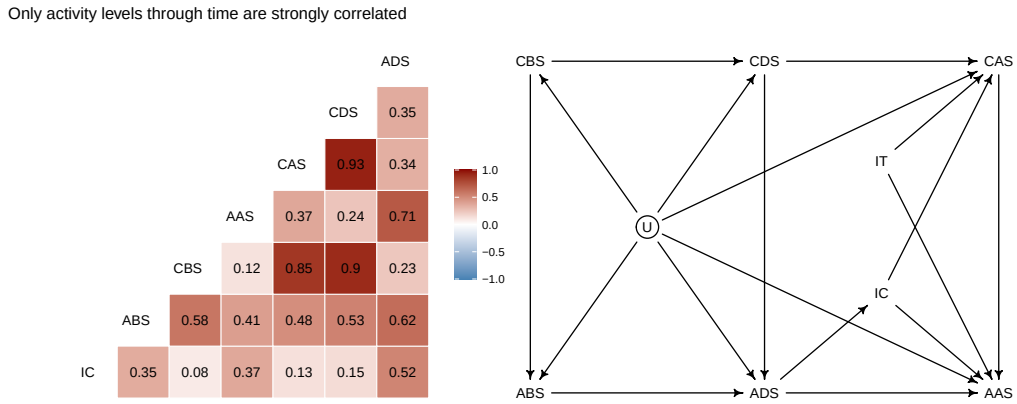


Figure 6: Correlations between predictors (left) and a plausible causal model (right) used in the before-and-after analysis.

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, $k \in \{1, \ldots, 7\}$). The situation, however, is somewhat different for the impact of the total number of cumulative interventions received so far. The trouble is, for example, that if a user received so far a number of interventions until yesterday, some of them had been received before yesterday and those had already impacted their aggression level yesterday. In other words, conditioning on lagged attacks leads to the post-treatment bias and should be avoided.[2]

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

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:

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

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

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

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

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

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

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

coming from a general distribution $\mathsf{Norm}(\bar{c}, \bar{\sigma_c})$ whose parameters are also to be estimated.

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

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

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

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

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

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

The models employing the second formula were superior in performance. It is not surprising that once attacks on a given day were removed from predictor, the overall aggression levels in the before period became predictive.

The price to pay, however, is that now to obtain a user-specific multiplicative interpretation of the impact of cumulative interventions, we need to put the two elements together while multiplying one by the user's overall aggression and only then exponentiate, that is we need to inspect, for instance, $exp(ic3emp_{\mathsf{userID[i]}} + icabst3emp_{\mathsf{userID[i]}} \times \mathsf{abst}[i])$, instead of simply looking at $exp(ic3emp_{\mathsf{userID[i]}})$.

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

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

That is, we take the resulting mean to be the result of the general average ($\alpha$) 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 impact on how the number of attacks during the treatment are 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 impact on the role played by the intervention count.

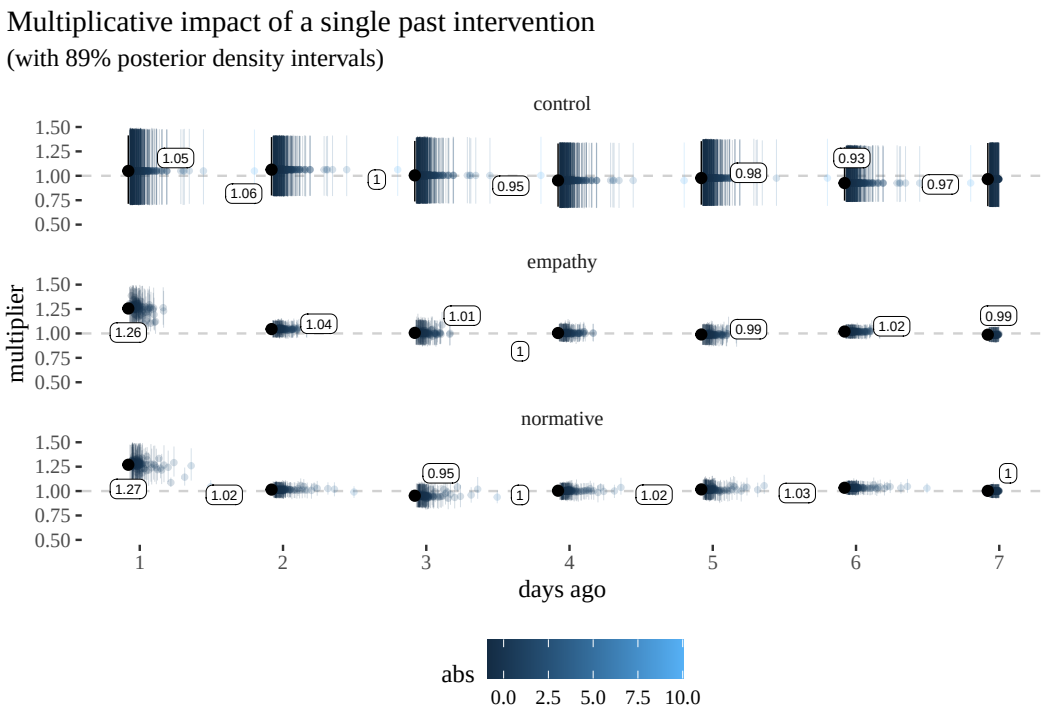


Figure 7: Impact of interventions received lag $1 \leq k \leq 7$ on attacks on a given day. High-level coefficients are pictured in black.

### *1.3. Results*

#### *1.3.1. 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 7 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.

#### *1.3.2. 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 8.

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.

*1.3.3. Long term before/after analysis*

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

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

Grouped before-and-after predicted change of attacks by the levels we just listed are visualized in Figure 10. 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 11), then versus ADS (attacks during, standardized, Figure 12).

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

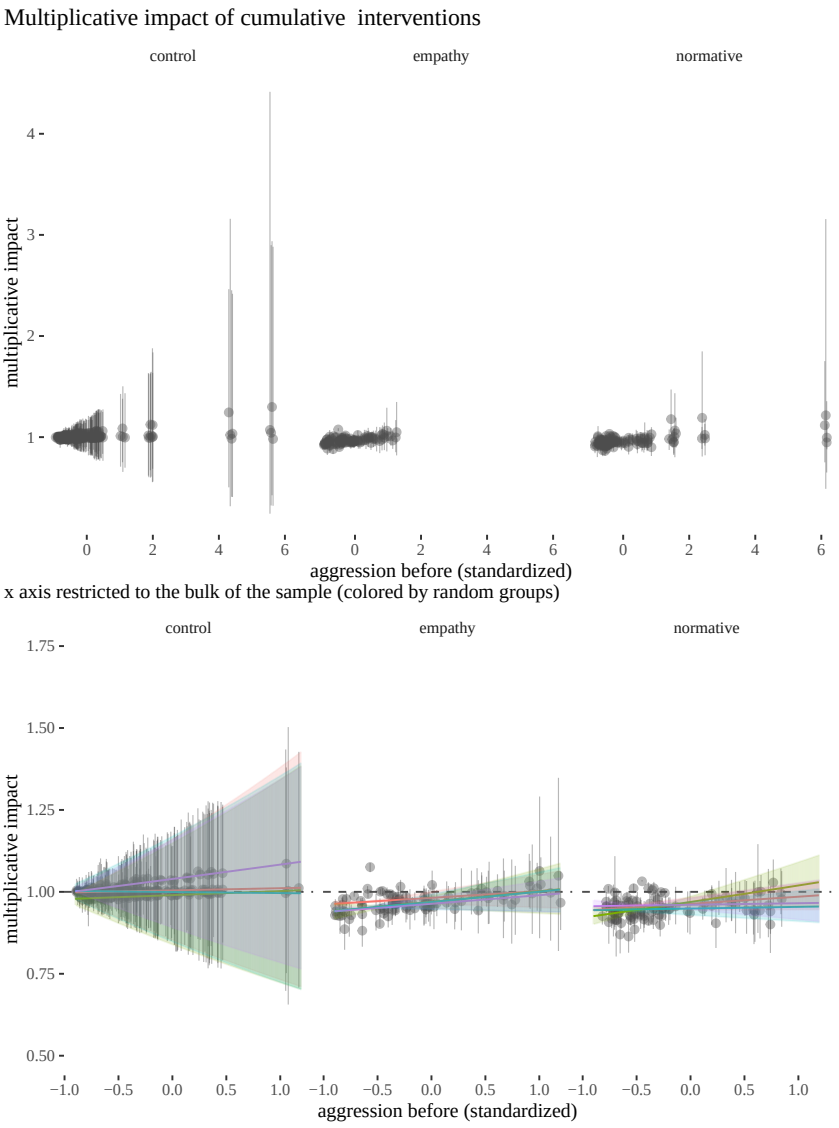


Figure 8: Multiplicative impact of cumulative interventions lag 3 on attacks. Individual users' coefficents only, full range (top), and with attention restricted to the bulk of the sample. Sub-sample coefficients depend on aggression and are represented as lines, colored by sub-sample. Note low number and high uncertainty for highly aggressive users, which motivate the restriction of the $x$ axis for inspection.

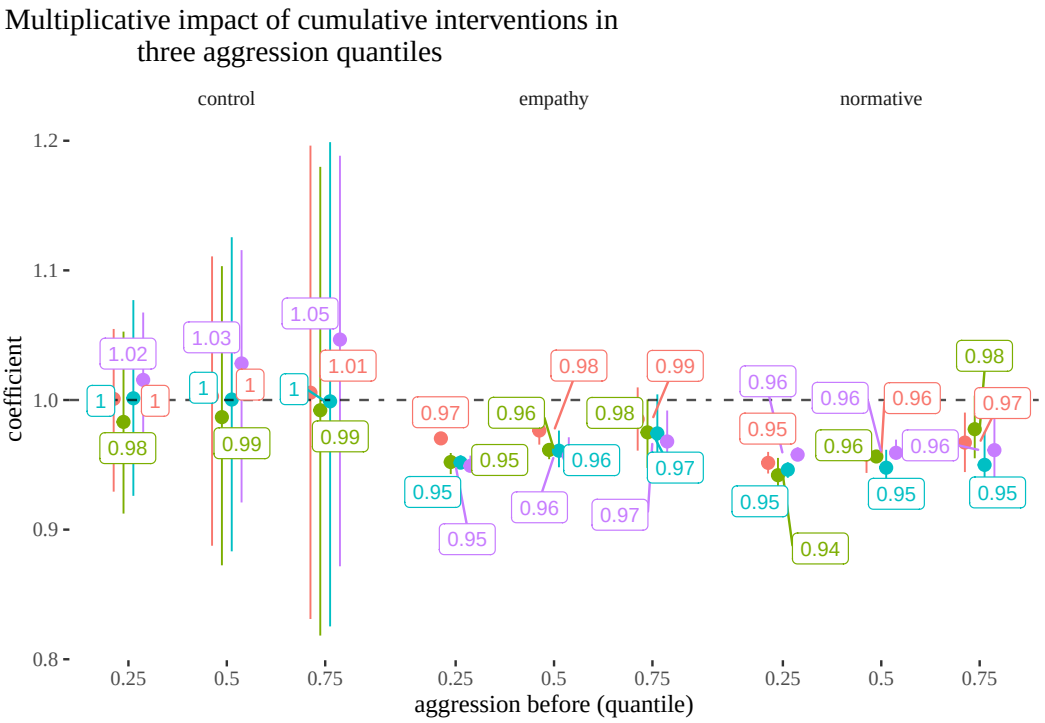


Figure 9: Multiplicative impact of cumulative interventions lag 3 on attacks. General level coefficents only, in three quantiles (.25, .5, .75). Colored by sub-sample.

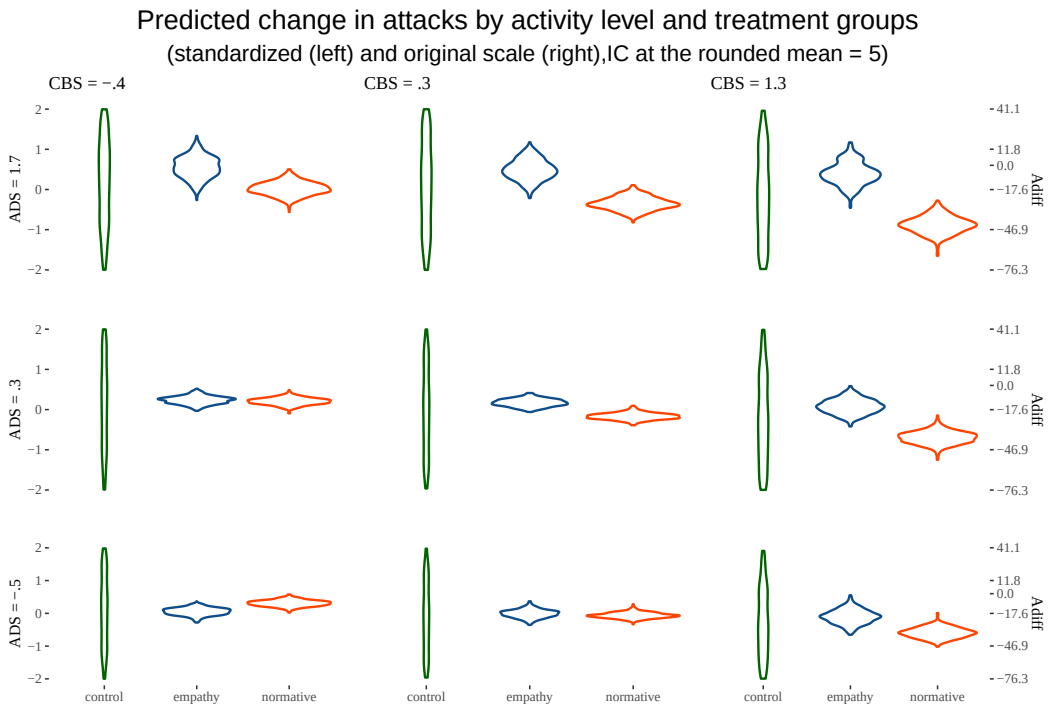


Figure 10: Predicted change in attacks, depending on user's activity level (CBS: comments before, standardized) and how aggressive overall they were (ADS: attacks during, standardized). The more aggressive and active the users, the higher the attacks drop in the normative group, slight drop correlated with emotive interventions for not too active users.

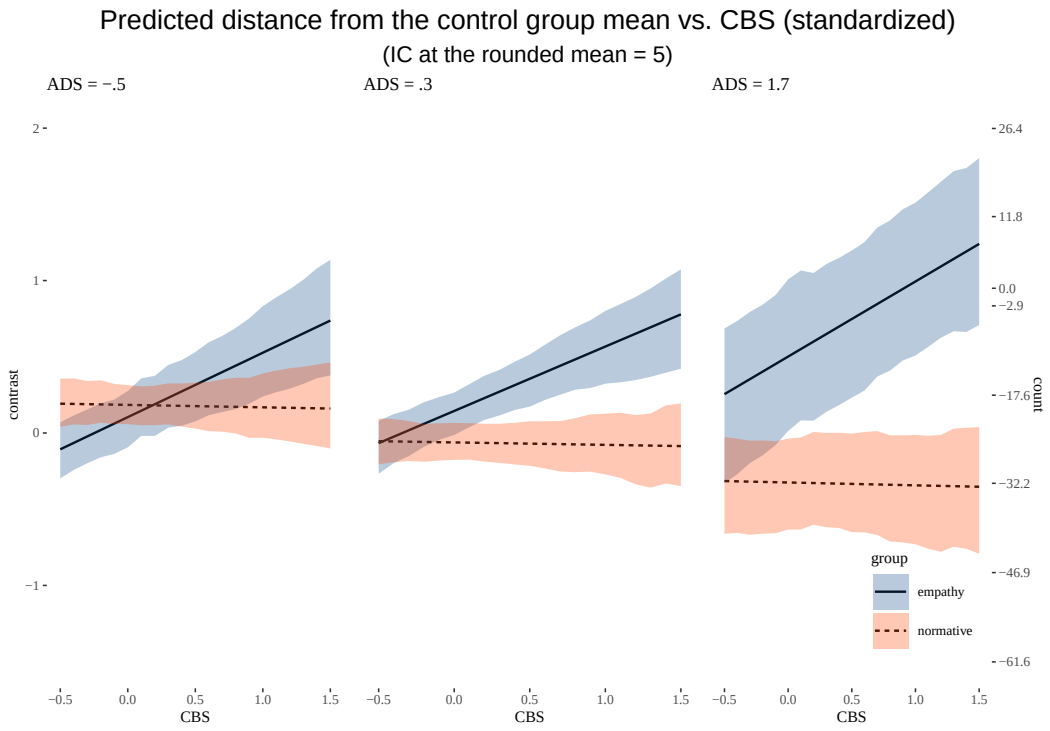


Figure 11: Predicted contrasts (difference in attacks as compared to the control group) for the two treatment groups vs activity before the treatment. Notice that empathetic interventions correlated with decreased attacks for less active users, but performed worse than normative interventions for more active users. Normative interventions, in contrast, seem to have better impact on more active users.

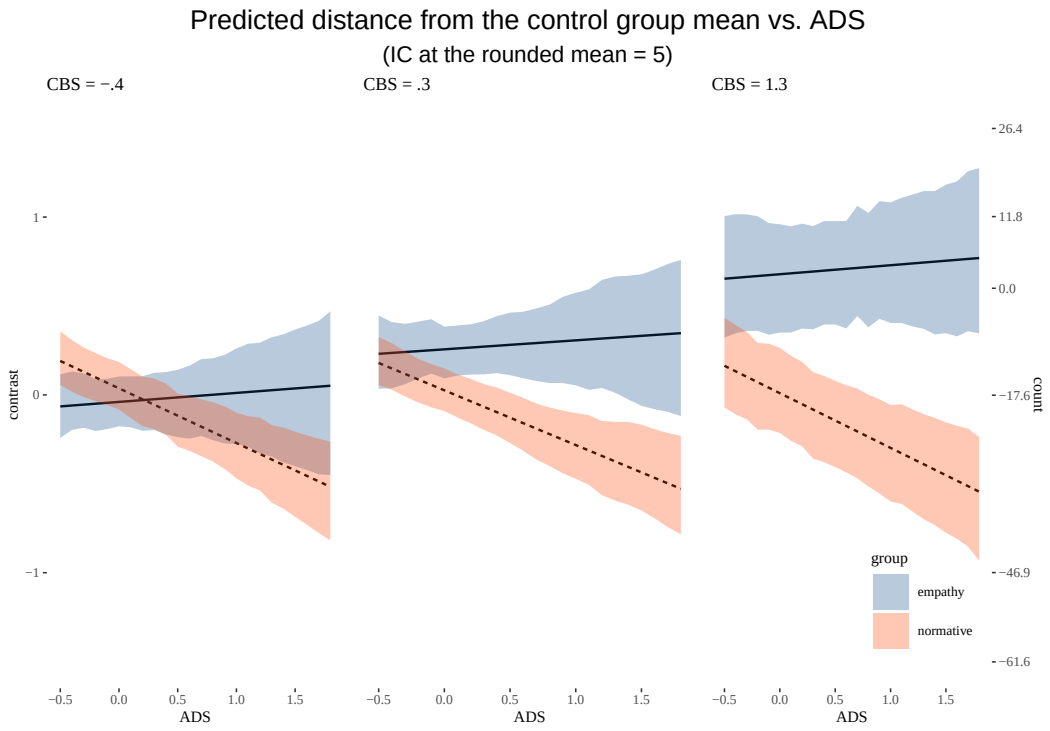


Figure 12: Predicted contrasts (difference in attacks as compared to the control group) for the two treatment groups vs aggression during the treatment period. Notice that empathetic interventions correlated with decreased attacks for less aggressive users, but performed worse than normative interventions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

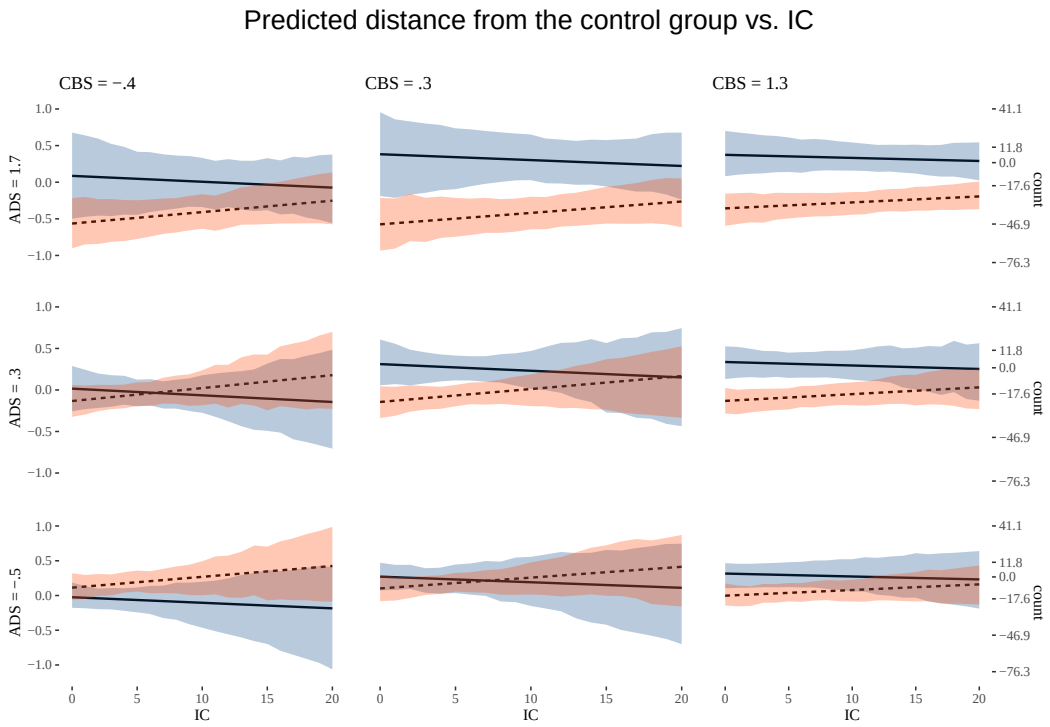


Figure 13: Contrasts (change in attacks as compared to the control group) vs the number of interventions received. Note that repeating empathetic interventions correlates with decreased attacks, while repeating normative interventions is counterproductive.

## 2. Discussion

## 3. Volunteer engagement and impact of competitions

### *3.1. The challenge of keeping volunteers engaged*

### *3.2. Volunteer activity data analysis*

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

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

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

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

Moreover, the model is hierarchical: the individual level parameters are drawn from distributions whose parameters are in turn to be estimated as well. Thus, *lbar* is the overall baseline for the whole group, *enthbar* is the overall estimated group enthusiasm coefficient, and *compbar* is the overall

estimated competition impact coefficient (all of them come with their own nuisance sigma parameters).
All of these parameters are given priors in a manner analogous to the introduction of priors for the other time series models, as explained in the appendix.[3]

Raw data and daily means are illustrated in Figure 14, and the individualized totals with the key coefficients based on the trained model are illustrated in Figure 15.

## Appendix A. Explanation of WAIC

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

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

where $S$ is the number of samples in the posterior, and $\Theta_s$ 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

[3]Interestingly, 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.

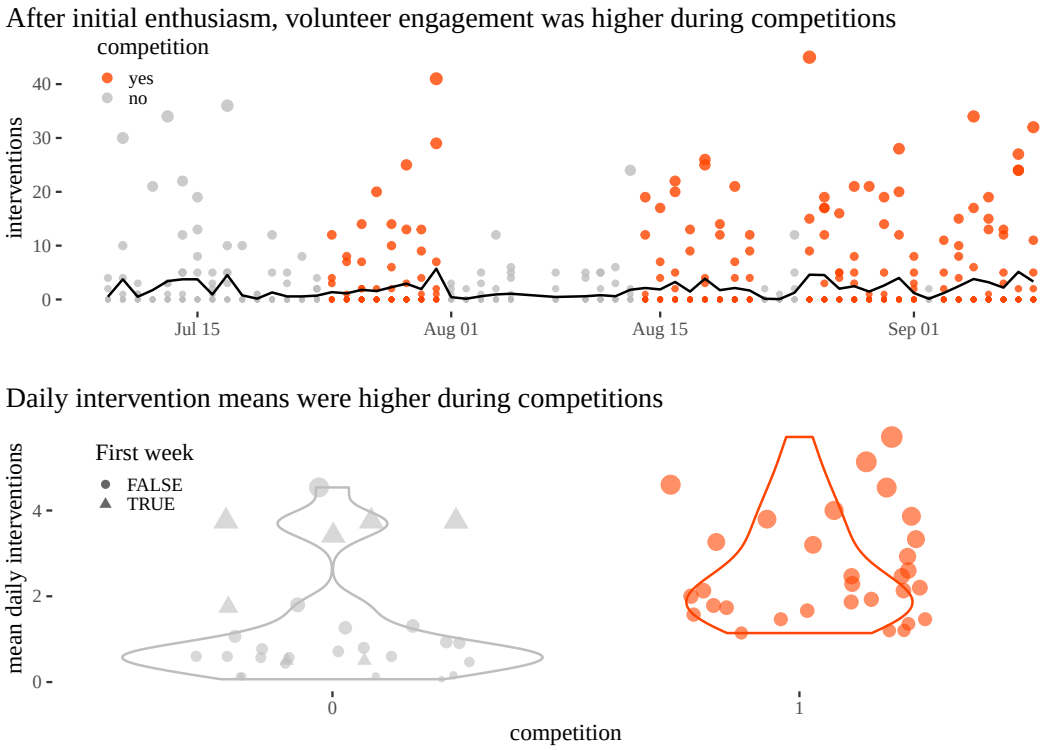


Figure 14: Daily individual voilunteer intervention counts accross time with competition periods marked (top) and daily group intervention means grouped by whether a competition was ongoing (bottom). Note most of high means in the non-competition period are in the first week.

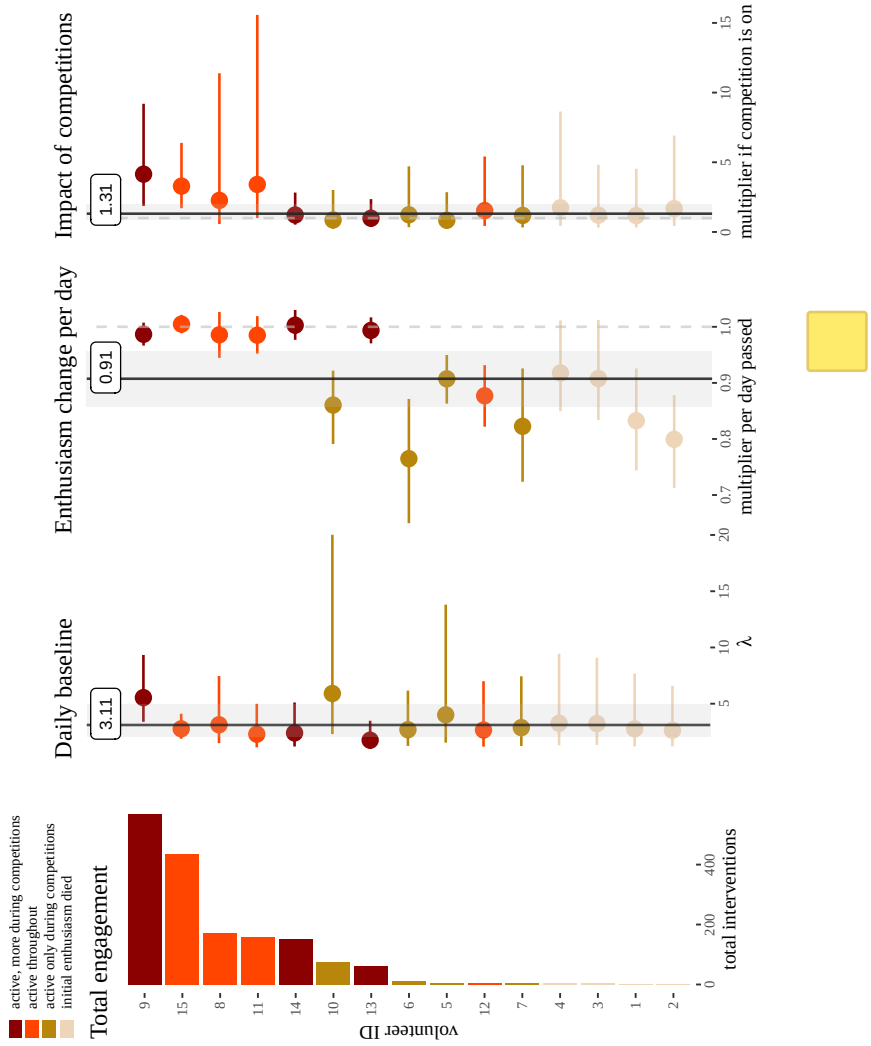


Figure 15: Volunteer total engagement with their daily baseline and multipliers for enthusiasm and impact of competition. Pointranges represent individual level coefficients, group coefficients are represented by black lines with shaded 89% HPDI areas.

Criterion (WAIC), which is an approximation of the out-of-sample deviance that converges to the cross-validation approximation in a large sample. It is defined as the log-posterior-predictive-density with an additional penalty proportional to the variance in the posterior predictions:

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

Thus to construct the penalty, we calculate the variance in log-probabilities for each observation and sum them up. Because of the analogy to Akaike's criterion, the penalty is sometimes called the effective number of parameters, $p_{\mathsf{WAIC}}$. 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 $\lambda$ for all the users, or user-specific $\lambda$s. The first Bayesian model has the following structure:

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

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

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

The priors were chosen using prior predictive check, so that the 89% density intervals reached between 0 and 34, with median around 1. Given

our prior experience with similar user datasets this is a fairly wide informative prior. The comparison, unsurprisingly, preferred the user-specific $\lambda$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.

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

Taking $\phi$ parameters to be user-relative also resulted in improvement:

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

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

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

Now, let us rethink the priors. The coefficients need to be exponentiated to be understood multiplicatively. For instance, the prior for $i1empOverall$ is $\text{Norm}(0, .2)$. To understand what priors for the exponentiated individual coefficients this entails, we can simulate: (1) draw 1e4 values $i1bar$ of the mean from $\text{Norm}(0, .2)$, (2) draw 1e4 values $i1sigmabar$ of the standard deviation parameter from $\text{Exp}(5)$, and each time (3) draw 1e4 parameters from $\text{Norm}(i1bar, i1sigmabar)$. The resulting distribution looks as in Figure B.16. This is still a very wide prior for the multiplicative impact of emphatetic interventions, centered around 1, allowing even extremely unlikely values close to 0 or 2 (upon reflection: you really should not expect a single intervention to reduce aggression to zero or to double it in everyone). In the cumulative model for computation reasons we will need to narrow down the distributions, but the general point hold: prior predictive check still ensures that they are centered around neutral values and that they allow for a very reasonable range of values.

## Appendix C. Model choice for the lo[redacted]erm analysis

let us elaborate on how we decided to use to the seemingly fairly complicated model we already described in the body of the paper. Once preliminary

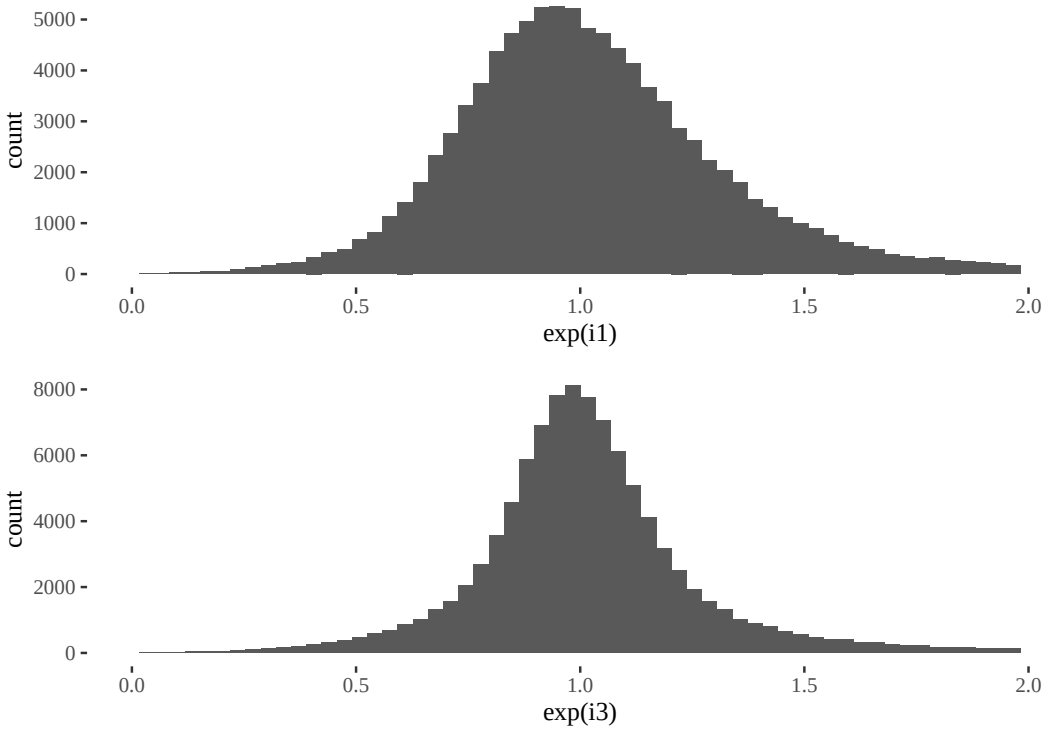


Figure B.16: Simulated priors for the individual i1 coefficients, and prior for the cumulative impact model with larger input variability (hence, the prior is more narrow to eliminate unrealistically huge impact).

causal considerations guided our restrictions on variable selection, we proceed by building models of increasing complexity, and comparing them in terms of Widely Acceptable Information Criterion (which we have already discussed). The models differ mostly in the underlying linear formulae. For computational ease we will here use quadratic approximations, while in the final analysis we will deploy Hamiltionian Monte Carlo. The names are meant to decode the model structure: the predictors are listed before dashes, whereas interactions are listed after dashes. The comparison results are in Table C.2 and plotted in Figure C.17. 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.

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

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

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

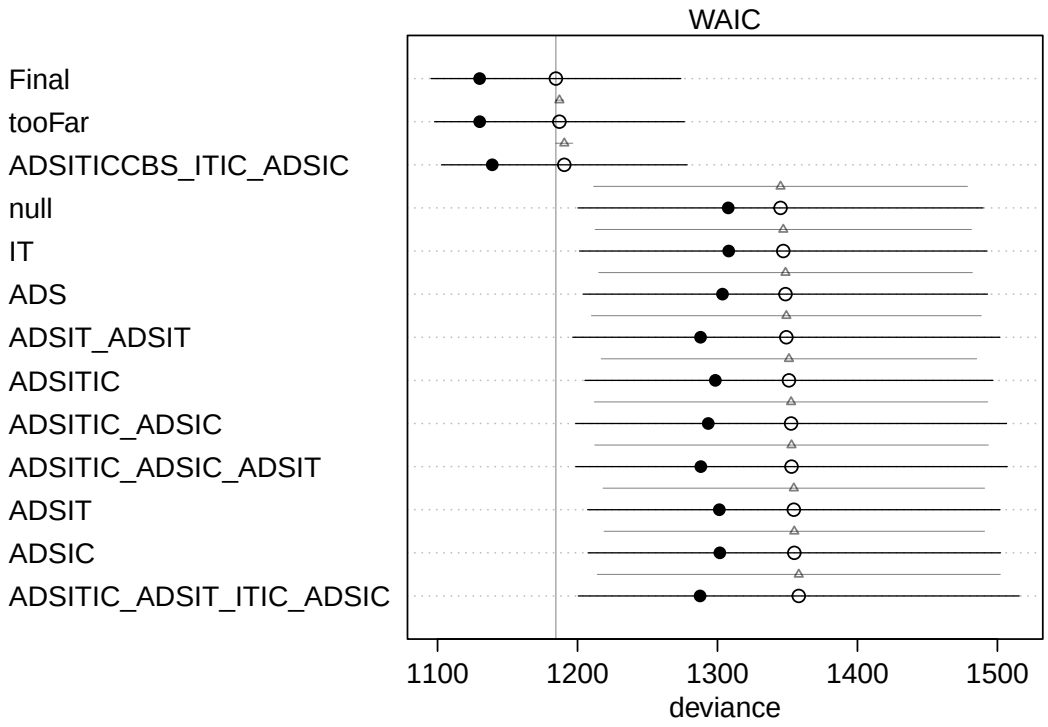


Figure C.17: Model comparison, WAIC scores. The filled points are the in-sample deviance values. The open points are the WAIC values. The line segments represent standard errors of the WAIC scores. really want however is the standard error of the difference in WAIC between the two models. The triangle is the difference to the top rated model, and the line segment going through it is the standard error of this difference.

| | 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.

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.

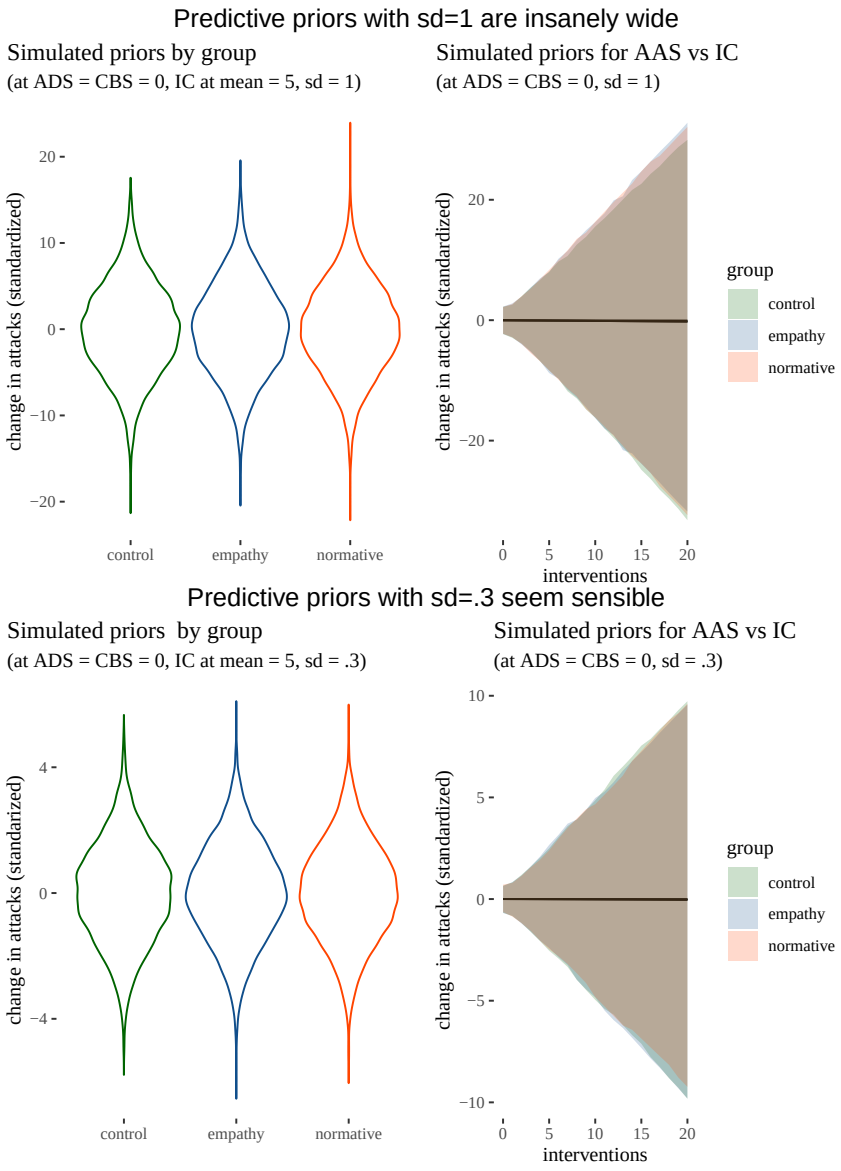


Figure C.18: Prior predictive check for two different sets of priors.