UNIVERSITY of CALIFORNIA PRESS | Collabra: Psychology

Young Reusser, A. I., Veit, K. M., Gassin, E. A., & Case, J. P. (2024). Responding to Online Toxicity: Which Strategies Make Others Feel Freer to Contribute, Believe That Toxicity Will Decrease, and Believe That Justice Has Been Restored? *Collabra: Psychology*, *10*(1). https://doi.org/10.1525/collabra.92328

Social Psychology

Responding to Online Toxicity: Which Strategies Make Others Feel Freer to Contribute, Believe That Toxicity Will Decrease, and Believe That Justice Has Been Restored?

Alison I. Young Reusser[1][a], Kristian M. Veit[2], Elizabeth A. Gassin[2], Jonathan P. Case[3]

[1] Psychology and Criminal Justice, Houghton University, Houghton, NY , US, [2] Behavioral Sciences, Olivet Nazarene University, Bourbonnais, IL, US, [3] Theology, Houghton University, Houghton, NY, US

Keywords: online discourse, benevolence, empathy, forgiveness, toxicity

https://doi.org/10.1525/collabra.92328

Collabra: Psychology

Vol. 10, Issue 1, 2024

When we encounter toxic comments online, how might individual efforts to reply to those comments improve others' experiences conversing in that forum? Is it more helpful for others to publicly, but benevolently (with a polite tone, demonstrated understanding of the original comment, and empathy for the commenter; Young Reusser et al., 2021), correct the post? Is going along with or joking along with the commenter in a benevolent way helpful? Or is retaliating – returning toxicity for toxicity – the best strategy? Using real Reddit conversation pairs – a toxic comment followed by a reply – as stimuli, we conducted a pilot study (n = 126 participants) and pre-registered experiment (n = 1357 participants) investigating the impact of three kinds of replies to online toxicity (benevolent correction, benevolent going-along, or retaliation) on observers' self-reported freedom to contribute to the conversation, their belief that the toxicity will be reduced, and their overall impression that justice has been restored. We found evidence that benevolently correcting the toxicity helped participants feel freer to contribute than retaliating against it. Benevolently correcting was also seen as the best option for dissuading the toxicity and restoring justice. These findings suggest that treating toxic commenters with empathy, understanding, and politeness while correcting their toxicity can be a useful strategy for online bystanders who want to intervene to improve the health of online discourse. Preregistered Stage 1 protocol: https://osf.io/hfjnb (date of in-principle acceptance: 01/23/2023).

You may have had the experience of reading through an engaging, lively discussion online and suddenly coming across a post that is toxic; in other words, hateful, aggressive, or disrespectful, potentially making you, and others, want to leave that discussion (Perspective, 2021). According to Pew Research, many U.S. adults have had similar experiences; 66% of U.S. respondents reported witnessing harassment online, from less severe cases (e.g., offensive name-calling) to more severe (e.g., physical threats; Pew Research Center, 2017). Forty-one percent of respondents reported personally experiencing such harassment (Pew Research Center, 2021). Pew also found that many U.S. adults prefer strict consequences for such users, with 51% of respondents saying permanently banning those who bully or harass others would be very effective in reducing the behavior. Many respondents (60%), though, say that those who witness harassing behavior should take on a "major role" in fixing the issue (Pew Research Center, 2017).

Toxic behavior online, then, is common and widely regarded as a problem that needs to be addressed, not only by platforms but by users themselves. Many researchers have studied online toxicity, identifying it (e.g., Wulczyn et al., 2017; Xia et al., 2020) and highlighting factors that can lead to it (e.g., Almerekhi et al., 2020). Another question, though, and one that has received less attention, is whether prosocial, positive responses to toxic behavior can help reduce toxicity.

Bao et al. (2021) asked participants to view hundreds of pairs of conversations from the discussion platform Reddit and select which of the two was more prosocial. They found that prosocial conversations tended to share information more, were rated higher by the community, received more engagement (e.g., total replies, sustained conversation from the same users), and included more polite content (e.g., compliments, laughter). Prosocial conversations were not simply, according to Bao et al. (2021), "the ab-

a Correspondence concerning this article should be addressed to Alison Young Reusser, Houghton University, One Willard Avenue, Houghton, NY 14744. Email: alison.youngreusser@houghton.edu

sence of antisocial behavior" (p. 1138). They found a small-to-moderate negative correlation between the likelihood that a conversation was prosocial and the number of toxic replies in that conversation, suggesting that kindness in response to toxicity might help conversations develop with a modestly kinder trajectory.

In previous work by the current research team (Young Reusser et al., 2021), we sought to understand prosocial behavior in the face of toxicity online. We developed a three-item scale to measure a construct we called reply benevolence. A benevolent post demonstrates understanding of the content of the original comment, empathy (care for, interest in, respect for, and concern for the well-being of the original commenter), and is politely, thoughtfully, and/or helpfully worded (Young Reusser et al., 2021). We asked 792 online volunteers to rate the benevolence of a single reply to the roughly 8,600 most-toxic comments in a dataset of about 11 million Reddit posts from the month of January, 2016 using this scale. We found that 37.83% of the replies were rated above the scale midpoint in benevolence. Benevolence in response to highly-toxic comments was a common strategy.

Since publishing that report, we explored that same Reddit dataset further to see what sorts of replies counted as benevolent. Our reasoning was that someone could appear empathic, understanding and polite because they are simply joking along with, rather than trying to counteract, the toxic commenter. In brief, we identified two distinct strategies in the 669 most-benevolent replies to toxic posts in our data: 50.24% of these replies corrected the toxic comment in some way (Benevolent Correction) and 37.10% of the replies went along with (that is, agreed with or joked along with) the toxic commenter (Benevolent Going-Along). These strategies were strongly negatively correlated (r = -.74, p < .001; see Supplementary Materials for further details).

Building on this descriptive work, we were interested in the current research in understanding any potential differences in effectiveness among three types of replies to toxic posts online: Benevolent Corrections, Benevolently Going-Along, and another understandable response to toxicity – retaliation (Retaliatory). We assessed three separate dependent measures we thought might be impacted by these replies: 1) how free participants feel to engage in the conversation, 2) their sense that toxicity will decrease, and 3) their sense that justice has been restored.

Dependent Measure One. How Free Do People Feel to Engage in the Conversation?

Our first research question is as follows: To what extent do Benevolently Correcting, Benevolently Going-Along, or Retaliatory responses to toxic comments online make observers feel freer to contribute to the conversation?

The Spiral of Silence Theory (Noelle-Neumann, 1977) holds that in situations where there is disagreement about public opinion, individual speakers will tend to first identify what position they are able to express while avoiding social isolation. If they realize that an expressed opinion doesn't have much social support, they will tend to express that opinion less and less. Perceiving that others are "with them," though, can make speakers more likely to contribute in a public forum. For example, Zerback and Fawzi (2017) found that participants who supported evicting immigrants were substantially more likely to post a comment to a fictitious Facebook group if previous commenters agreed with their position compared to if previous commenters disagreed with them. Those who opposed evicting immigrants, though, did not change their commenting behavior depending on prior comments.

According to Spiral of Silence Theory, a toxic comment may serve to indicate to a forum user that at least some other users are actively aggressive. This may signal that posting one's own opinion will result in further social isolation (here, possibly, in the form of an attack from the toxic commenter). If the toxic comment is left uncorrected, participants could feel less free to engage in the forum. This is consistent with some research on online toxicity; for instance, Mohan et al. (2017) found that the higher the percentage of toxic posts in a Reddit forum, the fewer posts and fewer unique users there were (see Salehabadi (2019) for a similar finding based on Twitter conversations). In addition, a 2015 Harassment Survey by the Wikimedia Support & Safety Team suggested that while half (51%) of respondents who witnessed personal attacks or harassment said it did not change their involvement in the community, 42% at least considered not contributing to the site, with only 4% saying their contributions increased (Wikimedia Support & Safety Team, 2015).

Other work suggests, though, that counter to Spiral of Silence Theory, toxic comments may increase forum engagement. Xia et al. (2020) found a small positive correlation between the toxicity of a Reddit comment and the number of direct replies to that comment across several topics. In their analyses of BBC news-related message boards, Chmiel et al. (2011) found that boards with more negative emotional content had more user engagement. They speculated, though, that angry exchanges "may encourage other users to adopt a similar tone" (p. 14). In other words, more engagement does not necessarily mean healthier conversation. Consistent with this, Xia et al. (2020) found that Reddit replies to toxic comments were slightly more likely to be toxic themselves. While Kolhatkar and Taboada (2017) found evidence that the toxicity of a comment was unrelated to that comment's ability to promote civil dialogue, they based this on direct replies to news articles. Their findings may not be applicable to interpersonal online conversation.

How might direct replies to toxic comments impact how free users feel to contribute, according to Spiral of Silence Theory? A reply that corrects the toxicity in a benevolent way might serve as a signal that the comment was an unpopular one, increasing the likelihood that users will speak up. A retaliatory reply might serve a similar function, setting a norm for others of speaking out against the toxicity, albeit in a more negative way. A reply that is benevolent but that simply goes along with the toxicity might actually serve to reinforce a sense that others *agree* with the toxic

commenter, potentially *decreasing* the likelihood that the participant will speak up.

Research from the forgiveness literature supports the idea that publicly correcting toxicity could free others up to contribute. Hershcovis et al. (2018) found that participants who reported confronting incivility in the workplace experienced more psychological forgiveness (forgiving, wanting good things to happen to them and for others to treat them fairly) for the uncivil coworker. Research by Gromet and Okimoto (2014) suggests that when a victim of workplace incivility forgave, observers not only believed their relationship was repaired, but felt more comfortable interacting with both the offender and the victim than if the victim did not forgive.

> Hypothesis 1. Participants will feel freer to contribute to a conversation initiated by a specific toxic comment after a Benevolent Correction or Retaliatory reply compared to a reply that Benevolently Goes Along with the toxic comment.

Dependent Measure 2. Will Toxicity Decrease?

Cialdini, Kallgren and Reno (1991) distinguished in their Focus Theory of Normative Conduct between descriptive norms, social norms which highlight what people typically do in a situation, and injunctive norms, those which highlight what people believe is appropriate. In a classic example of this (Cialdini et al., 1990), the likelihood that participants would litter was highest when a descriptive norm was highlighted; a confederate dropping litter in a littered environment suggests that littering is what people typically do. The likelihood of littering was lowest when an injunctive norm was highlighted; a confederate dropping litter in a clean environment suggests that littering is socially unacceptable.

What sorts of replies to online toxicity will make it appear socially unacceptable and thus less likely to occur? A person correcting a toxic post in a benevolent way might highlight two injunctive norms: that saying toxic things is wrong and treating others kindly is right. The injunctive norm to be kind to others might be especially salient to observers given how unexpected it might seem in that situation. If observers intuit this, they could believe the toxic commenter will be less likely to reoffend after a benevolent correction. A retaliatory reply could highlight the injunctive norm that saying toxic things is wrong, but might not highlight the importance of treating others kindly. Observers in this case might believe the toxic commenter will be less likely to reoffend, but perhaps not to the same extent. A reply that is benevolent but goes along with the toxic comment might not highlight injunctive norms at all, leading observers to feel that the toxic commenter's behavior will not change (or possibly will increase due to an injunctive norm that toxicity is socially acceptable here). Theoretically, then, benevolent corrections should be most effective for dissuading toxicity, and benevolently going along with the toxicity least effective, with retaliation somewhere in the middle.

Some evidence that corrective posts might dissuade toxicity exists. Hangartner et al. (2021) found that politely-worded corrective messages intended to elicit empathy from Twitter users who posted hate speech resulted in a small reduction in the amount of hate speech posted and slightly increased the likelihood that that user would delete their negative tweet. In an elementary school context, Saarento et al. (2013) found that when teachers were perceived to disapprove of bullying, students self-reported less victimization by other students. On the other hand, Wright et al. (2017) found that when Twitter users publicly replied to hate speech with corrective posts, responses to these corrective posts were at times positive (apologies), but sometimes negative (angry argument).

What about retaliation? Work by Molnar, Chaudhry and Loewenstein (2020) is consistent with the idea that retaliation could be viewed as a deterrent for future toxicity. When their participants learned that a target individual was unfair to them, the majority of those who retaliated monetarily preferred to include a message explaining the reason (e.g., "because you were unfair to your partner" (p. 7)). Further, in another experiment, participants who retaliated and explained their reasoning expected the target individual to treat others better in the future – in other words, some participants thought retaliation (with explanation) would dissuade the negative behavior.

Retaliation is an option some online users consider; in guided interviews, Ziegele et al. (2014) found that while most news website users said they would not respond to aggressive comments, some said they "felt challenged to rebuke the authors of these postings" (p. 1118). In qualitative interviews with 19 women from online video game forums, Cote (2017) found that one of several common responses to online harassment was being aggressive or sarcastic in return. Further, one interviewee argued that this stopped the harassment or changed it into more friendly banter.

Participants could believe that retaliation might not be an effective deterrent, though. While Benesh et al. (2016) argue that continued, civil conversation might slowly reduce the extremity of hate speech, they specifically discourage hostility and aggression, which they argue can entrench the original commenter and escalate the situation. Indeed, Herschovis et al. (2018) found that when participants confronted incivility in the workplace, incivility was more likely to reoccur. In other words, the injunctive norm being highlighted by a retaliatory response might be more that aggression is acceptable than that toxicity is not, potentially encouraging toxicity as a result.

> Hypothesis 2. Participants will believe more that toxicity has been dissuaded...
> a. when the replier Benevolently Corrects or Retaliates compared to Benevolently Going Along.
> b. when the replier Benevolently Corrects compared to either alternative (Benevolently Going Along or Retaliating).

Dependent Measure 3. Has Justice Been Restored?

Wenzel, Okimoto, Feather and Platow (2008) distinguish between two psychological motivations for bringing about

justice after a transgression: retributive and restorative. In retributive justice, moral order which has been disrupted by a perpetrator is reestablished by punishing them and giving them what they deserve. The punishment fixes the problem, regardless of the transgressor's remorse or lack thereof. In restorative justice, the transgressor is viewed as in conflict with the victim (or victims) and the community as a result of their actions. The goal of restorative justice is for all parties to gain "a shared understanding of the harm the offense has done and the values it violated" (p. 378). Importantly, while transgressors should accept accountability for their actions in this model, victims are also urged to see transgressors with benevolence, forgiving them and viewing them as "a morally worthy person capable of more than wrongdoing" (Govier, 1999, p. 60).

Accordingly, which sorts of replies to an online transgression – a toxic comment – will give participants the sense that justice has been restored? Perhaps a restorative-justice approach, correcting the comment in a benevolent way, will do so. When participants were asked to imagine that another student had lied to them to get out of a group project, their sense that the situation was fair and just was moderately higher if they imagined forgiving that student than if they did not, regardless of whether the offender apologized (Wenzel & Okimoto, 2010). Forgiving also reduced hostile emotions compared to not forgiving. If forgiveness is analogous to kindly, empathically correcting toxicity, perhaps observers might see benevolent correction as increasing fairness and justice.

Alternatively, a retributive-justice approach, a retaliatory response, could be seen as more restorative of justice. Liang et al. (2018) found that when participants were asked to stab an online voodoo doll representing a rude or otherwise hostile supervisor, their implicit sense of injustice was lower than participants who did not retaliate in this way. Wang and Todd (2021) found that when participants read a brief conversation between a negative target individual (a member of a White supremacist group) and another person who either condemned the target or empathized with them, participants had more respect for this second person if they condemned the target.

> Hypothesis 3. After reading a set of conversations, participants will feel that justice has been restored...
> a. after a Benevolent Correction compared to either a Benevolently Going-Along or Retaliatory reply.
> b. after a Retaliatory reply compared to a Benevolent reply of either kind (Correction or Going Along).

We specify which condition differences would support each hypothesis as well as what would disconfirm each hypothesis in the PCI RR study design below (Table 1).

Pilot Study

Method

We conducted a pilot study to estimate the smallest effect size of interest for use in a power analysis to determine an appropriate sample size for our main experiment. We also hoped to refine our materials and methods based on this study. The methods and materials used in the pilot are almost identical to those used in the main experiment, so we devote more time to them here. The pilot tested the same hypotheses as the main experiment. A pdf of our Qualtrics survey (pilotQualtrics.pdf) and deidentified pilot data can be found on the Open Science Framework (https://osf.io/6dwjx/?view_only=2b45b35cf37e46e5818a40bf79fc981d).

Participants

We collected pilot data from 126 participants recruited from psychology courses across two faith-based liberal arts colleges (college A, n = 66; college B, n = 60) in Fall of 2021 and Spring of 2022 to test our materials and design. Nine were dropped for failing an attention check, resulting in a final sample size of 117. The average social media usage per day in our sample was 2.87 hours (SD = 1.90; min = 0; max = 10) and the average comfort with offensive language on a scale from 0 (it makes me very uncomfortable) to 4 (it doesn't bother me) was 2.89 (SD = 1.07; min = 0; max = 4; see Figure 1). Participants were randomly assigned to either the Benevolent Correction condition (n = 40 after two were dropped), the Benevolent Going Along condition (n = 35 after six were dropped), or the Retaliatory condition (n = 42 after one was dropped). Using a chi-square goodness-of-fit test, we did not find evidence that significantly more participants were dropped from a given condition, $\chi^2(2) = 4.67$, p = .097.

Materials

Reddit Conversations. Participants in each condition read four separate conversations during the experiment in a randomized order. These were selected from Young Reusser et al.'s (2021) Reddit 2016 dataset. They were taken verbatim from the dataset and unedited, preserving any grammatical errors or typos. Each consisted of a toxic comment followed by a single reply – a comment-reply pair. All benevolent replies were rated as a 5 or above on the 1-6 benevolence scale described in Young Reusser et al. (2021). Three of the four benevolently correcting replies were rated above the scale midpoint in Correcting by undergraduate research assistants; one was researcher-selected. Three of the four benevolently going-along replies were rated above the scale midpoint in Going Along; one was researcher-selected. All four retaliatory replies were researcher-selected from a list of the least-benevolent replies and further, demonstrated a negative, aggressive, dismissive and/or rude tone. Comments which revealed a controversial opinion (e.g., a political argument) were excluded from consideration. We also tried to keep the lengths of the comments and replies manageable and similar across conditions, although since these were real conversations, we could not exactly match comment length. All conversations, their ratings and word counts can be found in Appendix A.

Per-Pair Ratings. After each comment-reply pair, participants rated their first impression of the toxic commenter using seven options from -3 (Very negative) to +3 (Very positive). They were then asked to consider both individuals in the conversation and indicate how free they felt

Table 1. PCI RR Study Design: Sampling and Analysis Plan, Findings That Would Support or Disconfirm Each Hypothesis

Question	Hypothesis	Sampling plan	Analysis Plan	Sensitivity rationale	Interpretation given different outcomes	Theory that could be shown wrong by the outcomes
Q1: To what extent do Benevolently Correcting, Benevolently Going Along, or Retaliatory responses to toxic comments online make observers feel freer to contribute to the conversation?	1. Participants will feel freer to contribute to a conversation initiated by a specific toxic comment after a Benevolent Correction or Retaliatory reply compared to a reply that Benevolently Goes Along with the toxic comment.	According to GPower, the sample size required to detect the smallest effect size of interest (*f* = .11) 90% of the time using an ANCOVA with a three-level factor (condition) and three covariates is 1049. We estimate based on the pilot that roughly 7% might be dropped for failing an attention check, so we added an additional 7% to our proposed sample, resulting in a target sample of 1122 participants.	Participants who fail both attention check questions will be dropped prior to all analyses. We will maintain the false discovery rate at 5% across all analyses using the Benjamini-Hochberg false discovery procedure (Benjamini & Hochberg, 1995). We will conduct a multilevel regression nesting ratings within comment-reply pair (1-12) and participant, predicting perceived freedom to contribute from condition (fixed factor, between-subjects; Benevolent Correction vs. benevolent Going-Along vs. Retaliatory reply) and three covariates (comfort with offensive language, first impression of toxic commenter, and willingness to self-censor). We will use two planned comparisons to test the differences between the 1) Benevolent Correction and Benevolent Going-Along and the 2) Retaliatory and Benevolent Going-Along condition means. A Bonferroni post hoc comparison will be used to test the difference between 3) the Benevolent Correction and Retaliatory means.	The smallest effect size of interest of *f* = .11 was drawn from research by Zerback and Fawzi (2017) who found that size for the difference in likelihood to post a comment when the majority is on vs. opposing your side.	Support for H1: The Benevolently Going Along condition's mean is lower than the other two at the .05 level. The two other conditions do not differ significantly. If the Benevolently Going Along condition's mean is similar to either other condition, this hypothesis would be disconfirmed. This hypothesis is agnostic as to the difference between the other two conditions. However, if the Benevolent Correction condition's mean is significantly higher than the Retaliatory condition's mean, that might suggest that the polite tone of the correction provides additional incentive to contribute. If instead the Retaliatory condition's mean is higher than the Benevolent Correction condition's mean, that might suggest that a negative tone is more likely to encourage others to respond. Note that we do not have a prediction as to the size of the effects for these hypotheses.	If H1 is supported, it would be consistent with Spiral of Silence Theory's (Noelle-Neumann, 1977) prediction that the apparent majority opinion leads those in the apparent minority to be less and less likely to speak. If H1 is not supported, it would suggest that in some cases, those who think they hold a minority opinion (e.g., not liking toxicity when someone has Benevolently Gone Along with a toxic comment) still want to speak up.
Q2: To what extent do Benevolent Corrections, Benevolent Going Along, or Retaliatory responses to toxic comments online make observers feel that the toxicity has been dissuaded?	Two possibilities: 2. Participants will believe to a greater extent that toxicity has been dissuaded… 1. when the replier Benevolently Corrects or Retaliates compared to	See above	Parallel multilevel regression analysis to Q1, with perceptions that toxicity has been dissuaded as the dependent measure. We will use planned comparisons to test the differences between 1) the Benevolent Correction and Benevolent Going Along conditions and the 2) Retaliatory and Benevolent Going Along conditions. We will use a Bonferroni post hoc test to compare the Benevolent Correction and Retaliatory conditions.		Support for H2a: The Benevolently Going Along condition's mean is lower than the other two at the .05 level. This hypothesis is agnostic as to the difference between the other conditions. If the Benevolently Going Along condition does not differ significantly from one or both of the other two conditions, this hypothesis would be disconfirmed. To be precise, though this hypothesis is agnostic to comparison between the Benevolent Correction and Retaliatory conditions, if the Benevolent Correction condition is higher than the Retaliatory, it might indicate that, similar to H2b, Benevolent corrections highlight the injunctive norms that saying toxic things is	If H2a is supported, it would be consistent with Cialdini et al.'s (1991) Focus Theory of Normative Conduct, potentially indicating that any correction of toxicity highlights an injunctive norm that toxicity is not socially acceptable, dissuading toxic posts.

Question	Hypothesis	Sampling plan	Analysis Plan	Sensitivity rationale	Interpretation given different outcomes	Theory that could be shown wrong by the outcomes
	Benevolently Going Along. 2. when the replier Benevolently Corrects compared to either alternative (Benevolently Going Along or Retaliating).				wrong and treating others kindly is right, whereas Retaliatory replies only highlight the first, serving as a weaker indication that toxicity has been dissuaded. If the Retaliatory condition mean is higher than the Benevolent Correction mean, that would suggest that the injunctive norm that toxicity is not socially acceptable is more salient if worded in a retaliatory way than in a benevolent way. Support for H2b: The Benevolent Correction condition's mean is higher than the other two at the .05 level. This hypothesis is agnostic as to the difference between the other conditions. If the Benevolent Correction condition does not differ significantly from one or both of the other two conditions, this hypothesis would be disconfirmed. To be precise, though this hypothesis is agnostic to comparison between the Benevolent Going-Along and Retaliatory conditions, if the Benevolent Going-Along condition mean is lower than the Retaliatory, that might suggest that Retaliatory replies are at least providing an injunctive norm that toxicity is not acceptable, whereas Benevolently Going Along is not. If the Retaliatory condition mean is lower than the Benevolent Going-Along mean, that might indicate that Retaliation is viewed as using the same tactics as the initial toxic post, reinforcing that toxicity is socially acceptable (and as such, not dissuaded). Note that we do not have a prediction as to the size of the effects for these hypotheses.	If H2b is supported, it would suggest that for participants to think toxicity will be dissuaded, the injunctive norm cannot only be that toxicity is not acceptable, but positively that behaving kindly is socially expected. If neither is supported, it would be inconsistent with the Theory of Normative Conduct.
Q3: After reading a set of conversations, to what extent will observers feel that justice	Two possibilities: 1. Benevolent Corrections (vs. Benevolently Going Along or Retaliating) will	See above	We will conduct an ANCOVA predicting perceived justice restoration from condition (between-subjects; Benevolent Correction vs. Benevolent Going Along vs. Retaliatory reply) and two covariates (comfort with offensive language and willingness to self-censor). We will use planned comparisons to test the differences between 1) Benevolent		Support for H3a: The Benevolent Correction condition's mean is higher than the other two at the .05 level. This hypothesis is agnostic as to the difference between the other conditions. If the Benevolent Correction condition's mean does not differ significantly from either of the other two conditions, this hypothesis would be disconfirmed.	If H3a is supported, it would suggest that participants are taking a restorative-justice approach to the online conflict. If H3b is supported,

Question	Hypothesis	Sampling plan	Analysis Plan	Sensitivity rationale	Interpretation given different outcomes	Theory that could be shown wrong by the outcomes
has been restored by Benevolent Corrections, Benevolent Going Along, or Retaliations?	make participants feel more that justice has been restored 2. Retaliatory responses (vs. Benevolently Correcting or Going Along) will make participants feel more that justice has been restored		Correction and Benevolent Going Along conditions and 2) Benevolent Correction and Retaliatory conditions. A Bonferroni post-hoc test will be used to compare the 3) Benevolent Going Along and Retaliatory conditions.		To be precise, though this hypothesis is agnostic to comparison between the Benevolent Going Along and Retaliatory conditions, if the Retaliatory mean is higher than the Benevolent Going Along mean, that might indicate that any correction, even a negative one, is seen as bringing about justice. If the Benevolent Going Along mean is higher than the Retaliatory mean, that might indicate that a kind tone is seen as restoring justice even without an explicit correction of the toxicity. Support for H3b: The Retaliatory condition's mean is higher than the other two at the .05 level. This hypothesis is agnostic as to the difference between the other conditions. If the Retaliatory condition does not differ significantly from either of the other conditions, this hypothesis would be disconfirmed. To be precise, though this hypothesis is agnostic to comparison between the Benevolent Correction and Benevolent Going Along conditions, if the Benevolent Correction mean is higher than the Benevolent Going Along mean, that might suggest that restoring a sense of justice is not only due to benevolence, but related to a correction of the toxicity. If the Benevolent Going Along mean is higher than the Benevolent Correction mean, that might indicate that correcting in a benevolent way might signal that the correction is somehow compromised, perhaps less authentic, and less restorative of justice. Note that we do not have a prediction as to the size of the effects for these hypotheses.	it would suggest that participants are taking a retributive-justice approach.
Manipulation check - benevolence	Ensure that the benevolent replies are rated as more benevolent than the		ANOVA predicting overall rated benevolence of a set of replies from condition (Benevolent Correction vs. Benevolent Going-Along vs. Retaliatory)		The two Benevolent conditions should be rated as more benevolent than the Retaliatory condition at the .05 level. If not, this would undermine our ability to interpret analyses related to RQ1, RQ2 and RQ3.	

Question	Hypothesis	Sampling plan	Analysis Plan	Sensitivity rationale	Interpretation given different outcomes	Theory that could be shown wrong by the outcomes
	retaliatory replies					
Manipulation check - correcting	Ensure that the Benevolently Correcting replies are rated as more correcting than the benevolently Going Along replies		Independent-samples *t*-test comparing the mean overall rating of a set of replies' attempt to correct the initial toxic comment between the Benevolent Correction and the Benevolent Going-Along conditions.		The Benevolent Correction condition should be rated as more correcting of the initial comment than the Benevolently Going Along condition at the .05 level. If not, this would call into question any analyses which suggest a difference between the benevolent correction and benevolent going-along conditions.	
Manipulation check – retaliatory	Ensure that the Retaliatory replies are rated as more retaliatory than either other condition		ANOVA predicting overall ratings of a set of replies' retaliatory ratings from condition (Benevolent Correction vs. Benevolent Going-Along vs. Retaliatory)		The Retaliatory condition should be rated as more retaliatory than the other two at the .05 level.	
Manipulation check – first impression of toxic commenter (pilot); perceived toxicity of initial comment (Main experiment)	Ensure the participant's first impression toxic commenter (pilot) or the perceived toxicity of the initial comment (Main experiment) is similar across conditions. If this is not the case, first impression will be controlled for.		Multilevel regression nesting ratings within comment-reply pair (1-12) and participant, predicting first impression from condition (fixed factor, between-subjects; Benevolent Correction vs. Benevolent Going Along vs. Retaliatory reply).		If the first impression of each toxic commenter differs by condition at the .05 level, this will be included as a covariate in the main analyses for RQ1, RQ2 and RQ3.	

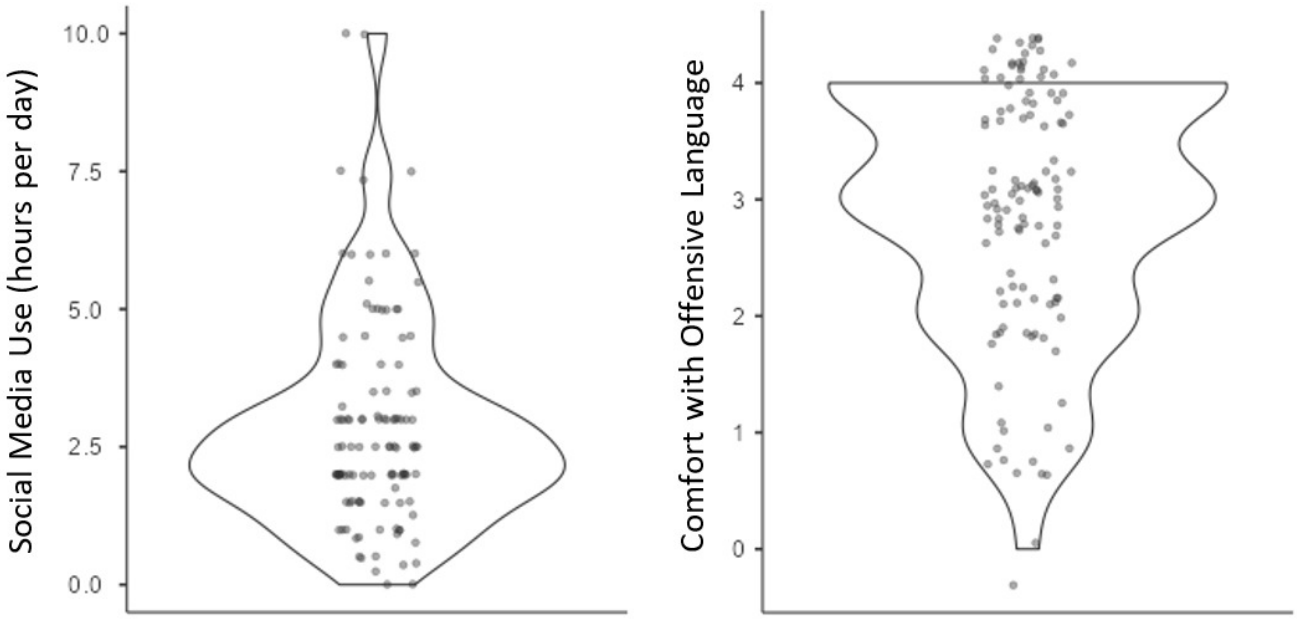

Figure 1. Distributions of Social Media Use and Comfort with Offensive Language in Pilot Study

Note. These violin plots represent the frequency distributions of the 117 pilot participants' social media use in hours per day and comfort with offensive language (0 = it makes me very uncomfortable; 4 = it doesn't bother me).

to contribute using three items (e.g., "How likely would you be to contribute to this conversation?") written by the experimenters. We dropped one reverse-worded item ("If you were to post to this forum, to what extent would you feel the need to hide what you really think from the rest of the group?") to improve the internal consistency of the scale from .65 to .70 and averaged responses to the remaining two items together. Participants responded using seven options from 0 (Not at all) to 6 (Very likely).

Participants then reported the extent to which they believed the reply addressed their concerns and discouraged the toxicity of the initial comment using four items (e.g., "The response is an appropriate way to address the toxicity of the first comment;" "The response will discourage the first commenter from continuing to post in the same negative tone as before", Cronbach's α = .88). They used seven options from -3 (Strongly disagree) to +3 (Strongly agree). These items can be found in Appendix B and means and standard deviations for these measures can be found in Table 2.

Overall Ratings. After reading all four conversations, participants completed an attention check where they were asked to select from four options something they had read in one of the previous conversations. They then rated their overall impression of whether the replies to the toxic comments they had read had restored justice using seven items adapted from Wenzel et al., 2010; e.g., "The resolution to the situation is fair," Cronbach's α = .91). All items except one were identical to the original scale, with one exception: the item "The resolution has restored justice" was modified to "The replies have restored justice" to make it more specific to this study. Participants responded using seven options from -3 (strongly disagree) to +3 (strongly agree).

Individual Differences. Participants then reported their Willingness to Self-Censor (Cronbach's α = .83), an eight-item scale developed by Hayes, Glynn and Shanahan (2005; e.g., "It is difficult for me to express my opinion if I think others won't agree with what I say") using five options from 1 (strongly disagree) to 5 (strongly agree). They rated their level of comfort with reading offensive language from 0 (it makes me very uncomfortable) to 4 (it doesn't bother me at all) and then estimated the total time in hours and minutes in an average day they were on social media for personal use. These were included so they could be controlled for in our analyses; a participant's tendency to avoid stating their true opinion could relate to how free they feel to contribute to any conversation and whether they feel that replying at all is a helpful strategy. Their level of comfort with offensive language could relate to their perception of the need to correct a toxic comment. Social media use was included to help characterize our sample (e.g., are these people who are used to online conversation?).

Procedure

After random assignment to condition, participants read through one toxic comment and one reply from four separate Reddit conversations in a randomized order. They provided ratings of each conversation (first impression of the toxic commenter, free to contribute (3 items, order randomized) and toxicity addressed (4 items, order randomized)), then an attention check, then their overall impression of the fairness/justice of the resolution (7 items, order not randomized). They then completed individual difference measures (Willingness to Self-Censor (Hayes et al. 2005; 8 items, order not randomized), comfort with offensive language and total time per day on social media), and were debriefed. 90% of participants finished in 30 minutes or less. This study had IRB approval from Olivet Nazarene University and Houghton University.

Table 2. Means and Standard Deviations for Key Variables in Pilot and Main Experiment

	Pilot study (n = 117)	Main Experiment (n = 1357)
First impression of toxic commenter (Pilot); Toxicity of toxic comment (Experiment)	-1.61 (1.11)	-0.75 (0.95)
Free to contribute per comment-reply pair	1.92 (1.48)	1.68 (1.77)
Toxicity addressed/dissuaded per comment-reply pair	-1.20 (1.50)	-0.78 (1.65)
Justice restored	-1.19 (1.28)	-0.52 (1.59)
Willingness to self-censor	3.11 (0.78)	3.06 (0.85)
Comfort with offensive language	2.89 (1.07)	3.04 (1.17)
Social media hours per day	2.87 (1.90)	2.56 (2.34)

Note: Standard deviations in parentheses.

Results

Per-Pair Ratings

Our first analyses were on the ratings of each separate conversation. These analyses were all multilevel regression models nesting ratings within pair (1-12; 4 each across three conditions) and participant. Condition (benevolent correction vs. benevolent going-along vs. retaliatory) was a between-subjects fixed factor in each model. Each analysis involved four ratings per 117 people, or 468 observations. We maintained the false discovery rate at 5% across all analyses using the Benjamini-Hochberg false discovery procedure (Benjamini & Hochberg, 1995).

First Impression of Toxic Commenter. To ensure that the toxicity of the initial comments was consistent across all three conditions, we conducted a multilevel regression predicting the first impression of the toxic commenter from condition (Intraclass Correlation Coefficient (ICC) for participant = 0.15, ICC for conversation pair = 0.20. This specifies the proportion of variation in first impressions explained by the cluster variables, conversation pair (1-12) and participant). The effect of condition was not significant, $F(2, 10.3) = 0.28$, $p = .76$; nor was the difference between Benevolent Going Along (M = -1.58, SE = 0.26, 95% CI [-2.16, -1.00]) and Benevolent Correction (M = -1.37, SE = 0.26, 95% CI [-1.95, -0.79]) conditions (planned comparison $t(10.4) = -0.57$, $p = .58$) or the difference between the Retaliatory (M = -1.63, SE = 0.26, 95% CI [-2.21, -1.05]) and Benevolent Correction conditions (planned comparison $t(10.1) = -0.71$, $p = .49$) or between the Benevolent Going Along and Retaliatory conditions ($t(10.2) = 0.14$, $p_{Bonferonni} = 1.00$. Note that the following analyses were conducted both including the covariates (willingness to self-censor and comfort with offensive language) and without, and the effect of condition is reported for both.

Free to Contribute. Using a multilevel model predicting how free participants felt to contribute to the conversation controlling for the first impression of the toxic commenter, willingness to self-censor, and comfort with offensive language ($ICC_{Participant} = 0.43$; $ICC_{pair} = 0.09$), we did not find a significant difference among conditions, $F(2, 22.3) = 1.87$, $p = .18$; without covariates, $p = .13$; see Figure 2. Planned comparisons suggested that the difference between the Benevolent Correction (M = 2.24, SE = 0.23, 95% CI [1.77, 2.71]) and the Retaliatory condition (M = 1.63, SE = 0.22, 95% CI [1.16, 2.10]) was moderate, albeit non-significant, $t(21.5) = -1.90$, $p = .071$, $r = .38$; without covariates, $p = .047$. The comparison between the Benevolent Going Along (M = 2.04, SE = 0.23, 95% CI [1.56, 2.52]) and Benevolent Correction conditions was small but non-significant, $t(23.0) = -0.62$, $p = .54$, $r = .13$; without covariates, $p = .45$. A post hoc comparison between the Benevolent Going Along and Retaliatory conditions was small but non-significant, $t(22.1) = 1.25$, $p_{Bonferroni} = .67$, $r = .26$; without covariates, $p_{Bonferroni} = .64$. Willingness to self-censor[1] was negatively related to how free participants felt to contribute, $b = -0.62$, $SE = 0.13$, 95% CI [-0.87, -0.37], $t(110.0) = -4.89$, $p < .001$. Comfort with

1 To check whether our individual difference measures, Willingness to Self-Censor and comfort with offensive language, differed by condition, we conducted two one-way ANOVAs predicting the measures from condition. Neither was significant (both p-values > .59), suggesting that our manipulation did not affect these individual differences.

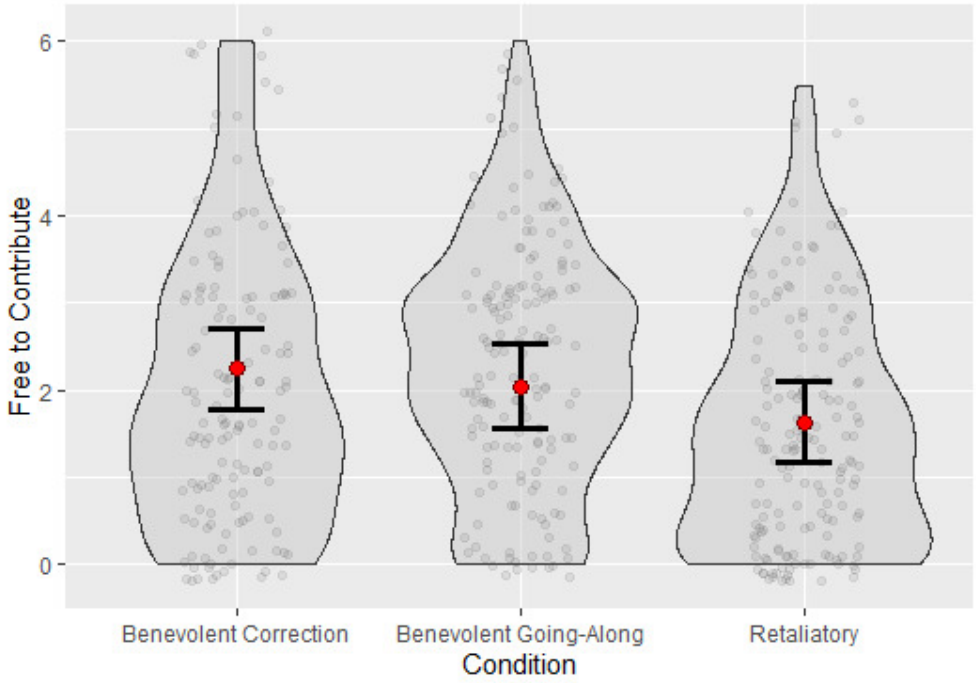

Figure 2. Free to Contribute Across Conditions in Pilot

Note. Error bars represent 95% confidence intervals. Red circles represent condition means. Grey dots indicate participant composite scores calculated by averaging responses to the Free to Contribute items.

offensive language was not a significant predictor (b = 0.08, SE = 0.09, 95% CI [-0.11. 0.27], t(110.0) = 0.86, p = .39).[2]

Toxicity Addressed/Dissuaded. Using a multilevel model predicting the extent to which participants felt the reply addressed and discouraged the initial toxicity controlling for willingness to self-censor and comfort with offensive language ($ICC_{Participant}$ = 0.41; ICC_{Pair} = 0.24), we again failed to find a significant difference among conditions, F(2, 13.1) = 5.10, p = .023 (not significant after the Benjamini & Hochberg (1995) false discovery rate procedure); without covariates, p = .016; see Figure 3.[3] Planned comparisons suggested that the difference between the benevolent correction (M = -0.46, SE = 0.31, 95% CI [-1.12, 0.21]) and benevolent going along (M = -1.76, SE = 0.31, 95% CI [-2.43, -1.08]) conditions was large and significant, t(13.3) = -2.96, p = .011, r = .63; without covariates, p = .007. The difference between the benevolent correction and retaliatory (M = -1.55, SE = 0.31, 95% CI [-2.21, -0.88]) conditions was also large, t(12.8) = -2.51, p = .026, r = .57, though non-significant after the false discovery rate procedure (without covariates, p = .020). A post hoc comparison suggested that the difference between the benevolent going-along and retaliatory conditions was small and not significant, t(13.1) = -0.48, $p_{Bonferonni}$ = 1.00, r = .13. Willingness to self-censor was not related to toxicity addressed, b = -0.19, SE = 0.12, 95% CI [-0.42, 0.04], t(110.1) = -1.61, p = .11. Comfort with offensive language was not related to toxicity addressed, b = 0.08, SE = 0.09, 95% CI [-0.09, 0.25], t(110.0) = 0.90, p = .37.

Overall Ratings: Justice Restored

Recall that after reading all four conversations, each participant provided their overall impression of whether the replies resolved the situation fairly. We conducted a between-subjects ANCOVA predicting this fair/just resolution rating from condition controlling for willingness to self-

2 Inspecting the confidence intervals around the condition means, we noticed that the interval appeared wider in the retaliatory condition (95% CI [0.84, 2.41]) than in either the benevolent correction (95% CI [1.77, 2.69]) or benevolent going along (95% CI [1.82, 2.26]) conditions. These confidence intervals were based on an analysis where we forgot to include pair as a nesting variable. However, based on that analysis, we conducted an exploratory multilevel model predicting first impression of the toxic commenter from condition, pair (1-12) and the interaction between condition and pair, nested within participant (ICC = 0.15). The interaction was significant, F(6, 342) = 16.03, p < .001, suggesting that conversation pairs differed in first impression across condition. Inspecting a plot of the first impressions across condition, one specific pair appeared more positive than the others: pair 9 (pair 1 of the retaliatory condition). The specific comment ("Do you realise how silly you sound?"), on inspection, did not appear to use to be especially negative, either. Further, excluding ratings of this particular pair from the analysis predicting free to contribute resulted in a significant condition effect (F(2, 31.4) = 5.17, p = .011). Based on this, we decided to replace pair 9 in the proposed experiment.

3 Similar to the free to contribute analysis, we also ran the multilevel model eliminating pair 9; this also resulted in a significant condition effect on toxicity addressed (p = .011).

Figure 3. Toxicity Addressed/Dissuaded Across Conditions in Pilot

Note. Error bars represent 95% confidence intervals. Red circles represent condition means. Scale reflects opinion that toxicity has been addressed/dissuaded from -3 (Strongly disagree) to +3 (Strongly agree). Grey dots indicate each participant's composite score calculated by averaging responses to the Toxicity Addressed/Dissuaded items.

censor and comfort with offensive language. We found evidence of at least one difference among the condition means, $F(2, 107) = 17.09$, $p < .001$, $\eta^2 = 0.23$, without covariates, $p < .001$; see Figure 4. Planned comparisons suggested, consistent with Hypothesis 3a, that benevolent corrections were rated as providing a more just/fair resolution ($M = -0.37$, $SE = 0.18$, 95% CI [-0.72, -0.02]) than benevolent going-along ($M = -1.51$, $SE = 0.19$, 95% CI [-1.88, -1.13]), $t(107) = -4.37$, $p < .001$, $d = 1.04$ (without covariates, $p < .001$); and retaliatory ($M = -1.74$, $SE = 0.17$, 95% CI [-2.07, -1.40]) replies, $t(107) = -5.53$, $p < .001$, $d = 1.25$ (without covariates, $p < .001$). We did not find evidence that the benevolent going-along and retaliatory replies differed ($p_{Tukey} = .64$; without covariates, $p_{Tukey} = .61$). Comfort with offensive language was not related to fair/just resolution ratings, $F(1, 107) = 3.72$, $p = .057$, $\eta^2 = 0.03$ and willingness to self-censor was not a significant predictor, $F(1, 107) = 1.47$, $p = .23$, $\eta^2 = 0.03$.

Discussion

Of our three primary dependent measures, we only found evidence of a condition effect for participants' overall ratings of the extent to which the replies to toxic comments restored justice. Consistent with Hypothesis 3a, when a Redditor replied in a benevolent way and corrected the initial toxic remark, participants felt justice had been restored more than either of the other two kinds of replies, which did not differ. The size of this effect was large. We did not find support in this pilot for either Hypothesis 1, regarding how free participants felt to contribute to the conversation,

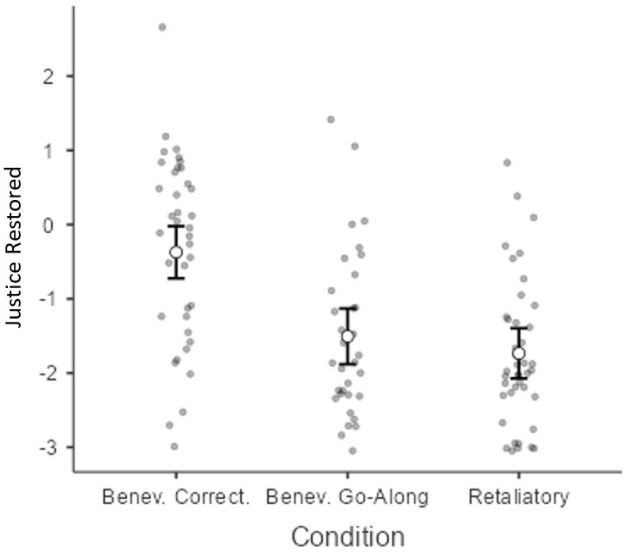

Figure 4. Overall Justice Restored Across Conditions in Pilot

Note. Error bars represent 95% confidence intervals. Scale reflects opinion that justice has been restored from -3 (Strongly disagree) to +3 (Strongly agree). Dots indicate each participant's composite score calculated by averaging their responses to the Justice Restored items.

or Hypotheses 2a and 2b, related to perceptions that toxicity had been dissuaded.

Main Experiment

Smallest Effect Size of Interest

To estimate the smallest effect size of interest for our main experiment, we used effect sizes from our pilot study as well as some effect sizes from the literature. The pilot effect sizes of interest related to condition differences for the three key dependent variables: how free participants felt to contribute (we mistakenly believed these ranged from r = .35 to r = .69),[4] the extent to which they thought toxicity would be dissuaded (we mistakenly believed these ranged from r = .33 to r = .79),[5] and their overall sense that justice had been restored (d = 1.04 to d = 1.25). Willingness to speak out in online settings, in Porten-Chee and Eilders (2015), was higher when personal opinion and the opinion climate clashed (β = .14 and β = .23), though they used a different measure than we plan to. Zerback and Fawzi (2017) similarly found an effect size of r = .11 for the difference in likelihood to post a comment when the group majority is on your side vs. opposing your position. One example in the literature on reducing toxicity suggests that actual posts become less toxic after humor is used (d = .38 and d = .36), though this isn't a measure of perceived toxicity reduction (Elsayed & Hollingshead, 2022). Strelan, Di Fiore, and Van Prooijen (2017) found an effect size of d = .93 for the difference in perceived justice between conditions where participants punish vs. cannot punish a transgressor. Schoenebeck, Haimson and Nakamura (2021) found an effect size of g = 1.07 for the perceived justice of mediation versus banning as a response to online toxicity targeting the participant. Taken together, we decided to use r = .11 (f = .11) as our smallest effect size of interest.

A power analysis using G*Power for an ANCOVA with three conditions, three covariates, an effect size of f = .11, a family-wise alpha of .05, and 90% power suggested a total sample size of 1049. Nine of the 126 pilot participants were dropped for failing an attention check, or 7% of the sample. We added an additional 7% to our sample size, resulting in a target sample of 1122 participants. We used the Benjamini-Hochberg procedure to keep the false discovery rate at 5% (Benjamini & Hochberg, 1995).

Method

Participants

We recruited 1122 participants using CloudResearch's Mechanical Turk Toolkit (Litman et al., 2017) to complete our experiment. Due to experimenter error, 238 additional people completed the study, resulting in a sample of 1360. Three respondents did not complete any questions and were dropped prior to analysis. Those with partial data were retained, though if they were missing data for an entire variable in an analysis, their data were not included in that particular analysis. Missing values were imputed using multiple imputation (Rubin, 1976; Van Buuren & Groothuis-Oudshoorn, 2011) for participants who answered at least 50% of the questions for a multi-item scale. Participants were paid $1.00 to participate and the study took roughly 15-30 minutes. No participants failed both attention checks resulting in a final sample of 1357.[6]

The sample consisted of 601 men, 740 women, 10 nonbinary individuals, two trans men, one trans woman and one who indicated "other." It was 72.79% White, 9.59% Black, 7.37% Asian, 2.51% Biracial or Multiracial, 6.19% Hispanic or Latino, and 7.15% unknown or other race.[7] The average social media use per day was 2.56 hours (SD = 2.34, min = 0, max = 24). The average comfort with offensive language on a scale from 0 (it makes me very uncomfortable) to 4 (it doesn't bother me) was 3.04 (SD = 1.17; min = 0; max = 4; see Figure 5). Participants were randomly assigned to either the Benevolent Correction condition (n = 451; none dropped), the Benevolent Going Along condition (n = 458; none dropped), or the Retaliatory condition (n = 448; none dropped). Since no participants were dropped for failing both attention checks, we did not conduct a chi-square goodness-of-fit test comparing the drop rates across conditions.

Modifications to Pilot Procedure and Materials

Our main experiment's methods draw heavily from our pilot study, with several modifications. First, we replaced one of the three items intended to measure participants' freedom to contribute ("If you were to post in this forum, to what extent would you feel the need to hide what you really think from the rest of the group?") because it did not correlate strongly with the other two in the pilot. The new question was adapted from Hampton, Shin and Lu (2017). The original item wording was "If the topic of the government's surveillance program came up, would you be very willing, somewhat willing, somewhat unwilling, or very unwilling to join the conversation?" To match the response options given for the other two items, we modified the wording to "How willing would you be to join this conversation?"

4 We initially overestimated these effect sizes because of an error in our original pilot multilevel analyses which has since been corrected in the manuscript. The actual effect sizes ranged from r = .13 to r = .38.

5 See footnote 4; the actual effect sizes ranged from r = .13 to r = .63.

6 In the Benevolent Correction condition, 369 got the first attention check correct (82%) and 81 got it wrong (18%). Our attention check in the Benevolent Going-Along condition mistakenly had two correct answers to choose between, so no one had an incorrect answer. In the Retaliation condition, 405 (90.4%) were correct and 43 (9.6%) were incorrect. Only five people (0.4%) got the second attention check question wrong, and those five people did not also get the first question wrong. No one was excluded.

7 We had a large number of people with race unknown because the race/ethnicity question was open-response. Many respondents provided only their ethnicity (e.g., Hispanic) without reporting their race.

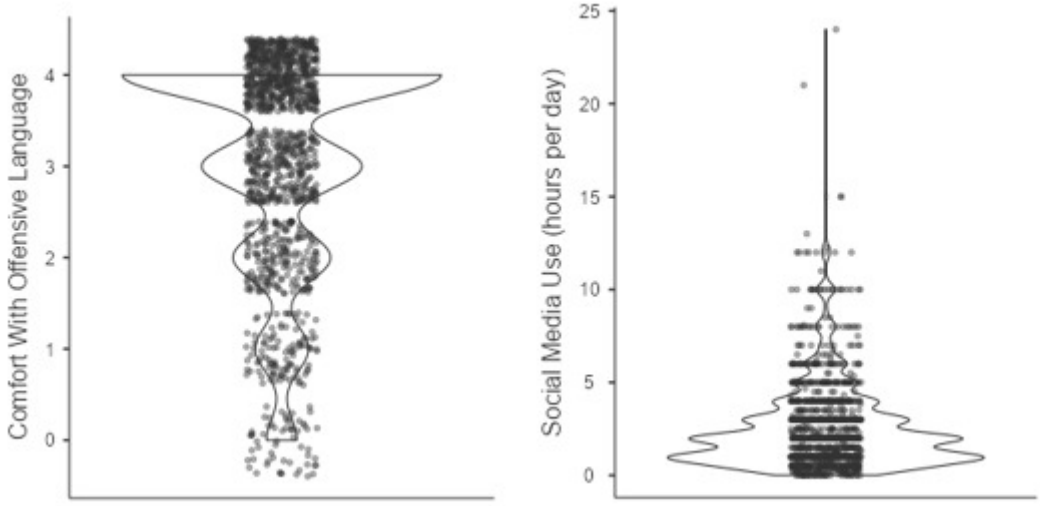

Figure 5. Distributions of Social Media Use and Comfort With Offensive Language in Main Experiment

Note. These violin plots represent 1357 participants' social media use in hours per day and comfort with offensive language (0 = it makes me very uncomfortable; 4 = it doesn't bother me).

Second, we replaced several of the twelve Reddit comment-reply pairs used in the pilot (see Appendix A) as follows. Pairs 2 and 3 in the benevolent correction condition were coded similarly by undergraduate research assistants in the extent to which the reply corrected and went along with the initial comment, so we replaced them with pairs in which the reply was rated as more correcting and not at all going along. Pairs 4, 7 and 11 were replaced based on a reviewer suggestion that we should avoid comments which mention specific topics which participants might have differing opinions on. Pair 9 was replaced because pilot participants' first impression of this toxic commenter was relatively positive compared to the other conversations (see Appendix A for the comment-reply pairs).

Third, based on reviewer feedback, we replaced the first-impression question intended to verify the toxicity of the initial comment in each pair with the more specific item wording used by Google's Perspective API when classifying the toxicity of online text (Perspective, 2021; see Appendix B). This we named perceived toxicity.

Fourth, to make our measure of perceptions that a reply has dissuaded the toxic commenter from continuing to post in the same negative tone more specific, we replaced items one and two, which were too general (e.g., asking whether toxicity had been addressed without defining what "addressed" means; asking whether someone would reconsider what they posted when that could have a variety of meanings), with more specific items (e.g., "The response will encourage the first commenter to post more positively in the future"). Fifth, due to experimenter error, we did not randomize the order of either the overall ratings of fair/just resolution items or the Willingness to Self-Censor scale (Hayes et al., 2005b) in the pilot. We randomized the order of the items within each scale here.

Finally, we added one additional attention check question before the overall ratings of fair/just resolution and three manipulation check items after the overall ratings of fair/just resolution but before the Willingness to Self-Censor scale (Hayes et al., 2005b). The first measured the extent to which the replies to the toxic comments demonstrated benevolence for the initial commenter. The second measured the extent to which the replies appeared to correct the initial comment, and the third measured the extent to which the replies retaliated against the original commenter (see Appendix B).

Otherwise, the experiment used identical materials and procedure to the pilot study. In brief, participants were randomly assigned to read four conversations (a toxic comment followed by a reply) in a randomized order from either the benevolent correct, benevolent going-along, or retaliatory condition. After each comment-reply pair, they reported their first impression of the toxic commenter, how free they felt to contribute to the conversation (3 items, order randomized, $\alpha = 0.93$) and how much they thought the toxicity had been dissuaded (4 items, order randomized, $\alpha = 0.71$). After responding to all four pairs, they completed two attention check questions, the first asking them which statement they saw in the previous conversations and the second instructing them to choose the color "green" from a list of colors rather than answering with their favorite color. They then reported their overall impression of whether justice had been restored (7 items, randomized order, $\alpha = 0.94$), completed two manipulation check questions (how correcting and how benevolent the replies in general were, order randomized), their Willingness to Self-Censor (Hayes et al., 2005b; 8 items, randomized order, $\alpha = 0.86$), comfort with offensive language (0 = it makes me very uncomfortable to 4 = it doesn't bother me at all), total time per day spent on social media for personal use, their age, gender, and race/ethnicity (open-response). Means and standard deviations for key variables are reported in Table 2.

Table 3. Factor Loadings for Free to Contribute (Factor 1) and Toxicity Dissuaded (Factor 2)

Factor	Indicator	Standardized Estimate	SE	Z	p
Toxicity Dissuaded	Item 1	0.83	0.02	68.2	< .001
	Item 2 (Reversed)*	0.15	0.03	10.3	< .001
	Item 3	0.79	0.02	63.9	< .001
	Item 4	0.77	0.02	62.2	< .001
Free to Contribute	Item 1	0.80	0.02	69.9	< .001
	Item 2	0.96	0.02	94.4	< .001
	Item 3	0.95	0.02	93.4	< .001

Note: *Item 2 from the Toxicity Dissuaded measure was deleted from the scale following factor analysis

Table 4. Intercorrelations of the Four Toxicity Dissuaded Items

	Item 1	Item 2 (Reversed)	Item 3
Item 2 (Reversed)	0.20*		
Item 3	0.64*	0.10*	
Item 4	0.63*	0.07*	0.65*

Note: Values are Pearson's r's. *$p < .001$.

Results

Scale Reliability, Unidimensionality and Composites

As with the pilot study, we assessed the reliability of the multiple-item measures (free to contribute, toxicity dissuaded, overall justice restored, willingness to self-censor) using Cronbach's alpha and averaged the items together since alpha was above .7 for each. We conducted a confirmatory factor analysis for each scale to provide evidence of unidimensionality. Using Rosseel's (2012) lavaan R package in jamovi (The jamovi project, 2022), we entered the items measuring freedom to contribute and toxicity dissuaded into a confirmatory factor analysis forcing the items to load onto one of two distinct factors. Model fit was satisfactory, RMSEA = 0.076, 95% CI [0.07, 0.08], CFI = 0.98, TLI = 0.97, $\chi^2(13) = 418$, $p < .001$,[8] but one item from the toxicity dissuaded scale, item two, did not load as strongly onto its latent factor as the others (standardized factor loading = 0.15; see Table 3). Brown (2015) suggests that loadings at or above 0.3 account for a sufficient amount of variance to be "salient" (p. 115), but this does not meet that criterion. Further, item two ("The response will encourage the first commenter to post in a more negative tone than before" (reverse-scored)) did not correlate strongly with the other scale items (all correlations with item two < .21; see Table 4) and removing it increased the scale reliability from $\alpha = .71$ to $\alpha = .84$. We therefore decided to remove item two from the toxicity dissuaded scale to improve its reliability.

We next entered the seven items measuring the perception that justice has been restored (Wenzel et al., 2010) into a confirmatory factor analysis forcing the items to load onto one factor. Wenzel et al. (2010) found in their original sample that all items strongly loaded onto a single factor. We similarly found that most items loaded strongly onto one factor (standardized estimates > 0.89) with one reverse-scored item, item 2, loading less strongly (standardized estimate = 0.40). Our model fit was good according to the CFI and TLI indicators (CFI = 0.98, TFI = 0.97), though not as good according to the RMSEA (RMSEA = 0.097, 95% CI [0.08, 0.11]. We retained all seven items. The above analyses, in sum, provided evidence of the unidimensionality of each dependent measure (free to contribute, toxicity dissuaded and justice restored).

The confirmatory factor analysis we ran on the eight-item Willingness to Self-Censor scale (Hayes et al., 2005) was less clear. When we constrained the items to load onto a single factor, model fit was poor, RMSEA = 0.13, 95% *CI* [0.12, 0.14], CFI = 0.90, TLI = 0.86. Each item, though, loaded strongly onto that factor (standardized estimates > 0.58). Since the original scale is intended to be unidimensional and has been validated (Hayes et al., 2005a) and used as a single-factor scale many times since (e.g., Etchegaray et al., 2019; Stoycheff, 2016), we chose to treat it as unidimensional here but will return to this issue in our discussion.

8 That the chi-square test of perfect model fit is significant is not necessarily indicative of an incorrect model given our large sample size. The larger the sample size, the more likely this test is to be significant, according to Babyak and Green (2010): "If the sample size is large, the *T* value will necessarily be large, and even small and possibly unimportant discrepancies between the model implied and observed covariance matrix will yield significance. It is our observation that tests of models are routinely significant—meaning that we conclude our model does not fit—when sample size exceeds 200."

Manipulation Checks

We first tested whether the comments in the two benevolence conditions were rated as more benevolent than the retaliatory condition using a between-subjects Welch's ANOVA predicting benevolence score from condition. We found a significant effect of condition, $F(2, 895) = 331$, $p < .001$, $\eta^2 = .43$. As expected, the retaliatory condition ($M = 1.21$, $SD = 1.62$) was rated as less benevolent than either the benevolent correction ($M = 3.95$, $SD = 1.57$; $t(1352) = 23.7$, $p_{Tukey} < .001$, $d = 1.72$) or the benevolent going-along condition ($M = 2.68$, $SD = 1.98$; $t(1352) = 12.7$, $p_{Tukey} < .001$, $d = 0.81$). The benevolent correction condition was also rated as more benevolent than the benevolent going-along condition, $t(1352) = -11.1$, $p_{Tukey} < .001$, $d = 0.71$.

Second, to test whether the benevolent correction condition conversations were perceived as more correcting than the benevolent going-along condition, we conducted an independent-samples *t*-test comparing the two group means. We found a significant difference between conditions, $t(903) = -25.3$, $p < .001$, $d = -1.68$. The benevolent correction replies were rated as more correcting ($M = 4.14$, $SD = 1.51$) than those in the benevolent going-along condition ($M = 1.38$, $SD = 1.77$).

Finally, we conducted a between-subjects Welch's ANOVA comparing the three conditions on how retaliatory they appeared to be. We found a significant effect of condition, $F(2, 883) = 432$, $p < .001$. The retaliatory replies were rated as more retaliatory ($M = 4.68$, $SD = 1.44$) than either the Benevolent Correction ($M = 2.19$, $SD = 1.92$), $t(1351) = -20.9$, $p_{Tukey} < .001$, or the Benevolent Going Along replies ($M = 1.71$, $SD = 1.96$); $t(1351) = -25.0$, $p_{Tukey} < .001$. These findings suggest that we successfully manipulated how benevolent and how correcting the Reddit conversations were.

Per-Pair Ratings

As with the pilot, we conducted multilevel regression models nesting ratings within pair (1-12) and participant predicting ratings of each separate conversation. Condition (benevolent correction vs. benevolent going-along vs. retaliatory) was a between-subjects fixed factor in each model. Each analysis involved four scores per 1356 people, or 5424 observations. We used the Benjamini-Hochberg procedure to keep the false discovery rate at 5% (Benjamini & Hochberg, 1995). As per a recommendation from Segerstrom (2019), analyses were conducted both including the covariates (perceived toxicity of the initial comment, willingness to self-censor, and comfort with offensive language) and without, and the effect of condition was reported for both. Summary descriptive statistics for each dependent measure across conditions can be found in Table 5.

First Impression of Toxic Commenter. We conducted a multilevel regression predicting the perceived toxicity of the initial comment from condition ($ICC_{Participant} = 0.56$; $ICC_{Pair} = 0.10$). The effect of condition was not significant, $F(2, 10.9) = 1.94$, $p = .19$. The initial comments did not differ in perceived toxicity between the two benevolent conditions ($p = .90$), the Retaliatory vs. the Benevolent Correction condition ($p = .13$) or the Retaliatory vs. the Benevolent Going Along condition ($p_{Bonferroni} = .32$). Because of this, we did not control for the perceived toxicity of the initial comment in any subsequent analyses.

Free to Contribute. We conducted a multilevel regression predicting how free participants felt to contribute to the conversation from condition controlling for willingness to self-censor and comfort with offensive language (ICC = 0.65). We found a significant difference among conditions, $F(2, 20.6) = 7.70$, $p = .003$; without covariates, $p = .003$; see Figure 6. Planned comparisons suggested that, inconsistent with Hypothesis 1, participants did not feel less free to contribute in the benevolent going-along condition ($M = 1.69$, $SE = 0.12$, 95% CI [1.45, 1.94]) than the benevolent correction condition ($M = 2.01$, $SE = 0.12$, 95% *CI* [1.76, 2.25]), $t(20.5) = 1.85$, $p = .078$, $d = 0.10$; without covariates, $p = .069$. Also inconsistent with Hypothesis 1, participants' freedom to contribute did not differ between the benevolent going-along condition and the retaliatory condition; in fact, the trend was in the opposite direction from our prediction (retaliatory $M = 1.34$, $SE = 0.12$, 95% *CI* [1.10, 1.59]), $t(20.6) = -2.08$, $p = .051$, $d = 0.11$; without covariates, $p = .052$. Also inconsistent with Hypothesis 1, participants felt more free to contribute in the benevolent correction condition than in the retaliatory condition according to a post hoc comparison, $t(20.7) = 3.92$, $p = .002$, $d = 0.22$, without covariates, $p_{Bonferroni} = .002$. We had expected these two conditions not to differ. Willingness to Self-Censor was not related to how free participants felt to contribute, $b = -0.07$, $SE = 0.05$, 95% *CI* [-0.17, -0.02], $t(1350) = -1.50$, $p = .135$, and comfort with offensive language was positively related to it, $b = 0.10$, $SE = 0.04$, 95% *CI* [0.03, 0.17], $t(1350) = 2.65$, $p = .008$.

Toxicity Dissuaded. We conducted a multilevel regression predicting the extent to which participants felt the reply discouraged future toxicity from condition controlling for willingness to self-censor and comfort with offensive language ($ICC_{Participant} = 0.64$; $ICC_{Pair} = 0.05$).We found a significant difference among conditions, $F(2, 14.6) = 29.69$, $p < .001$, without covariates, $p < .001$; see Figure 7. Planned comparisons suggested that, consistent with Hypothesis 2a and 2b, the benevolent going-along condition mean ($M = -1.25$, $SE = 0.13$, 95% *CI* [-1.52, -0.98]) was significantly, moderately lower than the benevolent correction condition ($M = 0.005$, $SE = 0.13$, 95% *CI* [-0.26, 0.27]), $t(14.5) = -7.10$, $p < .001$, $d = 0.53$, without covariates, $p < .001$. According to a post hoc comparison, inconsistent with Hypothesis 2a, the benevolent going-along did not differ from the retaliatory condition mean ($M = -1.08$, $SE = 0.13$, 95% *CI* [-1.35, -0.81]), $t(14.5) = -0.96$, $p_{Bonferroni} = 1.00$, $d = .07$, without covariates, $p_{Bonferroni} = 1.00$. A planned comparison suggested that, consistent with Hypothesis 2b, the benevolent correction condition mean was significantly higher than the retaliatory condition mean, $t(14.6) = -6.14$, $p < .001$, $d = .46$, without covariates, $p < .001$. Willingness to Self-Censor was positively related to the perception that toxicity was dissuaded, $b = 0.20$, $SE = 0.04$, 95% *CI* [0.12, 0.29], $t(1349.8) = 4.85$, $p < .001$. Comfort with offensive language was negatively related to the perception that toxicity was dissuaded,

Table 5. Descriptive Statistics for Dependent Measures By Condition In Main Experiment

	Benevolent Correction (n = 451)	Benevolent Going-Along (n = 458)	Retaliatory (n = 448)
Perceived toxicity of initial comment	-0.67 (0.11)	-0.66 (0.11)	-0.92 (0.11)
Free to contribute per comment-reply pair	2.01^{A} (0.12)	1.69^{AB} (0.12)	1.34^{B} (0.12)
Toxicity dissuaded per comment-reply pair	$0.005\ (0.13)^{A}$	-1.25^{B} (0.13)	-1.08^{B} (0.13)
Justice restored	0.64^{A} (0.06)	-1.06^{B} (0.06)	-1.14^{B} (0.06)

Note: Standard errors in parentheses. Means in the same row which share a superscript letter do not differ at the .05 level.

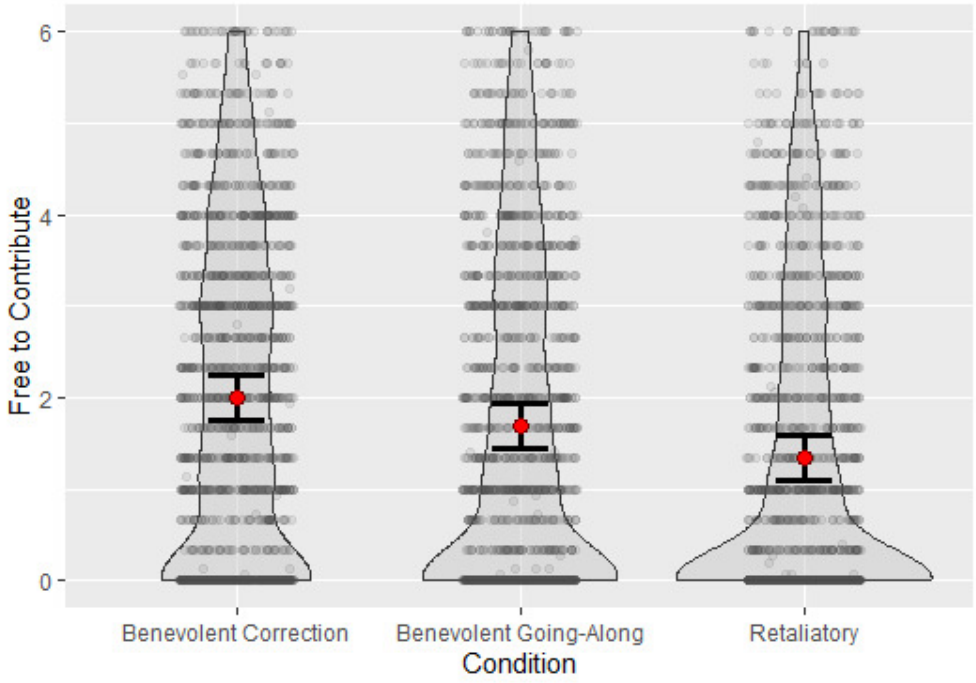

Figure 6. Free to Contribute Across Conditions in Main Experiment

Note. Error bars represent 95% confidence intervals. Red circles represent condition means; grey dots are data points jittered for visibility with density represented by darker greys.

b = -0.09, SE = 0.03, 95% CI [-0.15, -0.03], t(1349.8) = -2.91, p = .004. Hypothesis 2a was therefore disconfirmed but Hypothesis 2b was supported (see Table 6).

Word Count. Though we did not preregister controlling for the number of words in the comment-reply pairs, it did vary across condition (see Appendix A). When we included word count as a covariate, it did not change the significance or pattern of findings for either the Free to Contribute or Toxicity Dissuaded analyses reported above (see Supplementary Materials).

Overall Ratings: Justice Restored

We conducted a between-subjects ANCOVA predicting the overall rating of whether justice had been restored for all four conversations from condition controlling for willingness to self-censor and comfort with offensive language. We found evidence of at least one difference among the condition means, F(2, 1345) = 245.92, p < .001, η^2 = 0.27, without covariates, p < .001; see Figure 8. Planned comparisons suggested, consistent with Hypothesis 3a, that benevolent corrections were rated as providing a more just resolution (M = 0.64, SE = 0.06, 95% CI [0.51, 0.77]) than benevolently going-along (M = -1.06, SE = 0.06, 95% CI [-1.19, -0.94]), t(1345) = -18.8, p < .001, d = 1.25 (without covariates, p < .001). Also consistent with Hypothesis 3a, benevolent corrections were rated as providing a more just resolution than retaliatory (M = -1.14, SE = 0.06, 95% CI [-1.26, -1.01]) replies, t(1345) = -19.6, p < .001, d = 1.31 (without covariates, p < .001). We did not find evidence that the benevolent going-along and retaliatory replies differed (p_{Tukey} = .68; without covariates, p_{Tukey} = .59). Comfort with offensive language was positively related to just resolution ratings, b = 0.07, F(1, 1345) = 5.26, p = .022, η^2 = 0.003 and willingness to self-censor was not a significant predictor (p = .17).

Main Experiment Discussion

Hypothesis 1 and 3b were disconfirmed. We found support for Hypotheses 2a, 2b, and 3a. We discuss these specific findings below.

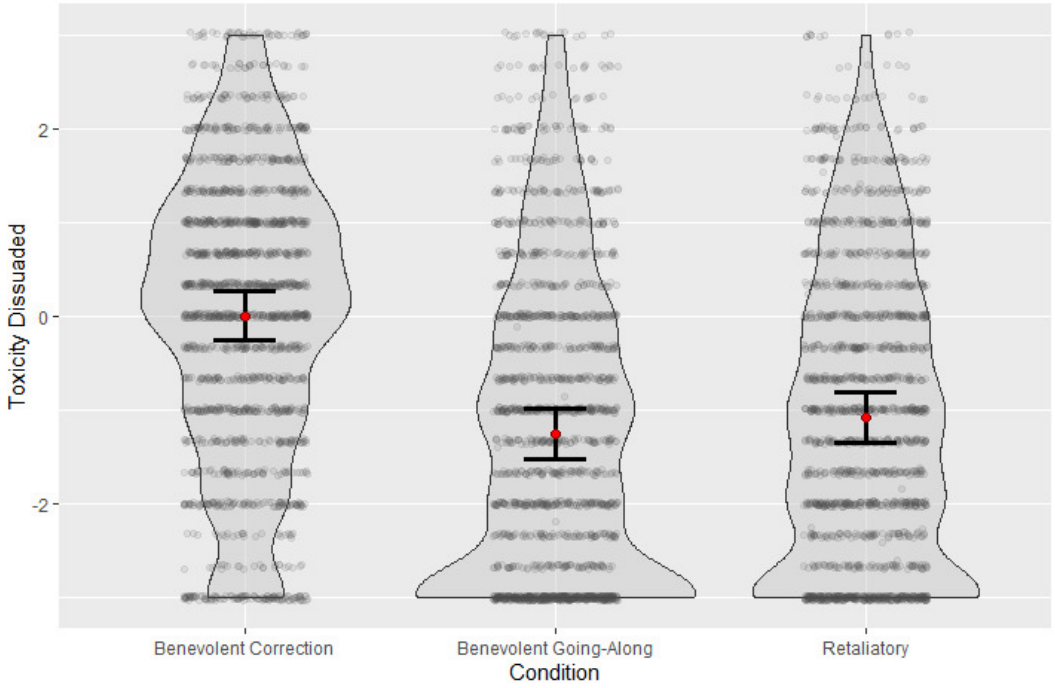

Figure 7. Toxicity Dissuaded Across Conditions in Main Experiment

Note. Error bars represent 95% confidence intervals. Scale reflects opinion that toxicity has been dissuaded from -3 (Strongly disagree) to +3 (Strongly agree). Red circles represent condition means; grey dots are data points jittered for visibility with density represented by darker greys.

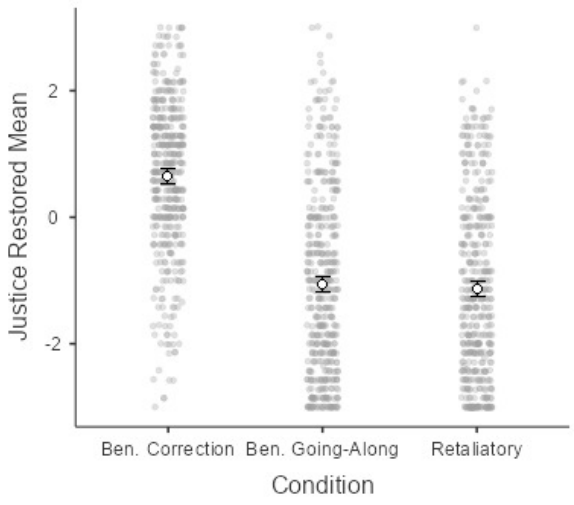

Figure 8. Overall Justice Restored Across Conditions in Main Experiment

Note. Error bars represent 95% confidence intervals. Scale reflects opinion that justice has been restored from -3 (Strongly disagree) to +3 (Strongly agree).

Hypothesis 1: Free to Contribute

Spiral of Silence Theory (Noelle-Neumann, 1977) holds that when individuals feel their viewpoint lacks social support, they are less likely to express it. We had predicted that how free participants feel to share *any* opinion following a toxic comment is potentially renewed if the toxic comment is corrected by another person (as with a retaliatory or benevolently-correcting reply). Reading a benevolent but non-correcting reply, on the other hand, may signal tacit approval of the toxicity, leaving others feeling they still cannot share their own viewpoint. Inconsistent with this, we did not find that correcting replies were more helpful than non-correcting ones. Instead, we found that benevolent corrections made participants feel freer to contribute than retaliating against the toxic comment. Note that this effect size was small; those in the benevolent correction felt only two-thirds of a scale point more free (M = 2.01 on a scale from 0 = not at all to 6 = very much) than those in the retaliatory condition (M = 1.34), but still below the scale midpoint of 3. .

Hypotheses 2a and 2b: Toxicity Dissuaded

Using the Focus Theory of Normative Conduct (Cialdini et al., 1991), we had predicted that leaving toxicity uncorrected would be perceived as least effective at dissuading toxicity (Hypothesis 2a) and that correcting toxicity in a benevolent way would be seen as especially likely to dissuade toxicity (Hypothesis 2b). Only Hypothesis 2b was supported, though participants did not see any of the three options as particularly effective fixes to the toxicity problem. Replies which benevolently corrected the toxicity resulted in a mean toxicity dissuaded rating near the scale midpoint, suggesting that participants had a neutral take on their effectiveness; they neither agreed nor disagreed that benevolent corrections would lead to less toxicity from that commenter. Replies which retaliated against or went along with the toxicity were seen as actively unhelpful by comparison and did not differ from each other.

Hypotheses 3a and 3b: Overall Justice Restored

Our participants preferred replies consistent with restorative justice over those consistent with retributive justice. Supporting Hypothesis 3a, they believed that justice had been restored most in the benevolent correction condition. The effect sizes of the difference between the benevolent correction condition mean and each other condition were large, corresponding to an increase in roughly one and a half scale points on a seven-point scale. Inconsistent with Hypothesis 3b, the other options (retaliating and not correcting at all) were seen as less effective in restoring justice and did not differ from each other.

General Discussion

We have overall evidence that certain methods of replying to online toxicity are perceived as more helpful than others. Participants clearly preferred replies which benevolently corrected toxic comments; these replies led participants to feel freer to contribute to the conversation than retaliatory replies, feel most that the toxicity had been dissuaded, and feel most that justice had been restored. That benevolent corrections were seen as best at restoring justice was corroborated by our pilot study data. Correcting a toxic comment in a toxic, retaliatory way did not increase freedom to contribute to the conversation relative to not correcting it at all (counter to our prediction) and was no more effective than a non-corrective, benevolent reply at helping participants feel that toxicity had been dissuaded or in restoring a sense of justice.

Freedom to Contribute After Someone Retaliates or Goes Along

Why did people not differ in how free they felt to join the conversation after reading a retaliatory vs. non-corrective (going-along) reply to toxicity? Wouldn't a retaliatory reply signal to observers that toxicity is not acceptable and encourage them to contribute? Spiral of Silence Theory predicts that on controversial issues, given the impression of a united majority, those who believe they hold a minority viewpoint become less and less likely to speak publicly (Noelle-Neumann, 1977). Eilders & Porten-Cheé (2015) argue that in online environments, user comments might serve as especially strong cues of public opinion and, if uncivil, may lead others to withhold their opinion to avoid social isolation. Exposure to toxic comments should, then, reduce freedom to contribute. Our data appear to support this; the average freedom to contribute across all conditions was well below the scale midpoint.

We had assumed that any correction to this toxicity, whether benevolent or retaliatory, would signal that toxicity is not socially acceptable, counteracting the perceived risk of social isolation the toxicity originally presented. The retaliatory replies in our stimuli, though, were also toxic; they were sarcastic, snide, and mean. They may have served to, in fact, bolster the sense that the forum was uncivil and suppress participants' desire to share their opinion.

Benevolent Corrections as Injunctive Norms

Participants believed most that the toxic commenter would stop posting toxic things after it was benevolently corrected and saw it as more effective than either other reply type. This is consistent with the possibility that benevolent corrections highlight an injunctive norm (Cialdini et al., 1991) – e.g., this is how we ought to behave in this forum – which participants see as especially likely to exert normative influence on the toxic poster's behavior. Given this expectation of relative success, it is interesting that benevolent correction is not a common strategy. Young Reusser et al. (2021), as reported above, found that roughly 38% of replies to the most toxic Reddit comments in January, 2016 were benevolent, and our further analyses suggested that only about half of these were also corrective (about one-fifth of all replies to the most toxic comments).

Similarly, Mathew et al. (2019) found that using an empathic, kind, polite or civil tone to counter hate speech (known as counterspeech) accounted for only 9% of the YouTube comments in their sample, whereas hostile responses to hate speech occurred more frequently, about 30% of the time. Clearly, believing a reply strategy is effective does not guarantee a person will use it. Future research might focus on a) whether benevolent corrections actually *do* discourage toxicity, something our data do not speak to, and b) if so, how the use of benevolent corrections might be encouraged.

Retaliating Not Seen As An Effective Deterrent

Our participants did not, on average, rate the retaliatory replies as any more effective at dissuading the toxicity than replies which did not correct the commenter at all. This is inconsistent with our prediction that a retaliatory reply might highlight an injunctive norm (toxicity is not acceptable) but a reply which goes along with the toxic comment does not (indicating tacit approval of the toxicity). Perhaps, as Benesh et al. (2016) argue, our respondents believed that hostility might entrench the commenter, leading them to be as or even more toxic than before. Our data suggests that in order to dissuade toxic commenters, a responder should correct in a polite rather than aggressive way, something that could be understandably difficult if the responder finds the toxic comment offensive or feels attacked by it.

Implications of Benevolently Correcting Someone's Toxicity

The benevolent corrections in our stimuli do not direct the toxic commenter to leave the forum; rather they directly but respectfully call the person out for the toxicity (e.g., "That's not a very nice thing to say") and discourage the toxicity (e.g., "If you don't have anything nice to say, don't say it"). In other words, the person who made the toxic comment is tacitly "allowed" by such a replier to continue to post in the forum as long as their behavior changes. This highlights the possibility that benevolently correcting toxicity could allow everyone in a particular forum to move forward and have a healthier conversation.

Table 6. Summary of Main Experiment Results Per Each Hypothesis

Measure	Hypothesis	Interpretation given different outcomes
Freedom to contribute	H1. Participants will feel freer to contribute to a conversation initiated by a specific toxic comment after a Benevolent Correction or Retaliatory reply compared to a reply that Benevolently Goes Along with the toxic comment.	Support for H1: The Benevolently Going Along condition's mean is lower than the other two at the .05 level. **It did not differ from either other condition. Hypothesis 1 disconfirmed.** However, if the Benevolent Correction condition's mean is significantly higher than the Retaliatory condition's mean, that might suggest that the polite tone of the correction provides additional incentive to contribute. **This was the case.** If instead the Retaliatory condition's mean is higher than the Benevolent Correction condition's mean, that might suggest that a negative tone is more likely to encourage others to respond. **This was not the case.**
Perception that the toxicity has been dissuaded	Two possibilities: H2. Participants will believe to a greater extent that toxicity has been dissuaded… a. when the replier Benevolently Corrects or Retaliates compared to Benevolently Going Along. b. when the replier Benevolently Corrects compared to either alternative (Benevolently Going Along or Retaliating).	Support for H2a: The Benevolently Going Along condition's mean is lower than the other two at the .05 level. **Benevolently Going Along mean was lower than benevolent correction but NOT lower than retaliatory. Hypothesis 2a disconfirmed.** Support for H2b: The Benevolent Correction condition's mean is higher than the other two at the .05 level. **This was the case. Hypothesis 2b confirmed.**
Perception that justice has been restored	Two possibilities: a. Benevolent Corrections (vs. Benevolently Going Along or Retaliating) will make participants feel more that justice has been restored b. Retaliatory responses (vs. Benevolently Correcting or Going Along) will make participants feel more that justice has been restored	Support for H3a: The Benevolent Correction condition's mean is higher than the other two at the .05 level. **This was the case.** Support for H3b: The Retaliatory condition's mean is higher than the other two at the .05 level. **This was not the case.**
Manipulation check - benevolence	Ensure that the benevolent replies are rated as more benevolent than the retaliatory replies	The two Benevolent conditions should be rated as more benevolent than the Retaliatory condition at the .05 level. **This was the case.**
Manipulation check - correcting	Ensure that the Benevolently Correcting replies are rated as more correcting than the benevolently Going Along replies	The Benevolent Correction condition should be rated as more correcting of the initial comment than the Benevolently Going Along condition at the .05 level. **This was the case.**
Manipulation check – retaliatory	Ensure that the Retaliatory replies are rated as more retaliatory than either other condition	The Retaliatory condition should be rated as more retaliatory than the other two at the .05 level. **This was the case.**
Manipulation check – first impression of toxic commenter (pilot); perceived toxicity of initial comment (Main experiment)	Ensure the participant's first impression toxic commenter (pilot) or the perceived toxicity of the initial comment (Main experiment) is similar across conditions. If this is not the case, first impression will be controlled for.	If the first impression of each toxic commenter differs by condition at the .05 level, this will be included as a covariate in the main analyses for RQ1, RQ2 and RQ3. **The first impression did vary by condition and this was controlled for in the main analyses.**

Restoring Justice After Toxicity

Our largest effect sizes were found when comparing the three reply types on their perceived ability to restore justice and fairness in the forum. Benevolent corrections were clearly seen as the best option in both our well-powered, online sample and our small undergraduate pilot sample. Consistent with a restorative justice approach, participants on average agreed that responding to toxicity with under-

standing, politeness, and empathy restored justice, whereas they disagreed that going along with or retaliating against the toxicity did so.

Recent research on other online platforms has uncovered a similar preference for restorative over retributive justice outcomes. Xiao, Cheshire and Salehi (2022) interviewed 28 West Coast undergraduate students asking how online harm they had experienced (e.g., harassment, trolling, offensive name-calling) on platforms like Instagram, Facebook, and Tiktok could be addressed. While many of their participants brought up retribution of some kind (e.g., reporting or calling out the harm-doer), the majority did not. Further, those who did mention retribution placed it as a secondary concern behind things like making sense of the situation and receiving support from others.

Limitations and Future Directions

Since we recruited our main sample from Mechanical Turk via CloudResearch, it is not representative of the general U.S. population and our findings cannot be generalized that widely. White participants, for example, are over-represented (see Peer et al., 2023 for typical demographics of various online sampling platforms). However, participants spent, on average, more than two hours per day on social media. Their responses should provide some idea of how other social media users might view responses to online toxicity.

To increase ecological validity, we presented participants with real comments sampled from Reddit. However, they were not long conversations, consisting only of a single toxic comment followed by one reply. Future studies should use longer conversations as stimuli to see whether the condition differences we found are still apparent, get weaker or stronger, etc. as the conversation continues. In addition, these comments were from a particular month in a particular year on a specific social media site and may not be representative of other times or platforms. Future studies could manipulate the platform type to see if the norms of a given site change whether, for instance, a benevolent correction is always seen as the most just response.

While we tried to ensure that the toxic comments were as similar as possible in toxicity across conditions, those in the Retaliatory condition were rated as more toxic on average. We statistically controlled for comment toxicity, but future studies could use stimuli that better control for this confound. Since they are unedited, the comments in the three conditions also differ in word count; word count is therefore a potential confound. Benevolent corrections were longer on average (22.75 words) than replies that benevolently went along (16.25) or retaliated (16.5 words). When we statistically controlled for word count, our findings were unchanged (see Supplementary Materials).

We intentionally selected conversations based on reviewer feedback which were not obviously about a particular topic so that any differences across condition couldn't be explained by familiarity with or agreement/disagreement with a particular position. This means, though, that participants might not have been as invested in the comments as they would be in their own online conversations. They also may have found it difficult to understand what the initial commenter was talking about (other than that they were being mean about something). Future work could sacrifice ecological validity to craft stimuli that are easier to understand, which center on interesting or controversial topics, but control for potential confounds like word count, topic area, familiarity, etc.

It is important to note that our dependent measures of how free people felt to contribute and whether toxicity had been dissuaded were measuring their expectations of future events rather than change in an actual conversation. Future research should follow participants through a longer conversation to see whether they actually contribute more in a conversation if someone has just benevolently corrected a toxic comment. Research could also include individuals who post toxic things to see whether a benevolent correction actually does change their behavior compared to other sorts of replies.

Willingness to Self-Censor Unidimensionality

One last limitation has to do with whether Willingness to Self-Censor (Hayes et al., 2005b) was unidimensional in our sample. As mentioned above, confirmatory factor analysis suggested poor fit for a single-factor model. However, the scale was written to measure a single construct and has been validated; Hayes et al., 2005a, for instance, found that individuals who scored higher on the scale were more sensitive to whether other people in a hypothetical conversation all agreed or all disagreed with them when deciding whether they were willing to share their own opinion. Hayes, Uldall and Glynn (2010) found evidence that only participants who scored high on the scale self-censored during in-person conversations with confederates who disagreed with them. Even assuming the scale is not unidimensional, none of our findings changed when our covariates – including Willingness to Self-Censor – were removed from analyses.

Conclusion: Intervening When Others are Toxic

Our data provide evidence that when toxicity occurs online, bystanders can play a key role in addressing the problem. If they correct the toxicity in a way that is polite, understanding, and empathic, they can help others in the conversation feel freer to contribute, believe that the toxicity will decrease, and think that justice has been restored compared to retaliating or going along with the toxicity. Since bystander intervention in these cases is seen by many to be an important strategy (e.g., Pew Research Center, 2017; Xiao et al., 2022), knowing what sorts of responses are seen as effective can lead to better experiences for all.

Contributions

Contributed to conception and design: AYR, KMV, EAG, JPC

Contributed to acquisition of data: AYR, KMV

Contributed to analysis and interpretation of data: AYR
Drafted and/or revised the article: AYR
Approved the submitted version for publication: AYR, KMV, EAG, JPC

Funding

This work is funded by a Networking grant from the Council of Christian Colleges and Universities.

Conflicts of Interest

The authors have no conflicts of interest to disclose.

Data Accessibility Statement

All the stimuli, measures, participant data (for the pilot; for the main experiment), jamovi output and analysis scripts can be found on this paper's project page on the Open Science Framework: (https://osf.io/6dwjx/).

Submitted: November 13, 2023 PST, Accepted: December 01, 2023 PST

References

Almerekhi, H., Kwak, H., Salminen, J., & Jansen, B. J. (2020). Are these comments triggering? predicting triggers of toxicity in online discussions. In *Proceedings of The Web Conference 2020* (pp. 3033–3040). https://doi.org/10.1145/3366423.3380074

Babyak, M. A., & Green, S. B. (2010). Confirmatory factor analysis: An introduction for psychosomatic medicine researchers. *Psychosomatic Medicine*, *72*(6), 587–597. https://doi.org/10.1097/psy.0b013e3181de3f8a

Bao, J., Wu, J., Zhang, Y., Chandrasekharan, E., & Jurgens, D. (2021). Conversations gone alright: Quantifying and predicting prosocial outcomes in online conversations. In *Proceedings of the Web Conference 2021* (pp. 1134–1145). https://doi.org/10.1145/3442381.3450122

Benesch, S., Ruths, D., Dillon, K. P., Saleem, H. M., & Wright, L. (2016). Considerations for successful counterspeech. *Dangerous Speech Project*. https://dangerousspeech.org/wp-content/uploads/2016/10/Considerations-for-Successful-Counterspeech.pdf

Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society: Series B (Methodological)*, *57*(1), 289–300. https://doi.org/10.1111/j.2517-6161.1995.tb02031.x

Brown, T. A. (2015). *Confirmatory factor analysis for applied research*. Guilford publications.

Chmiel, A., Sobkowicz, P., Sienkiewicz, J., Paltoglou, G., Buckley, K., Thelwall, M., & Hołyst, J. A. (2011). Negative emotions boost user activity at BBC forum. *Physica A: Statistical Mechanics and Its Applications*, *390*(16), 2936–2944. https://doi.org/10.1016/j.physa.2011.03.040

Cialdini, R. B., Kallgren, C. A., & Reno, R. R. (1991). 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* (Vol. 24, pp. 201–234). Academic Press. https://doi.org/10.1016/s0065-2601(08)60330-5

Cialdini, R. B., Reno, R. R., & Kallgren, C. A. (1990). A focus theory of normative conduct: Recycling the concept of norms to reduce littering in public places. *Journal of Personality and Social Psychology*, *58*(6), 1015–1026. https://doi.org/10.1037/0022-3514.58.6.1015

Cote, A. C. (2017). "I can defend myself" women's strategies for coping with harassment while gaming online. *Games and Culture*, *12*(2), 136–155. https://doi.org/10.1177/1555412015587603

Eilders, C., & Porten-Cheé, P. (2015). The spiral of silence revisited. In *Political Communication in the Online World* (pp. 88–102). Routledge. https://doi.org/10.4324/9781315707495-7

Elsayed, Y., & Hollingshead, A. B. (2022). Humor reduces online incivility. *Journal of Computer-Mediated Communication*, *27*(3). https://doi.org/10.1093/jcmc/zmac005

Etchegaray, N., Scherman, A., & Valenzuela, S. (2019). Testing the hypothesis of "impressionable years" with willingness to self-censor in Chile. *International Journal of Public Opinion Research*, *31*(2), 331–348. https://doi.org/10.1093/ijpor/edy012

Govier, T. (1999). Forgiveness and the Unforgivable. *American Philosophical Quarterly*, *36*(1), 59–75. https://www.jstor.org/stable/20009953

Gromet, D. M., & Okimoto, T. G. (2014). Back into the Fold: The Influence of Offender Amends and Victim Forgiveness on Peer Reintegration. *Business Ethics Quarterly*, *24*(3), 411–441. https://doi.org/10.5840/beq20147814

Hampton, K. N., Shin, I., & Lu, W. (2017). Social media and political discussion: when online presence silences offline conversation. *Information, Communication & Society*, *20*(7), 1090–1107. https://doi.org/10.1080/1369118x.2016.1218526

Hangartner, D., Gennaro, G., Alasiri, S., Bahrich, N., Bornhoft, A., Boucher, J., Demirci, B. B., Derksen, L., Hall, A., Jochum, M., Munoz, M. M., Richter, M., Vogel, F., Wittwer, S., Wüthrich, F., Gilardi, F., & Donnay, K. (2021). Empathy-based counterspeech can reduce racist hate speech in a social media field experiment. *Proceedings of the National Academy of Sciences*, *118*(50). https://doi.org/10.1073/pnas.2116310118

Hayes, A. F., Glynn, C. J., & Shanahan, J. (2005a). Validating the willingness to self-censor scale: Individual differences in the effect of the climate of opinion on opinion expression. *International Journal of Public Opinion Research*, *17*(4), 443–455. https://doi.org/10.1093/ijpor/edh072

Hayes, A. F., Glynn, C. J., & Shanahan, J. (2005b). Willingness to self-censor: A construct and measurement tool for public opinion research. *International Journal of Public Opinion Research*, *17*(3), 298–323. https://doi.org/10.1093/ijpor/edh073

Hayes, A. F., Uldall, B. R., & Glynn, C. J. (2010). Validating the willingness to self-censor scale II: Inhibition of opinion expression in a conversational setting. *Communication Methods and Measures*, *4*(3), 256–272. https://doi.org/10.1080/19312458.2010.505503

Hershcovis, M. S., Cameron, A.-F., Gervais, L., & Bozeman, J. (2018). The effects of confrontation and avoidance coping in response to workplace incivility. *Journal of Occupational Health Psychology*, *23*(2), 163–174. https://doi.org/10.1037/ocp0000078

Kolhatkar, V., & Taboada, M. (2017). Constructive language in news comments. In *Proceedings of the first workshop on abusive language online* (pp. 11–17). https://doi.org/10.18653/v1/w17-3002

Liang, L. H., Brown, D. J., Lian, H., Hanig, S., Ferris, D. L., & Keeping, L. M. (2018). Righting a wrong: Retaliation on a voodoo doll symbolizing an abusive supervisor restores justice. *The Leadership Quarterly*, *29*(4), 443–456. https://doi.org/10.1016/j.leaqua.2018.01.004

Litman, L., Robinson, J., & Abberbock, T. (2017). TurkPrime.com: A versatile crowdsourcing data acquisition platform for the behavioral sciences. *Behavior Research Methods*, *49*(2), 433–442. https://doi.org/10.3758/s13428-016-0727-z

Mathew, B., Saha, P., Tharad, H., Rajgaria, S., Singhania, P., Maity, S. K., Goyal, P., & Mukherjee, A. (2019). Thou shalt not hate: Countering online hate speech. In *Proceedings of the international AAAI conference on web and social media* (Vol. 13, pp. 369–380). https://doi.org/10.1609/icwsm.v13i01.3237

Mohan, S., Guha, A., Harris, M., Popowich, F., Schuster, A., & Priebe, C. (2017). The impact of toxic language on the health of reddit communities. In *Canadian Conference on Artificial Intelligence* (pp. 51–56). Springer. https://doi.org/10.1007/978-3-319-57351-9_6

Molnar, A., Chaudhry, S., & Loewenstein, G. F. (2020). *"It's not about the money. It's about sending a message!" Unpacking the components of revenge* (CESifo Working Paper No. 8102). https://doi.org/10.2139/ssrn.3541450

Noelle-Neumann, E. (1977). Turbulences in the climate of opinion: Methodological applications of the Spiral of Silence Theory. *Public Opinion Quarterly*, *41*(2), 143–158. https://doi.org/10.1086/268371

Peer, E., Rothschild, D., & Gordon, A. (2023). *Behavioral Lab 3.0: Towards the next generation of online behavioral research*.

Perspective. (2021). *Attributes and Languages*. https://developers.perspectiveapi.com/s/about-the-api-attributes-and-languages

Pew Research Center. (2017). *Online Harassment 2017*. https://www.pewresearch.org/internet/2017/07/11/online-harassment-2017/

Pew Research Center. (2021). *The State of Online Harassment*. https://www.pewresearch.org/internet/2021/01/13/the-state-of-online-harassment/

Porten-Cheé, P., & Eilders, C. (2015). Spiral of silence online: How online communication affects opinion climate perception and opinion expression regarding the climate change debate. *Studies in Communication Sciences*, *15*(1), 143–150. https://doi.org/10.1016/j.scoms.2015.03.002

Rosseel, Y. (2012). lavaan: An R package for structural equation modeling. *Journal of Statistical Software*, *48*(2), 1–36. https://doi.org/10.18637/jss.v048.i02

Rubin, D. B. (1976). Inference and missing data. *Biometrika*, *63*(3), 581–592. https://doi.org/10.1093/biomet/63.3.581

Saarento, S., Kärnä, A., Hodges, E. V. E., & Salmivalli, C. (2013). Student-, classroom-, and school-level risk factors for victimization. *Journal of School Psychology*, *51*(3), 421–434. https://doi.org/10.1016/j.jsp.2013.02.002

Salehabadi, N. (2019). *The impact of toxic replies on Twitter conversations* [Doctoral dissertation]. The University of Texas at Arlington.

Schoenebeck, S., Haimson, O. L., & Nakamura, L. (2021). Drawing from justice theories to support targets of online harassment. *New Media & Society*, *23*(5), 1278–1300. https://doi.org/10.1177/1461444820913122

Segerstrom, S. C. (2019). Statistical guideline #3: Designate and justify covariates a priori, and report results with and without covariates. *International Journal of Behavioral Medicine*, *26*(6), 577–579. https://doi.org/10.1007/s12529-019-09811-5

Stoycheff, E. (2016). Under surveillance: Examining Facebook's spiral of silence effects in the wake of NSA internet monitoring. *Journalism & Mass Communication Quarterly*, *93*(2), 296–311. https://doi.org/10.1177/1077699016630255

Strelan, P., Di Fiore, C., & Prooijen, J.-W. V. (2017). The empowering effect of punishment on forgiveness. *European Journal of Social Psychology*, *47*(4), 472–487. https://doi.org/10.1002/ejsp.2254

The jamovi project. (2022). *jamovi* (Version 2.3). [Computer Software]. https://www.jamovi.org

Van Buuren, S., & Groothuis-Oudshoorn, K. (2011). mice: Multivariate imputation by chained equations in R. *Journal of Statistical Software*, *45*(3), 1–67. https://doi.org/10.18637/jss.v045.i03

Wang, Y. A., & Todd, A. R. (2021). Evaluations of empathizers depend on the target of empathy. *Journal of Personality and Social Psychology*, *121*(5), 1005–1028. https://doi.org/10.1037/pspi0000341

Wenzel, M., & Okimoto, T. G. (2010). How acts of forgiveness restore a sense of justice: Addressing status/power and value concerns raised by transgressions. *European Journal of Social Psychology*, *40*(3), 401–417. https://doi.org/10.1002/ejsp.629

Wenzel, M., Okimoto, T. G., Feather, N. T., & Platow, M. J. (2008). Retributive and restorative justice. *Law and Human Behavior*, *32*(5), 375–389. https://doi.org/10.1007/s10979-007-9116-6

Wenzel, M., Okimoto, T. G., Feather, N. T., & Platow, M. J. (2010). Justice through consensus: Shared identity and the preference for a restorative notion of justice. *European Journal of Social Psychology*, *40*(6), 909–930. https://doi.org/10.1002/ejsp.657

Wikimedia Support & Safety Team. (2015). *Harassment Survey 2015*. https://upload.wikimedia.org/wikipedia/commons/5/52/Harassment_Survey_2015_-_Results_Report.pdf

Wright, L., Ruths, D., Dillon, K. P., Saleem, H. M., & Benesch, S. (2017). Vectors for counterspeech on twitter. In *Proceedings of the first workshop on abusive language online* (pp. 57–62). https://doi.org/10.18653/v1/w17-3009

Wulczyn, E., Thain, N., & Dixon, L. (2017). Ex machina: Personal attacks seen at scale. In *Proceedings of the 26th international conference on world wide web* (pp. 1391–1399). https://doi.org/10.1145/3038912.3052591

Xia, Y., Zhu, H., Lu, T., Zhang, P., & Gu, N. (2020). Exploring antecedents and consequences of toxicity in online discussions: A case study on Reddit. *Proceedings of the ACM on Human-Computer Interaction*, *4*(CSCW2), 1–23. https://doi.org/10.1145/3415179

Xiao, S., Cheshire, C., & Salehi, N. (2022). Sensemaking, support, safety, retribution, transformation: A restorative justice approach to understanding adolescents' needs for addressing online harm. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems* (pp. 1–15). https://doi.org/10.1145/3491102.3517614

Young Reusser, A. I., Veit, K. M., Gassin, E. A., Case, J. P., & Reusser, G. M. (2021). Assessing the prevalence of benevolence in response to online toxicity on Reddit. *Technology, Mind & Society 2021 Conference Proceedings*. https://doi.org/10.1037/tms0000023

Zerback, T., & Fawzi, N. (2017). Can online exemplars trigger a spiral of silence? Examining the effects of exemplar opinions on perceptions of public opinion and speaking out. *New Media & Society*, *19*(7), 1034–1051. https://doi.org/10.1177/1461444815625942

Ziegele, M., Breiner, T., & Quiring, O. (2014). What creates interactivity in online news discussions? An exploratory analysis of discussion factors in user comments on news items. *Journal of Communication*, *64*(6), 1111–1138. https://doi.org/10.1111/jcom.12123

Appendices

Appendix A. Reddit conversations used in Pilot Study and Main Experiment

Condition	Pair #	Comment	Reply	Word Count	Correcting rating (1 = not at all correcting; 5 = very much) of reply	Going Along rating (1 = not at all; 5 = very much) of reply
Benevolent Correction	1	The only relation to the thread is you and your stupidity.	I understand you don't like my opinions. But really, to keep questioning my character, etc. We just don't agree. That's all.	32	4	0
	2	That's how you assert dominance. If you're an asshole.	lol, I was joking. :P	14	3.67	3.00
	2*	Stop being dumb, it doesn't matter where you use it.	Hey, insulting people isn't necessary or valuble to the conversation. If you don't have anything nice to say, don't say it.	31	4.67	0
	3	Your edits are dumb.	I found the edits kinda funny actually..	11	3.67	1.33
	3*	Hi, loser.	That's not a very nice thing to say.	10	4.33	0
	4	That's why I want to play a cracked version, and not pay $60 for it, idiot.	That's fine. No need to resort to childish name calling…relax. And if you pay full price for games these days that's just silly. You can buy this game for half that.	47	4	0.67
	4*	The guy who commented above you is an idiot.	Not necessarily, he just missed a line. It happens	18	4.33	0.67
Benevolent Going Along	5	I literally cannot believe you are this stupid.	Im not one for name calling but for once I'm glad someone said it.	22	0.33	4.67
	6	I have compassion for you and all other brain damage sufferers	Haha this interaction is hilarious	16	0	4
	7	HOW DO YOU DROP THAT YOU IDIOT?!	Looking down field before catching it. Gotta grab that, agree!	17		Researcher-selected
	7*	What a fucking idiot.	Agreed. Not sure how he lost control, doesn't even look like he's going that fast.	19	0	4.67
	8	Yeah, I hate our fanbase, so knee jerk. You all suck and need other hobbies.	You said it my brother.	20	0	5
Retaliatory	9	Do you realise how silly you sound?	Calls people silly but provides zero evidence for anything he says… Nice.	19	Researcher-selected	
	9*	.Aren't you quite the smug prick?	"Well hello there pot." "Fuck off kettle, you black bastard."	16	Researcher-selected	
	10	Because you're a dick	So, your ignorance is caused by me?	11	Researcher-selected	
	11	You dumb bastard. It's not a schooner… it's a Sailboat	A schooner IS a sailboat, idiot!	16	Researcher-selected	
	11*	No, you're just stupid. And a pussy.	Its really hard to take such an angry little fella like you seriously. Keep going though.	23	Researcher-selected	
	12	So vicious. You slay me with your pathetic insults.	You "slay" yourself with your unmitigated ignorance.	16	Researcher-selected	

Note. Pair numbers with asterisks were used in the main experiment as replacements for those in the pilot

Appendix B. Measures developed for Pilot and Main Experiment

Free to contribute from 0 (Not at all) to 6 (Very likely/Very willing)
1. How likely would you be to express your *true* opinion to this group?
2. How likely would you be to contribute to this conversation?
3. (Pilot) If you were to post to this forum, to what extent would you feel the need to hide what you really think from the rest of the group?
3. (Main experiment) How willing would you be to join this conversation?
Toxicity dissuaded from -3 (Strongly disagree) to +3 (Strongly agree)
1. (Pilot) The response fixes any concerns I have about the first comment. 1. (Main experiment) The response will encourage the first commenter to post more positively in the future.
2. (Pilot) The response is an appropriate way to address the toxicity of the first comment.
2. (Main experiment) The response will encourage the first commenter to post in a more negative tone than before.* [this item did not correlate strongly with the other scale items and was removed] 3. The response will discourage the first commenter from continuing to post in the same negative tone as before.
4. (Pilot) The response will make the first commenter reconsider what they initially posted. 4. (Main experiment) The response will make the first commenter believe their initial post was inappropriate.
Manipulation check questions
First comment toxicity: (Pilot) What is your first impression of the person who made the first comment above? (-3 (Very Negative) to 0 (Neutral) to +3 (Very Positive). (Main experiment) Please rate the first comment you read above on the following scale: -2 (Very toxic - a very hateful, aggressive or disrespectful comment that is very likely to make you leave a discussion), -1 (Toxic - a rude, disrespectful or unreasonable comment that is somewhat likely to make you leave a discussion), 0 (Neither), +1 (Healthy contribution - a reasonable, civil or polite contribution that is somewhat likely to make you want to continue a discussion) or +2: (Very healthy contribution - A very polite, thoughtful or helpful contribution that is very likely to make you want to continue a discussion) **Reply benevolence, correction, retaliation:** (Main experiment only) Consider the replies to the toxic comments you read. Overall, to what extent did these replies… (participants responded from 0 (not at all) to 6 (extremely)
1. …demonstrate benevolence (politeness, understanding, and empathy) for the initial commenter?
2. …appear to correct the initial comment? 3. …appear to retaliate against the initial commenter?

*Reverse-scored item

Supplementary Materials

Supplemental Material

Download: https://collabra.scholasticahq.com/article/92328-responding-to-online-toxicity-which-strategies-make-others-feel-freer-to-contribute-believe-that-toxicity-will-decrease-and-believe-that-justice-ha/attachment/192636.docx?auth_token=LAtbvQsvqBSAe04xIXNS

Peer Review History

Download: https://collabra.scholasticahq.com/article/92328-responding-to-online-toxicity-which-strategies-make-others-feel-freer-to-contribute-believe-that-toxicity-will-decrease-and-believe-that-justice-ha/attachment/192637.docx?auth_token=LAtbvQsvqBSAe04xIXNS