

Emotion Regulation Contagion Drives Reduction in Negative Intergroup Emotions

Corresponding Author: Dr Amit Goldenberg

This file contains all reviewer reports in order by version, followed by all author rebuttals in order by version.

Version 0:

Reviewer comments:

Reviewer #1

(Remarks to the Author)

Thank you for the opportunity to review this manuscript. The presented work reports findings from a large-scale intervention study that examines the impact of an emotion regulation intervention in groups, comparing different proportions of group members who received the intervention.

This is a well-designed study, and I enjoyed reading this manuscript. The work has several strengths, including its rigor, pre-registered design, and consideration of different proportions and dosage effects. I believe it builds a solid foundation for future intervention studies with important real-world applications. My comments are restricted to some minor suggestions and things that should be clarified in the manuscript.

Introduction/Conceptual Background:

1. The introduction would benefit from more conceptual clarity on relevant group processes of emotional contagion and amplification. Similarly, what are the proposed/assumed mechanisms given that the experiment did not seem to include face-to-face interaction? Is face-to-face interaction deemed important for these group processes? Additionally, was emotional contagion thought to be achieved mainly through the increased mimicking of the trained participants' emotion regulation strategies (e.g., reappraisal)? More elaboration in the manuscript would be helpful.

Experiment/Method:

2. Instructions/interventions: It seems like both groups knew they would receive an emotion intervention of some kind (either reappraising emotions or "observing" how they naturally unfolded). How do the authors think this could have affected the results? In general, I also wondered whether the control condition may have triggered emotional acceptance to some extent?
3. Was there an experimenter guiding participants through the task? How did the experiment ensure that participants actually read each other's responses?
4. Impact on general sentiments towards Palestinians: Was general sentiment assessed at a single time point or multiple times (pre/post task)? If only 1 time point was used, what was the rationale? Please clarify in the manuscript and make sure to use appropriate language (I assumed more than 1 timepoint since the word "change" is used but, based on the Methods section, I am not sure). And were the groups comparable in terms of pre-existing sentiments toward Palestinians?
5. Maybe I missed it, but more information on the dropout rate would be helpful.

Discussion:

6. It seems participants did not interact with each other directly (e.g., no camera on or oral interaction). I would suggest to more explicitly state this lack of face-to-face interaction in the discussion and discuss its implications for potential transfer effects in real-world contexts. As a potential strength, I wonder whether this intervention could be particularly relevant for social media contexts, as a context in which emotional contagion is common without face-to-face interaction
7. It would be helpful to briefly outline how follow-up studies could more clearly distinguish emotional contagion from conformity processes.

Minor/specific comments:

- P.8: "These measures were designed to test whether there were changes in emotions could be seen when participants know that their ratings will be shown to others, in order to eliminate the possibility that compliance was driving participants' emotions." -> eliminate is quite a strong word here, consider rephrasing.
- Figure 2: Perhaps use the terms "treated/non-treated participants" in the figure legend, for consistency/to make it more self-

explanatory.

- Method section, sample description: I would suggest to add % for genders.
- Discussion: "...but also other, more general psychological interventions": Perhaps add a very brief example
- In the abstract: "Results indicated an exponential relationship between the proportion of treated participants and the group reduction in emotions" -> consider briefly elaborating this relationship in plain language
- In Supplemental Material (p.8; control instructions): "In the current study, we will teach you a method that can help you change your emotional response, especially when it comes to reducing negative emotions and in a group context." -> was this really part of the control instructions or could this be a copy-paste error?
- General: There are a few instances where the language sounds vague and should be edited for clarity. For example:
 - o "Relatively strict experimental design"
 - o "...this model was very similar to others" (similar in terms of model fit?)
 - o "We had a similar model with an interaction between proportion and condition which is reported in SI."

Reviewer #2

(Remarks to the Author)

This paper describes an impressively large experimental study examining the role of emotion regulation in group settings. The authors test the idea that individual level interventions of reappraisal can alter (untreated) group member's reappraisal in a "contagion" process. There were many things to like about this experiment including the large sample, the intriguing question, sophisticated design and analyses, and transparent and comprehensive preregistration.

There were a few issues that arose while reading the paper that I have enumerated below.

1. To what extent are the authors observing conformity/compliance versus emotion regulation contagion? While reading the paper the open question that nagged at me was to what extent were participants responding differently based on compliance/conforming to others in the group versus the contagion process the authors claim. The authors raise the possibility compliance as the first limitation on page 21, but it was also clear in the results section that they worried about this as well. For example, when attempting to address this issue in the "general sentiments towards Palestinians" section they write "finding differences in the expression of sentiments towards Palestinians would be another indication that **genuine contagion** occurred" (emphasis mine). This seems to be the critical question here, was there emotion regulation contagion in that people felt differently when others in their group received an emotion regulation manipulation or were the untreated participants altering their responses to be more in line with those who were reappraising and offering (likely) more complex, multidimensional understanding of the conflict? The authors point to the sentiment data to show that when responses were private there was an effect of the proportion of reappraisers showing that negative sentiments decreased with more reappraisers in a group. Notably though, the effects observed were much smaller than when examining the public responses, there was not an overall main effect when comparing treated to untreated individuals, and there was an interaction between proportion and condition (peeked out a $t=1.97$) with control participants showing less negative sentiments with greater proportion of reappraisers in their group (page 16). The supplemental material is maybe more concerning with null effects for changes in attitudes towards Palestinians (SI: page 24) or concessions towards Palestinians (SI: page 25), showing no changes in private attitudes immediately following the group interaction.

These weaker and null effects don't completely undermine the effects here—the authors are examining emotion regulation process in a group, which they do show robust effects. However, these findings do bring up questions regarding the authenticity of the public pronouncements and the extent to which untreated individuals are genuinely influenced by the number of reappraisers versus parroting the majority or persuasive sentiment. You can see some influence of "majority opinions" in the authors' data in figure 4C. When looking at the figure on the far left – one reappraiser and five untreated group members, the reappraiser became more similar to the untreated group members, consistent with a conformity effect. Do the authors have any data that can address the question that these effects reflect genuine changes in emotion regulation among the untreated participants?

2. Word cloud. The authors use text analyses and then present a word cloud to provide a qualitative visual of the differences as a function of the experimental manipulation (figure 4a). I did not see the value of the word cloud for the paper. For a conference talk, this is a cute way to present an easy visual but for a scientific paper, especially one that can provide a quantitative answer to the question of whether the two conditions resulted in different word choices, this seemed beneath the rigor of the work. The word cloud is presented with only the two extreme conditions (no reappraiser versus six reappraisers) so it provides limited value. I would vote to remove this from the main paper.

Minor clarifications:

1. On page 7, the authors describe the experimental design as showing participants stimuli depicting "either Palestinian resistance to the Israeli occupation or Palestinian violence against Israel." Because of the language used here – "either/or" – I assumed this was a between subjects factor. That is, the authors were studying emotion regulation contagion as a function of stimuli where Palestinians were either attackers or defenders. It turns out that this "condition" was not mentioned again, and it isn't mentioned in the preregistration either (unless I missed it). I assume I was wrong, and this wasn't a condition that was manipulated in the experiment. The authors might want to clarify what was meant by this sentence given other readers might also mistakenly believe there was another manipulated condition. If this was just an effort to use a variety of stimuli to show that there was not something unique about one type of stimuli (Palestinians only presented offensively) it might be worth explaining that.

Reviewer #3

(Remarks to the Author)

This is an extremely interesting paper with immediate and far-reaching public policy implications. I recommend that it be accepted for publication. I have a number of comments for how to improve the clarity and impact of the findings prior to publication. While I suspect I will not be alone in some of my confusions/questions, I do not think that addressing any of these is required for publishing this paper. I think it is already sufficiently clear and will be received with great interest by the broader research community. My numbered comments are below:

1. The authors are understandably cautious in the extent to which they frame their results as supporting particular theories of contagion. That said, their experiment is well setup to test and support complex contagion theory, which holds that peer reinforcement is required to trigger spreading of a counter-normative emotion. Their finding that treating more participants generates more spillover is consistent with this of course. But more specifically, Panel C of Figure 4 suggests that when the number of treated agents is 1/6 (a simple contagion), there are many trials where similarity to reappraisal is markedly lower than all cases where a complex contagion is satisfied (treated $\geq 2/6$). This suggests negative social influence, which is consistent with the dynamics of fractional thresholds in complex contagions (e.g., see Centola & Macy, 2007; Guilbeault & Centola 2021). Negative influence suggests that in the absence of sufficient peer reinforcement, the treated subjects may actually revert to the negative (or at least neutral, the ceasing of the positive) emotion. If the authors could show support for this, this would make for a very cool empirical contribution to complex contagion theory (perhaps better left for future work). One possible way to measure this would be to look at whether individuals' similarity to reappraisal decreases overtime in the conditions with 1/6 treated. Furthermore, the finding that the greatest movement toward reappraisal happens between 60-80% regulators suggests we are dealing with quite a complex contagion that requires considerable reinforcement (both panel B and C of fig. 4). These patterns are highly consistent with complex contagion theory. I'd encourage the authors to highlight some of these connections in the discussion, since I think they would increase the impact of their findings for the contagion literature.

References:

-Centola, Damon, and Michael Macy. "Complex Contagions and the Weakness of Long Ties." *American Journal of Sociology* 113, no. 3 (2007): 702–34. <https://doi.org/10.1086/521848>.
-Guilbeault, Douglas, and Damon Centola. "Topological Measures for Identifying and Predicting the Spread of Complex Contagions." *Nature Communications* 12, no. 1 (July 20, 2021): 4430. <https://doi.org/10.1038/s41467-021-24704-6>.

2. Throughout, the authors use the term "control participant" to refer to both untreated participants in networks exposed to treated participants and participants in an "only-control" condition (no reappraisal in the group). I find this a bit confusing. It seems more appropriate to describe the only-control groups as the control condition. Non-treated agents in the groups with the treated participants are in the treatment condition, they just happen to not have received the treatment directly but rather as a spillover. I found this particularly confusing when the authors talked about interactions in their models where the movement toward reappraisal was strongest among control participants; I found myself asking, but weren't the control participants not exposed to treated participants? Then I had to go through the manuscript again to clarify. I'd encourage the authors to refine the terminology to avoid similar confusions among other readers.

3. This might be out of scope and better left for future work, but it would be super cool to see if the way that treated individuals communicate (or say, the extent of reappraisal or positivity they communicate) is mediated by the presence of communications from other people who have been treated. The presence of others may create a safe space for communicating counter-normative affect. I am reminded of Traeger et al.'s (2020) paper showing that injecting bots that communicate vulnerable messages can help steer a group discussion in a civil direction by making it psychologically safe for people to communicate similarly vulnerable sentiments.

Reference

-Traeger, Margaret L., Sarah Strohkorb Sebo, Malte Jung, Brian Scassellati, and Nicholas A. Christakis. "Vulnerable Robots Positively Shape Human Conversational Dynamics in a Human–Robot Team." *Proceedings of the National Academy of Sciences* 117, no. 12 (March 24, 2020): 6370–75. <https://doi.org/10.1073/pnas.1910402117>.

4. I think the word cloud analysis could be tighter. Why not show the top n most frequent words in each group and color them by negative and positive, and show there is more positive terms more highly ranked in the positive condition? This could also be evaluated for statistical significance (difference in average ranking of positive items) in a frequentist sense. But regardless, I buy the results here.

5. I found it a bit confusing that the authors fit different functional forms to different aspects of the diffusion process, without elaborating on the qualitative implications. At a group level, they emphasize the exponential, which makes sense. But I don't feel like the quadratic is sufficiently explained. The quadratic of course typically evokes a u-shaped curve. But the authors describe this functional form as a kind of step function writing that this model "suggested that there is relatively little change in the emotions of the non-treated participants until a certain proportion, at which point changes become much greater with every increase..." Qualitatively, this makes good sense to me, but it's not clear that only the quadratic functional form would capture this, so a little more elaboration on this point (even just discursively) would be helpful. This would help clarify in what way the authors show a "non-linear reduction in conflict-related negative emotions" by clarifying the kind of non-linearity observed (is it flat until a tipping point and then curve positive, like a hockey stick, or something more u-shaped).

6. One question concerning the movement toward the pure reappraisal vector – how much of this is driven by a shift in affect vs. the adoption of reappraisal-specific content language (e.g., increased use of the term “interpretation” or “reappraisal” itself), which could theoretically happen without a shift in affect. I think it would be sufficient to show in the appendix a handful of representative examples of people’s actual messages – from the treated participants and/or the untreated by infected participants – whose messages scored as being particularly close to the reappraisal vector. This would add some nice color.

7. I’d also mention in the discussion that an important direction for future research is to examine the effects of anonymity in this setting. If people have access to the identities of other people – or, perhaps, if they are knowingly interacting with people they have an actual history with, like friends or family – this could significantly impact their willingness to express either strong positive or negative views (depending on what they expect to be the normative stance of their network). For instance, we’ve shown that exposure to identifying information can meaningfully change people’s willingness to use social information (Guilbeault et al. 2018), consistent with broader empirical studies of social media (Aral et al. 2023). I don’t think this takes away from the current results, which stand alone, but I think this marks an important future direction worth noting.

References

- Guilbeault, Douglas, Joshua Becker, and Damon Centola. “Social Learning and Partisan Bias in the Interpretation of Climate Trends.” *Proceedings of the National Academy of Sciences* 115, no. 39 (September 25, 2018): 9714–19. <https://doi.org/10.1073/pnas.1722664115>.
- Taylor, Sean J., Lev Muchnik, Madhav Kumar, and Sinan Aral. “Identity Effects in Social Media.” *Nature Human Behaviour* 7, no. 1 (January 2023): 27–37. <https://doi.org/10.1038/s41562-022-01459-8>.

Best of luck with this revision. Thank you for the pleasure of reviewing this excellent work.

-Doug Guilbeault

Reviewer #4

(Remarks to the Author)

The topic of this paper, to better understand the social dynamics of the regulation of negative emotions in groups, is compelling. The results that increasing the proportion of treated participants produces more emotional regulation among non-treated group members is important and novel, especially given that the authors could not predict even the direction of the effects.

I have a number of suggestions for improving the paper.

I was left with a number of questions regarding what the authors thought was spreading here, and the nature of the relationship between emotions and emotional regulation language: are the authors making a claim that it is the emotional regulation that is spreading initially, and that impacts how participants report their negative emotions (which implies a mediation analysis using the semantic analysis results) or is what is spreading the emotions (emotional ratings) themselves? I think this distinction matters because we know from other contagion literature that what is being spread matters for how it spreads (e.g., Centola’s work on complex contagions). Relatedly, I’m unsure whether the authors would argue that what is happening in the intervention is the same or fundamentally different than the spread (or convergence?) of emotions in a group with different initial proportions of negative emotions (“naturally occurring” vs. the product of the intervention).

The authors need to further develop, and consider possible tests for, the potential mechanisms by which emotions/emotional regulation spread in a group. I would have loved to see much more of the authors’ thinking (which I believe appears in some of their other work) on whether and how the spread of emotions (or emotional regulation) differs from the spread of social behaviors (simple or complex), attitudes, perceptions in groups, etc. Are the models we already have for these types of diffusion processes adequate in this case or not? Should we think about this as an influence process (participants are learning what is acceptable text and emotional ratings in a group) or actually a “contagion” where seeing other people with negative emotions gives a participant more negative emotions? (I could see this argument holding for negative emotions, but the opposite, seeing more neutral emotions and “catching” those seems far less plausible, suggesting asymmetrical processes by proportion of treated participants in a group.) Also, should we characterize emotions/the use of emotional regulation as “spreading” as in diffusion, or something more like converging? I think the convergence language makes more sense here, but it is hard to tell from these analyses and without the group-level results, which I wish the authors had included. (I’d also like to know the N of the number of groups, which doesn’t appear in the manuscript or SI.)

I appreciated the testing of multiple functional forms of the relationship between the proportion of treated participants and the results are interesting, but they aren’t a substitute for the elaboration of nature of spread/convergence in this case (especially given how minor the AIC differences were between them); e.g., should we think about spread/convergence as a threshold effect, and what inference processes in the group might give rise to that? As the authors know, there are many types of spread/contagion/spillover processes and I would have liked a more precise definition of which they believed they were observing and why.

Another significant concern I had was whether there was differential attrition from the experiments as a function of condition (proportion of treated individuals in a group) or as a function of having more negative emotions (on average). In the SI, the authors report starting with 5040 groups participants, and, after exclusions and attrition, ended with 2,830 people, indicating

massive amounts of attrition. I would like to see the rate of attrition by condition (and by average emotion ratings) to ensure that the results are not significantly impacted by this attrition (e.g., are those with the most negative emotions at the outset more likely to drop out? Are those in low proportion conditions more likely to drop out?). (Also unclear was whether, in the analyses, the number of group members varied across trials. Whether they did or did not, the authors should discuss whether we should expect that the process of influence/spread/convergence differs at 50% treated group members when the group size is 4 vs. 6, for example.)

More minor comments:

The front end of the paper describing the procedures should be rewritten for clarity, e.g. the description of the control condition on p. 6, that the group size was not only 6 and what the distribution of group sizes were (related to above comment), etc. There were a number of key details about the procedures (the sequential randomization of group to condition and participant to condition; the existence of 20 experimental trials, etc.) that only appear in the Methods section that would have helped me understand the analytical choices better in the first reading.

The authors claimed that treating “25-40% of the group with a reappraisal intervention results in significant and reliable emotion contagion at the group level” (p. 14). Based on Figure 2B, it looks like the effect only become “reliable” (which I believe to mean consistently significant and in the same direction) at 40%. Please clarify.

The authors make a claim about this study testing the proportion of the network treated in a fully connected network (p. 5). Given that the authors do not use any information about the nature of the relationship among participants, the comparison to a network study is not appropriate here.

Version 1:

Reviewer comments:

Reviewer #1

(Remarks to the Author)

Thank you for your thorough revisions! This is an excellent manuscript, and I think it's ready for publication.

Reviewer #2

(Remarks to the Author)

I reviewed the previous submission of this paper. I appreciated how thoroughly the authors addressed various issues raised by me and the other reviewers and editor. I especially appreciated the additional study added to attempt to show that compliance wasn't the only driving factor. Additionally, the authors offer a fair summary that while compliance could still be driving some changes it is not the only factor at play.

I think the authors have delivered a fascinating and important manuscript that has broad appeal

Reviewer #3

(Remarks to the Author)

The authors have addressed my comments very thoughtfully and effectively. I have no remaining concerns. I think the revised paper is ready for publication, which I firmly recommend. Best of luck!

Reviewer #4

(Remarks to the Author)

The authors' revisions to the manuscript were very thorough and persuasive. I appreciate the new data collection now provided in the SI to address concerns about possible compliance effects. One minor note is that the text (e.g., the abstract and the discussion) implies more certainty about the functional form of the effects as exponential than the authors acknowledge in the rebuttal to reviewers. It may be warranted to clarify the language around the claim.

Open Access This Peer Review File is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made.

In cases where reviewers are anonymous, credit should be given to 'Anonymous Referee' and the source.

The images or other third party material in this Peer Review File are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.

To view a copy of this license, visit <https://creativecommons.org/licenses/by/4.0/>

Reviewer 1

R1.1 Thank you for the opportunity to review this manuscript. The presented work reports findings from a large-scale intervention study that examines the impact of an emotion regulation intervention in groups, comparing different proportions of group members who received the intervention.

This is a well-designed study, and I enjoyed reading this manuscript. The work has several strengths, including its rigor, pre-registered design, and consideration of different proportions and dosage effects. I believe it builds a solid foundation for future intervention studies with important real-world applications. My comments are restricted to some minor suggestions and things that should be clarified in the manuscript.

Thank you for your careful and kind review.

R1.2 Introduction/Conceptual Background: The introduction would benefit from more conceptual clarity on relevant group processes of emotional contagion and amplification. Similarly, what are the proposed/assumed mechanisms given that the experiment did not seem to include face-to-face interaction? Is face-to-face interaction deemed important for these group processes? Additionally, was emotional contagion thought to be achieved mainly through the increased mimicking of the trained participants' emotion regulation strategies (e.g., reappraisal)? More elaboration in the manuscript would be helpful.

Thank you for this comment. Traditionally, emotion contagion is achieved through two primary mechanisms. The first is simple mimicry, in which an emotional expression activates synchronous behavior on the part of the perceiver, which in turn activates affective processes (Hatfield et al., 1994). The notion of mimicry has been conceptually extended to the idea of category activation, in which exposure to emotional expressions primes an emotion category, which in turn leads to activation of specific emotional processes (Niedenthal et al., 2009; Peters & Kashima, 2015). The second mechanism, and the one that is more relevant to the current paper, is social appraisal, in which individuals use others' emotions as a guide for their own emotion appraisals, leading to similar emotional experiences (Clément & Dukes, 2017; Manstead & Fischer, 2001). Thus, it's likely that both emotion contagion (as seen in our pilot) and emotion regulation contagion (seen in the main study) are mainly driven by social appraisals. We do not think that mimicry is what drove contagion in this case for a few reasons. First, as you seem to have indicated, mimicry is likely to occur more in cases where people are exposed to others' emotions in more modes than just text. Second, people saw others' ratings to specific stimuli after they provided their own ratings, which suggests that mimicry, which often is limited to real-time contagion, is less likely to occur.

However, we believe that social appraisal can't be the only process driving our findings, as participants apply reappraisal to novel stimuli without seeing others' appraisals of those stimuli. As a reminder, participants in our studies learn about other people's ratings and

text responses in trial 1, but then seem to implement them in trial 2, before seeing others' responses. Therefore, in addition to social appraisal, which impacts participants' general understanding of the context, participants also seem to actually learn to apply the strategy of reappraisal. This is also supported by our language analysis, which suggests that reappraisal itself is more likely to be used when others are using reappraisal. Therefore, in addition to social appraisal, it seems that our participants also learn to conduct reappraisal through social learning, a phenomenon which has mainly been examined in developmental psychology (Morris et al., 2007; Silvers, 2022; Sorce et al., 1985).

We appreciate you pushing us toward greater clarity on this topic. In the introduction, we added a clarification of the potential mechanisms (p.4-5):

In the current project, we examined the possibility of leveraging the process of emotion regulation contagion to change group emotions. We define emotion regulation contagion as the spread of emotion regulation strategies from people who have received emotion regulation interventions to those who are not treated. We see emotion regulation contagion of cognitive reappraisal, as tested in the current project, as driven by two processes. First, regulation contagion may occur as a result of social appraisals, which are defined as the adoption of other people's appraisal of a certain situation in a way that impacts one's emotional response. In the current case, non-treated participants would adopt the reappraisals of the treated participants to conflict situations. Second, we believe that regulation contagion is driven by social learning, such that non-treated participants would learn how to generate reappraisals to similar stimuli and apply this knowledge to new stimuli.

We wish to distinguish the process of emotion regulation contagion from two similar processes (See Figure 1). The first is emotion contagion, where the mere expression of emotions of the treated participants is impacting the emotions expressed by the non-treated participants. Although emotion contagion is likely playing a role in impacting non-treated participants' emotions, we examined whether non-treated participants are also adopting the use of reappraisal, in addition to adjusting their emotion expressions to match those of other groups members. The second related process is compliance, where no real change in emotion experiences occurs, but merely a change in expressed emotions as a result of social pressure. Here again, although compliance may impact the results, we will show some evidence that non-treated participants are changing their emotions privately.

We created a new figure (Figure 1) for a clear summary of these mechanisms and their relationships.

We also now discuss how our empirical evidence based on linguistic analysis can shed light on these mechanisms on page 23:

Analyzing participants' text using Semantic Projection Analysis provided evidence for change in language produced by the non-treated participants as a function of the proportion of treated participants, providing support for linguistic contagion between the

treated and non-treated participants. The linguistic analysis results provide evidence for the social learning mechanism, such that the non-treated participants seemed to learn the cognitive reappraisal strategy from other treated individuals and use it in their newly generated texts.

It is also important for us to clarify, in light of your comment on face-to-face interactions, that both social appraisal and social learning do not require face-to-face interactions. This is somewhat in contrast to classic mimicry processes which do require exposure to others' expression of emotion in a more direct way. We agree, however, that the medium of social interactions can shape the emotion contagion process and that text-based interactions limit the impact of some non-verbal information such as facial expressions and body postures. Nonetheless, we argue that the proposed mechanisms are still valid under the current context. As research on emotions on digital platforms shows, textual information can serve as emotional cues that influence others' affect in the absence of non-verbal cues (Cheshin et al., 2011; Guillory et al., 2011; Kramer et al., 2014). Combining with related comments from you and Reviewer 3 (See R1.7 and R3.4), we added this point to the discussion of the paper (p.26):

On a related note, social interactions in the current study were solely based on text communication. We believe this communication form can make our findings particularly relevant for digital contexts such as social media, but we also recognize that the absence of non-verbal information such as facial expressions and body postures may shape the contagion process of emotions and emotion regulation. Besides, in more natural conversations, reappraisal training might not only change how group members react to stimuli but also how people respond to each other's comments and the quality of discourse. Future studies should examine regulation contagion in more naturalistic environments such as face-to-face interactions in work teams.

R1.3 Experiment/Method, Instructions/interventions: It seems like both groups knew they would receive an emotion intervention of some kind (either reappraising emotions or “observing” how they naturally unfolded). How do the authors think this could have affected the results? In general, I also wondered whether the control condition may have triggered emotional acceptance to some extent?

We chose the active control of observing because we considered it to be a more conservative choice for a control condition. If we only had an empty control, we would not be able to know if reappraisal worked merely because participants were asked to pay more attention to their emotions, which is a byproduct of any emotion regulation intervention including reappraisal, or because they actually learned how to use reappraisal. Another reason for choosing the observing control was due to the fact that a similar control condition has been used in a previous large-scale study ($N = 21,644$) and was also compared in that study to an empty control condition in which participants were just told to observe and react naturally (Wang et al., 2020). Results did not indicate any difference in emotional intensity between the active observing control (which was similar to what we did) and the passive control. Adding the active control condition in this study also served a technical goal. Because we had to coordinate 6 participants who completed the task

synchronously, we wanted the instruction part to be more or less of the same length so that it was easier to make the active control similar in length to the reappraisal training.

In terms of speculating regarding the effect of the control manipulation, based on existing literature, having participants reflect on their own feelings has been shown to generally reduce negative emotions, which was also the original goal of using this manipulation in the first place (Frattaroli, 2006). Although, as we mentioned, a previous study has not found a significant difference between the active control and a more passive control (Wang et al., 2020), we would expect that if anything our control condition would lead to a reduction in negative emotions compared to a passive control as a result of the effect of emotional acceptance as you pointed out. It is, however, definitely possible, that in the context of an intractable conflict, having people reflect on their emotions may lead to stronger emotions, which could be a limitation of the study. To discuss this limitation and the rationale for our choices, we modified the manuscript in the following way (p.29):

We chose to use this active control condition – in which participants were asked to engage with their emotions – because it has been used in previous studies exploring reappraisal. It is, however, worth mentioning that a recent study that compared observing to a more passive non-treated condition in which participants were merely instructed to respond did not find consistent differences between the two conditions. This can be a limitation particularly in the context of a conflict where having people reflect on their emotions may lead to stronger emotions.

R1.4 Was there an experimenter guiding participants through the task? How did the experiment ensure that participants actually read each other's responses?

There was no experimenter who guided the participants. Although we did not have direct evidence showing that participants read all the responses, we designed the experiment so that participants had to at least pay attention to the progress of the task and generate ratings and text responses in order not to be automatically kicked out of the experiment. The way the experiment was set up and explained to participants was that after five trials of missing ratings or responses, participants would be automatically kicked out.

More specifically, in the experiment, participants had 10 seconds to read other's responses and additional 20 seconds to provide their own responses. Due to the short intervals between producing the ratings and texts, we assume this would force participants to stay engaged during the task, keeping their attention focused on their screen. Given the short intervals, and the risk of getting kicked out of the task for missing trials, participants likely stayed in the task to not miss the next step. Indeed, of the 20 trials that participants were asked to complete, the average amount of responses that participants provided was 18.00. It is therefore unlikely that participants were quick enough to leave the feedback stage and come back to the next text collection.

Regardless of the design of the task, we also believe that the supplementary study that we added to the project provides clear evidence that participants were actually paying attention

to the text. In this new study (described in detail in SI), participants were only exposed to texts produced by other group members while their own ratings remained private. The study results suggest control participants' ratings were still reduced merely by being exposed to the text, suggesting that text had an impact on participants' own emotions. See detailed response in R2.2.

R1.5 Impact on general sentiments towards Palestinians: Was general sentiment assessed at a single time point or multiple times (pre/post task)? If only 1 time point was used, what was the rationale? Please clarify in the manuscript and make sure to use appropriate language (I assumed more than 1 timepoint since the word “change” is used but, based on the Methods section, I am not sure). And were the groups comparable in terms of pre-existing sentiments toward Palestinians?

The general sentiment towards Palestinians was measured only after the task. By using the word “change”, we meant to refer to the difference between the reappraisal and control condition. Following your suggestion, we modified our language to be more rigorous by making it clear that it was about between-condition comparison instead of pre-post differences. We now say (p.15-16):

“To examine the effect of the treatment on the sentiments towards Palestinians following the task we averaged both the proportion of treated participants and the group size across all trial [...] Looking first at negative emotions and exploring the main effects, results suggested that increasing the proportion of participants treated with reappraisal led to lower negative sentiments for both treated and non-treated participants [...] the relationship between proportion of treated participants and negative sentiment was stronger for the control participants.”

We decided not to have a pre-measure of general sentiments towards Palestinians because we were afraid that such a measure would impact the quality of reappraisal training. This decision was based on other research that suggested that participants are less interested in learning about ways to down-regulate their emotions when they know that the context is intergroup conflicts (Porat et al., 2016). We also didn't add this measure after the training and before the actual task, because general sentiments towards Palestinians are more stable than real-time reactions to stimuli and we were concerned that having these general sentiments would anchor the ratings in the task.

Regarding your question about the pre-existing difference in participants' sentiments towards Palestinians, although we did not directly measure sentiments toward Palestinians at the beginning of the task, we did measure general political attitudes participants held before the task on a 7-point Likert scale (1 = extreme right, 7 = extreme left). Results of a Welch two-sample t-test showed no statistically significant difference between treated and non-treated people's political attitudes ($t = 0.08$, $df = 2144.7$, $p = .933$). We also showed that this political attitude measure is a significant predictor of general sentiment towards Palestinians ($\beta = 0.24$, $p < .001$, 95% CI [0.20, 0.29]). Therefore, we believe this indicates that there is unlikely to be a significant preexisting difference in attitudes toward Palestinians between the two conditions.

To further clarify this point, we added this short explanation when reporting the results of the general sentiments (p.16):

Note that we decided not to have a measure of general sentiments towards Palestinians before the task because we were afraid that such a measure could impact the quality of reappraisal training.

R1.6 Maybe I missed it, but more information on the dropout rate would be helpful.

Our detailed explanation of the dropout rate was part of the supplementary information, but we have now added it to the main manuscript. We now provide a comprehensive report of all the participants that we recruited, every participant that we dropped, and the reason for doing so in the method section (p.27-28):

Our aim was to recruit 2,800 participants, 400 in each proportion group. Participants could start the task only after being grouped with five other participants. If after 5 minutes of waiting we were not able to group participants, they were transferred to complete a different task. Therefore, our initial sample was much larger than the one we were able to group. Our initial recruited sample was 6,161 participants. Of those participants, 5040 were able to be grouped in the initial round and started the task, but may have dropped in many different stages. The list of 5040 included participants who started the task twice. We allowed participants who did not start the actual rating phase (1604) to run through the task again. As preregistered, our dropout criteria were (numbers may be overlapping): (1) Participants who failed the reading check in which they were asked to choose the option that best describes the strategy that they were assigned to before starting the actual task (146 participants). (2) Participants who did not provide text responses, provided nonsensical text responses, or provided the exact same text response (10 participants). (3) groups who reported that they did not see the images or who had technical issues (6 groups). Participants were also automatically dropped from the study if they did not complete 15 of the 20 trials. After all of these omissions we had 2,830 who completed the whole task. We removed all participants who were the only person remaining in the group, leaving us with 2,659 participants which is our final sample. (Gender: 1,158 males (43.6%), 1495 females (56.2%), 6 other or refused to say (0.2%); Age: $M = 42.07$, $SD = 14.618$).

R1.7 Discussion: It seems participants did not interact with each other directly (e.g., no camera on or oral interaction). I would suggest to more explicitly state this lack of face-to-face interaction in the discussion and discuss its implications for potential transfer effects in real-world contexts. As a potential strength, I wonder whether this intervention could be particularly relevant for social media contexts, as a context in which emotional contagion is common without face-to-face interaction

We agree that text-based interactions may not be easily translated into other face-to-face interactions. We also appreciate your suggestion that our findings could be particularly useful in the social media context where emotion contagion is widely observed and especially impactful. We added a paragraph that discusses this limitation and potential implications to the discussion section (p.26) (see details in our previous response R1.2):

On a related note, social interactions in the current study were solely based on text communication. We believe this communication form can make our findings particularly relevant for digital contexts such as social media, but we also recognize that the absence of non-verbal information such as facial expressions and body postures may shape the contagion process of emotions and emotion regulation. Besides, in more natural conversations, reappraisal training might not only change how group members react to stimuli but also how people respond to each other's comments and the quality of discourse. Future studies should examine regulation contagion in more naturalistic environments such as face-to-face interactions in work teams.

R1.8 It would be helpful to briefly outline how follow-up studies could more clearly distinguish emotional contagion from conformity processes.

We agree that a more detailed discussion of contagion versus other processes is warranted and we added it to the introduction and discussion. In the introduction, we added (p.4-5):

We wish to distinguish the process of emotion regulation contagion from two similar processes (See Figure 1). The first is emotion contagion, where the mere expression of emotions of the treated participants is impacting the emotions expressed by the non-treated participants. Although emotion contagion is likely playing a role in impacting non-treated participants' emotions, we examined whether non-treated participants are also adopting the use of reappraisal, in addition to adjusting their emotion expressions to match those of other groups members. The second related process is compliance, where no real change in emotion experiences occurs, but merely a change in expressed emotions as a result of social pressure. Here again, although compliance may impact the results, we will show some evidence that non-treated participants are changing their emotions privately.

In the discussion, we added (p.25):

A second limitation is the concern that changes in ratings throughout the task may have been driven by compliance with other group members, rather than real changes in emotion. We directly addressed this concern with a supplementary study where participants gave ratings privately showing changes in their ratings despite not being seen by other group members. Note that we do not think this study rules out the possibility that compliance may be playing some role in shaping participants' ratings. Instead, we believe that it provides strong support for the idea that compliance can't be the only factor driving performance. Additionally, participants' general sentiments towards Palestinians after the task, when participants knew that their ratings would not be observed by other participants, were also reduced with the increased proportion of treated participants, providing another evidence that compliance was not the only factor driving the results. Nevertheless, future studies

should use other non-self-report methods to examine changes in emotion as a result of regulation contagion (see for example Oveis et al. (2020)).

Based on your recommendation and suggestions of reviewer 2 (see response R2.2) we also added an additional study that provides more clarity on the process of contagion, and specifically suggests that the effects are beyond mere compliance. We appreciate this feedback and hope that you will find the revised version clearer on these issues.

R1.9 Minor/specific comments:

- **P.8: “These measures were designed to test whether there were changes in emotions could be seen when participants know that their ratings will be shown to others, in order to eliminate the possibility that compliance was driving participants’ emotions.” -> eliminate is quite a strong word here, consider rephrasing.**

We changed the word “eliminate” to “reduce”.

- **Figure 2: Perhaps use the terms “treated/non-treated participants” in the figure legend, for consistency/to make it more self-explanatory.**

We changed the legend in all figures to “treated/non-treated participants” to make the terminology consistent in the paper.

- **Method section, sample description: I would suggest to add % for genders.**

We added the percentage of each gender on p.28: “*Gender: 1,158 males (43.6%), 1495 females (56.2%), 6 other or refused to say (0.2%)*”.

- **Discussion: “...but also other, more general psychological interventions”: Perhaps add a very brief example**

We believe the contagion effect we suggested can be applied to a variety of psychological interventions such as those for group belonging (Walton & Cohen, 2011), health behaviors (D. A. Kim et al., 2015), and individual mindsets (Yeager et al., 2019). We now added (p.23):

A variety of interventions, such as growth mindset, social belonging, and health behaviors, could benefit from thinking about their impact at the collective rather than individual level.

- **In the abstract: “Results indicated an exponential relationship between the proportion of treated participants and the group reduction in emotions” -> consider briefly elaborating this relationship in plain language**

We now say: *“The results showed that as the proportion of participants who received the treatment increased, the reduction in emotions within the group grew at an increasing rate.”*

• **In Supplemental Material (p.8; control instructions): “In the current study, we will teach you a method that can help you change your emotional response, especially when it comes to reducing negative emotions and in a group context.” -> was this really part of the control instructions or could this be a copy-paste error?**

This is an actual part of the control instruction. We decided to use an active control condition for three reasons (also provided in detail in response R1.3): First, we saw this as a conservative option, as we wanted to eliminate the possibility that merely attending one’s emotions would lead to the desired effect. Second, this condition has been tested as a control condition in a previous large-scale experiment (Wang et al., 2020) and was found to be no different than an empty control. Third, we wanted to equate the time participants spent reading instructions, and having an active control was a good solution to that issue.

• **General: There are a few instances where the language sounds vague and should be edited for clarity. For example:**
o **“Relatively strict experimental design”**

We now clarify this by saying (p.6): *“We used a text-based interaction design to precisely restrict the influence processes to written communication and to standardize the experimental setting across groups.”*

o **“...this model was very similar to others” (similar in terms of model fit?)**

We added (p.13) *“in terms of model fit”*.

o **“We had a similar model with an interaction between proportion and condition which is reported in SI.”**

We further elaborate on this model (p.12):

As preregistered, in another model, we included an interaction term between proportion and condition and tested the simple effects in each condition. In that model, we found a statistically significant main effect of the exponential term of proportion ($\beta = .01$, 95% Confidence Intervals $[-.07, .08]$, $t(3126.45) = -3.19$, $p = .840$). The coefficient of the interaction term was not statistically significant (See SI for details).

Reviewer #2 (Remarks to the Author):

R2.1 This paper describes an impressively large experimental study examining the role of emotion regulation in group settings. The authors test the idea that individual level interventions of reappraisal can alter (untreated) group member's reappraisal in a "contagion" process. There were many things to like about this experiment including the large sample, the intriguing question, sophisticated design and analyses, and transparent and comprehensive preregistration. There were a few issues that arose while reading the paper that I have enumerated below.

Thanks for your comments and review!

R2.2 To what extent are the authors observing conformity/compliance versus emotion regulation contagion? While reading the paper the open question that nagged at me was to what extent were participants responding differently based on compliance\conforming to others in the group versus the contagion process the authors claim. The authors raise the possibility compliance as the first limitation on page 21, but it was also clear in the results section that they worried about this as well. For example, when attempting to address this issue in the "general sentiments towards Palestinians" section they write "finding differences in the expression of sentiments towards Palestinians would be another indication that **genuine contagion** occurred" (emphasis mine). This seems to be the critical question here, was there emotion regulation contagion in that people felt differently when others in their group received an emotion regulation manipulation or were the untreated participants altering their responses to be more in line with those who were reappraising and offering (likely) more complex, multidimensional understanding of the conflict? The authors point to the sentiment data to show that when responses were private there was an effect of the proportion of reappraisers showing that negative sentiments decreased with more reappraisers in a group. Notably though, the effects observed were much smaller than when examining the public responses, there was not an overall main effect when comparing treated to untreated individuals, and there was an interaction between proportion and condition (eeked out a $t=1.97$) with control participants showing less negative sentiments with greater proportion of reappraisers in their group (page 16). The supplemental material is maybe more concerning with null effects for changes in attitudes towards Palestinians (SI: page 24) or concessions towards Palestinians (SI: page 25), showing no changes in private attitudes immediately following the group interaction.

These weaker and null effects don't completely undermine the effects here—the authors are examining emotion regulation process in a group, which they do show robust effects. However, these findings do bring up questions regarding the authenticity of the public pronouncements and the extent to which untreated individuals are genuinely influenced by the number of reappraisers versus parroting the majority or persuasive sentiment. You can see some influence of "majority

opinions” in the authors’ data in figure 4C. When looking at the figure on the far left – one reappraiser and five untreated group members, the reappraiser became more similar to the untreated group members, consistent with a conformity effect. Do the authors have any data that can address the question that these effects reflect genuine changes in emotion regulation among the untreated participants?

We completely agree that the question of compliance is the main limitation of the initial version of the manuscript. As you specified, there are some aspects that may provide some support to a genuine contagion, but these findings do not eliminate the possibility that compliance was driving the process. Having these concerns ourselves, and in light of your comments, we decided to add another study to the supplementary materials of the paper that directly addresses the question of compliance in the best way we could.

In our supplementary study (see SI), we designed a study that was similar to the original study reported in the paper with a few differences. The main and most important difference was that participants provided both text responses and ratings in responses to the stimuli, but while they saw each other’s text responses, they provided their ratings to the picture in private. We reasoned that if we found a reduction in participants’ private ratings during the task, it would be hard to argue that compliance was driving the effect.

The other differences were the following: First, we could not run this study in Israel due to the tumultuous and rapidly changing political situation, so instead, we ran this study with a US sample and provided general stimuli that elicited negative emotions using IAPS pictures (Lang et al., 1997). Second, participants believed that they were doing the experiment with other people, but in fact, we collected text responses prior to the study and presented them to participants as if they were generated in real time. Our texts were generated by participants who either completed our observing or reappraisal training and were asked to perform the task in the same way as presented in the current study. However, in the actual experiment reported here, we had only one real participant in each group, while the other texts were chosen from previous runs. This was done to save costs and make sure that we fully controlled the content that participants were exposed to. Third, all of our participants went through the observing training, and we manipulated the number of regulators in the group to be either zero (all observing) or three who were reappraisers.

Results of our supplementary experiment showed that participants’ private ratings were lower when they completed the task with three treated participants compared to non-treated participants. If compliance was the only process driving our findings, we should have not seen any reduction in participants’ ratings when they were doing the task with other regulators. Although compliance may change the texts that participants provided, given that these texts were thought to be shared by others, it’s very unlikely that they would have impacted the ratings that participants provided privately.

It is important to note that we do not think that this study rules out the possibility that compliance may be playing some role in shaping participants’ ratings. In fact, it’s very likely that it is. However, we believe that this additional study provides strong support for

the idea that compliance can't be the only factor driving performance. We further discussed this point in the limitation in the paper (p.25):

A second limitation is the concern that changes in ratings throughout the task may have been driven by compliance with other group members, rather than real changes in emotion. We directly addressed this concern with a supplementary study where participants gave ratings privately showing changes in their ratings despite not being seen by other group members. Note that we do not think this study rules out the possibility that compliance may be playing some role in shaping participants' ratings. Instead, we believe that it provides strong support for the idea that compliance can't be the only factor driving performance.

We appreciate you pushing us to further clarify this issue and provide this additional study.

R2.3 Word cloud. The authors use text analyses and then present a word cloud to provide a qualitative visual of the differences as a function of the experimental manipulation (figure 4a). I did not see the value of the word cloud for the paper. For a conference talk, this is a cute way to present an easy visual but for a scientific paper, especially one that can provide a quantitative answer to the question of whether the two conditions resulted in different word choices, this seemed beneath the rigor of the work. The word cloud is presented with only the two extreme conditions (no reappraiser versus six reappraisers) so it provides limited value. I would vote to remove this from the main paper.

We appreciate your opinions on the value of the word cloud figure. Following the editor's (E3) and reviewer 3's suggestions (R3.5), we created a more informative figure where we showed the top 100 most frequent words in each group and colored them by negative and positive. This figure qualitatively shows that there are more positive terms more highly ranked in the reappraisal condition. We further conducted a statistical test to examine the difference in the frequency of positive sentiment words between the conditions. Results of a Mann-Whitney U Test showed that the frequency ranking of positive words used by participants in the all-treated groups was significantly higher than that in all non-treated groups, $p = .013$. We believe this can provide additional support to our claim about the semantic differences in participants' texts under different treatments.

Nonetheless, given divergent opinions from the editor and reviewers, we've decided to move the word cloud to SI.

R2.4 Minor clarifications:

1. On page 7, the authors describe the experimental design as showing participants stimuli depicting "either Palestinian resistance to the Israeli occupation or Palestinian violence against Israel." Because of the language used here – "either/or" – I assumed

this was a between subjects factor. That is, the authors were studying emotion regulation contagion as a function of stimuli where Palestinians were either attackers or defenders. It turns out that this “condition” was not mentioned again, and it isn’t mentioned in the preregistration either (unless I missed it). I assume I was wrong, and this wasn’t a condition that was manipulated in the experiment. The authors might want to clarify what was meant by this sentence given other readers might also mistakenly believe there was another manipulated condition. If this was just an effort to use a variety of stimuli to show that there was not something unique about one type of stimuli (Palestinians only presented offensively) it might be worth explaining that.

Thank you for pointing this out! We did not manipulate the stimuli and all groups of participants viewed the same set of pictures containing Palestinian resistance to the Israeli occupation and Palestinian violence against Israel in random order. We made the language clearer in the paper as you suggested by saying: “*pictures containing Palestinian resistance to the Israeli occupation and Palestinian violence against Israel*” (p.8).

Reviewer #3 (Remarks to the Author):

R3.1 This is an extremely interesting paper with immediate and far-reaching public policy implications. I recommend that it be accepted for publication. I have a number of comments for how to improve the clarity and impact of the findings prior to publication. While I suspect I will not be alone in some of my confusions/questions, I do not think that addressing any of these is required for publishing this paper. I think it is already sufficiently clear and will be received with great interest by the broader research community. My numbered comments are below:

Thank you for the careful read and the wonderful comments!

R3.2 The authors are understandably cautious in the extent to which they frame their results as supporting particular theories of contagion. That said, their experiment is well setup to test and support complex contagion theory, which holds that peer reinforcement is required to trigger spreading of a counter-normative emotion. Their finding that treating more participants generates more spillover is consistent with this of course. But more specifically, Panel C of Figure 4 suggests that when the number of treated agents is 1/6 (a simple contagion), there are many trials where similarity to reappraisal is markedly lower than all cases where a complex contagion is satisfied (treated $\geq 2/6$). This suggests negative social influence, which is consistent with the dynamics of fractional thresholds in complex contagions (e.g., see Centola & Macy, 2007; Guilbeault & Centola 2021). Negative influence suggests that in the absence of sufficient peer reinforcement, the treated subjects may actually revert to the negative (or at least neutral, the ceasing of the positive) emotion. If the authors could show support for this, this would make for a very cool empirical contribution to complex contagion theory (perhaps better left for future work). One possible way to measure

this would be to look at whether individuals' similarity to reappraisal decreases overtime in the conditions with 1/6 treated. Furthermore, the finding that the greatest movement toward reappraisal happens between 60-80% regulators suggests we are dealing with quite a complex contagion that requires considerable reinforcement (both panel B and C of fig. 4). These patterns are highly consistent with complex contagion theory. I'd encourage the authors to highlight some of these connections in the discussion, since I think they would increase the impact of their findings for the contagion literature.

References:

- Centola, Damon, and Michael Macy. "Complex Contagions and the Weakness of Long Ties." *American Journal of Sociology* 113, no. 3 (2007): 702–34. <https://doi.org/10.1086/521848>.
- Guilbeault, Douglas, and Damon Centola. "Topological Measures for Identifying and Predicting the Spread of Complex Contagions." *Nature Communications* 12, no. 1 (July 20, 2021): 4430. <https://doi.org/10.1038/s41467-021-24704-6>.

Thank you so much for bringing in the perspective of complex contagion! We agree that emotion regulation could be a complex contagion process and we found the complex contagion theory very helpful in interpreting the mechanism of the findings when there is only 1 regulator in the group. Indeed, expressing emotions and appraisals different from the majority in the group is a counter-normative behavior that might trigger punishment. Therefore, one likely requires a certain amount of social affirmation in the group for contagion to occur. As panel C of Fig. 4 shows, when regulators were 1/6 of the group, the semantic similarity between participants' produced text and reappraisal seems to have decreased over time, which is consistent with what the complex theory would suggest. We did not validate this trend with statistical analysis, as the number of participants in each specific proportions group was too small to provide analysis with suitable power. However, we discussed the connection between our findings and the complex theory in the discussion section (p.23):

As panel B of Figure 3 shows, when only 1 out of 6 members in the group was a regulator, the similarity to reappraisal language in the texts of both treated and non-treated participants appeared to decrease over time. These findings are aligned with complex contagion theory such that the success of complex contagion depends upon interaction with multiple sources of affirmation (Centola & Macy, 2007; Guilbeault & Centola, 2021).

Thank you for pointing out the connection!

R3.3 Throughout, the authors use the term “control participant” to refer to both untreated participants in networks exposed to treated participants and participants in an “only-control” condition (no reappraisal in the group). I find this a bit confusing. It seems more appropriate to describe the only-control groups as the control condition. Non-treated agents in the groups with the treated participants are in the treatment condition, they just happen to not have received the treatment directly but rather as a

spillover. I found this particularly confusing when the authors talked about interactions in their models where the movement toward reappraisal was strongest among control participants; I found myself asking, but weren't the control participants not exposed to treated participants? Then I had to go through the manuscript again to clarify. I'd encourage the authors to refine the terminology to avoid similar confusions among other readers.

Thanks for pointing this confusion out. We revised the paper to make sure that the terminology was consistent such that the participants who received the reappraisal training are called "treated participants" and the participants who did not receive reappraisal training (i.e., receive the observing instructions) are called "non-treated participants". We now use the term control to only refer to the group where there are no treated participants.

R3.4 This might be out of scope and better left for future work, but it would be super cool to see if the way that treated individuals communicate (or say, the extent of reappraisal or positivity they communicate) is mediated by the presence of communications from other people who have been treated. The presence of others may create a safe space for communicating counter-normative affect. I am reminded of Traeger et al.'s (2020) paper showing that injecting bots that communicate vulnerable messages can help steer a group discussion in a civil direction by making it psychologically safe for people to communicate similarly vulnerable sentiments.

Reference

-Traeger, Margaret L., Sarah Strohkorb Sebo, Malte Jung, Brian Scassellati, and Nicholas A. Christakis. "Vulnerable Robots Positively Shape Human Conversational Dynamics in a Human-Robot Team." *Proceedings of the National Academy of Sciences* 117, no. 12 (March 24, 2020): 6370–75. <https://doi.org/10.1073/pnas.1910402117>.

This is a great point and it's very likely that having participants who were trained in reappraisal could have impacted the quality of discourse. However, because our design was based on a somewhat controlled structure, where participants provided texts and then saw each other's text without being able to add any texts, it was very hard to measure the impact of communication more broadly. We do agree that this could be a fantastic future direction and have added this idea to the discussion (p.26).

Besides, in more natural conversations, reappraisal training might not only change how group members react to stimuli but also how people respond to each other's comments and the quality of discourse.

R3.5 I think the word cloud analysis could be tighter. Why not show the top n most frequent words in each group and color them by negative and positive, and show there is more positive terms more highly ranked in the positive condition? This could also be evaluated for statistical significance (difference in average ranking of positive items) in a frequentist sense. But regardless, I buy the results here.

Thank you so much for the suggestion of revising the word cloud figure! We identified the sentiment of each word with the ‘tidytext’ package in R (Silge & Robinson, 2016). As you suggested, we displayed the top 100 most frequent words in each condition and color them by negative and positive sentiments.

We conducted a Mann-Whitney U Test to test the difference in the ranking of positive sentiment words in the whole sample between the two conditions. Results showed the ranking of positive words was significantly higher in the all-treated condition than all non-treated condition ($W = 282.0, p = .013$).



Please note that there are some key differences in word frequency between the original and new figures. We realized an important detail was missing when we reported the original word cloud: each unique word appeared only once, in the condition where its frequency was higher, even if it was present in both conditions. For instance, in the original plot, the word 'anger' was shown only in the non-treated condition, where it was mentioned more frequently, despite being the most common word in both conditions. Recognizing that this could cause confusion, we corrected this in the new figure by displaying words in both conditions if they appeared in the top 100 most frequent words in both.

Nonetheless, given divergent opinions from the editor and reviewers and the potential confusion the graph can cause, we've decided to move the word cloud to SI.

R3.6 I found it a bit confusing that the authors fit different functional forms to different aspects of the diffusion process, without elaborating on the qualitative implications.

At a group level, they emphasize the exponential, which makes sense. But I don't feel like the quadratic is sufficiently explained. The quadratic of course typically evokes a u-shaped curve. But the authors describe this functional form as a kind of step function writing that this model "suggested that there is relatively little change in the emotions of the non-treated participants until a certain proportion, at which point changes become much greater with every increase..." Qualitatively, this makes good sense to me, but it's not clear that only the quadratic functional form would capture this, so a little more elaboration on this point (even just discursively) would be helpful. This would help clarify in what way the authors show a "non-linear reduction in conflict-related negative emotions" by clarifying the kind of non-linearity observed (is it flat until a tipping point and then curve positive, like a hockey stick, or something more u-shaped).

Thank you for this comment. We modified the paper to provide a bit more detail on why we chose these specific models. More specifically, the linear model was included to test for a simple, proportional relationship where emotions change consistently with the proportion of treated participants. The quadratic model was employed to examine the possibility of an inverted U-shaped relationship, where emotions would initially decrease or increase with treatment in some form of reactance (Goldenberg et al., 2014) but then reverse as the proportion continues to rise. The cubic model was considered to capture situations where contagion may be particularly potent either in low or high proportions. This could be driven by the way people represent the number of group members who are regulating, as previous research on social perception revealed a cubic model in the way people represent collective information (Goldenberg et al., 2022). The logarithmic model was included to test the hypothesis that emotions might change rapidly at first as the proportion of treated participants increases but then taper off, indicating diminishing returns of the proportion of treated participants. Finally, the exponential model was chosen to explore the possibility that there is relatively little change in emotions when the proportion of treated participants is low, but the impact becomes more pronounced as the proportion increases, suggesting a potential threshold effect. We further clarified these choices in the actual manuscript (p. 11-12):

We chose five different alternative models based on the following rationale. The linear model was included to test for a simple, proportional relationship where emotions change consistently with the proportion of treated participants. The quadratic model was included to examine the possibility of an inverted U-shaped relationship, where emotions would initially increase with treatment in some form of reactance, but then reverse as the proportion continues to rise. The cubic model was considered to capture situations where contagion may be particularly potent either in low or high proportions. This could be driven by the way people represent the number of group members who are regulating, as previous research on social perception revealed a cubic model in the way people represent collective information. The logarithmic model was included to test the hypothesis that emotions might change rapidly at first as the proportion of treated participants increases but then taper off, indicating diminishing returns of the proportion of treated participants. Finally, the exponential model was chosen to explore the possibility that there is relatively little change in emotions when the proportion of treated participants is low, but the impact

becomes significantly more pronounced as the proportion increases, suggesting a potential threshold effect.

R3.7 The question concerning the movement toward the pure reappraisal vector – how much of this is driven by a shift in affect vs. the adoption of reappraisal-specific content language (e.g., increased use of the term “interpretation” or “reappraisal” itself), which could theoretically happen without a shift in affect. I think it would be sufficient to show in the appendix a handful of representative examples of people’s actual messages – from the treated participants and/or the untreated by infected participants – whose messages scored as being particularly close to the reappraisal vector. This would add some nice color.

We appreciate your suggestion to show the actual texts that have the highest reappraisal similarity score. We added a table that illustrates the ten most representative texts according to their similarity to the reappraisal vector in the SI. Almost identical responses were combined.

1. It seems that the soldier is not hurt.
2. At least they don’t have / they are not holding any weapons.
3. It seems that they are stopping.
4. He doesn’t seem badly injured.
5. At least they are kept in their own zone.
6. It seems like he is walking and not on a stretcher.
7. Perhaps the soldier is there to take care of the wounded.
8. At least they caught some of the killers.
9. They were able to get him alive.
10. I hope that some survived.

R3.8 I’d also mention in the discussion that an important direction for future research is to examine the effects of anonymity in this setting. If people have access to the identities of other people – or, perhaps, if they are knowingly interacting with people they have an actual history with, like friends or family – this could significantly impact their willingness to express either strong positive or negative views (depending on what they expect to be the normative stance of their network). For instance, we’ve shown that exposure to identifying information can meaningfully change people’s willingness to use social information (Guilbeault et al. 2018), consistent with broader empirical studies of social media (Aral et al. 2023). I don’t think this takes away from the current results, which standalone, but I think this marks an important future direction worth noting.

References

- Guilbeault, Douglas, Joshua Becker, and Damon Centola. “Social Learning and Partisan Bias in the Interpretation of Climate Trends.” *Proceedings of the National Academy of Sciences* 115, no. 39 (September 25, 2018): 9714–
19. <https://doi.org/10.1073/pnas.1722664115>.

-Taylor, Sean J., Lev Muchnik, Madhav Kumar, and Sinan Aral. "Identity Effects in Social Media." *Nature Human Behaviour* 7, no. 1 (January 2023): 27–37. <https://doi.org/10.1038/s41562-022-01459-8>.

Thank you for your suggestions on discussing the influence of people's identity! We agree that people's identity will have a profound impact on their motivations to experience and control certain emotions, which in turn, can shape the emotion contagion process. This is especially interesting because the literature on emotions and anonymity has somewhat shifted in the last few years, where initial findings suggested that more anonymity will always lead to stronger negative emotional expressions but more recent work suggests more complex findings in this space (H.-M. Kim et al., 2021; Lapidot-Lefler & Barak, 2012; Rösner & Krämer, 2016). The present experiment was conducted among people who shared a strong social identity and could therefore have a relatively strong motivation to express stronger negative emotions toward Palestinians in groups. Linked to your previous point on the complex contagion theory, emotion regulation in this case may require more social affirmation to be able to spread since there could be a stronger normative expectation on expressed emotions given their identity. In other contexts where people don't have such identity information, we might expect the threshold of the proportion of treated group members for emotion regulation to work to be lower. We added it as a limitation of the paper (p.26):

A fourth limitation of the paper exists in its context of intergroup conflicts where people's group identities were particularly salient. This might influence people's willingness to take in social information and to express strong emotions in the group. In other contexts, where people's group identity was less salient, we might expect a different threshold of the proportion of treated group members for emotion regulation to work.

Best of luck with this revision. Thank you for the pleasure of reviewing this excellent work.
-Doug Guilbeault

Reviewer #4 (Remarks to the Author):

R4.1 The topic of this paper, to better understand the social dynamics of the regulation of negative emotions in groups, is compelling. The results that increasing the proportion of treated participants produces more emotional regulation among non-treated group members is important and novel, especially given that the authors could not predict even the direction of the effects.

Thank you for your review!

R4.2 I have a number of suggestions for improving the paper. I was left with a number of questions regarding what the authors thought was spreading here, and the nature of the relationship between emotions and emotional regulation language: are the authors making a claim that it is the emotional regulation that is spreading initially, and that

impacts how participants report their negative emotions (which implies a mediation analysis using the semantic analysis results) or is what is spreading the emotions (emotional ratings) themselves? I think this distinction matters because we know from other contagion literature that what is being spread matters for how it spreads (e.g., Centola's work on complex contagions).

We agree that the question of whether it is emotion or emotion regulation that spreads is a central concern for this paper. Starting with emotion contagion, your comment seemed to suggest that it may only be the ratings that are driving the effect. While the initial version of the paper attempted to address this limitation, we agree that more evidence were needed. In our revision, we have addressed this limitation by adding another study in which participants were exposed to others' texts as in the initial study, but provided their ratings privately. We found that even in this task, where participants were only exposed to texts produced by treated participants, they also rated the pictures as less intense compared to being exposed to non-treated participants. We provide a detailed description of the study in response R2.2.

Having established that the texts that participants were exposed to were able to lead to some change in participants' ratings, the question of emotion contagion versus emotion regulation still remains to some extent. It's possible that what led to the reduction in ratings was exposure to lower emotions rather than exposure to reappraisals. We believe that our text analysis provides a partial response to this issue. Non-treated participants who were exposed to responses of treated participants were more likely to use texts that were similar to reappraisal in their own responses. This is especially striking given that participants' text responses were provided to novel pictures, before seeing others' responses.

In summary, we agree that both the exposure to ratings and to emotions in general played a role in impacting participants' responses, but we also believe that both the new study and our existing analysis provided evidence that emotion regulation also spread between participants, at least to some extent. To make sure that the issue is properly addressed we revised the limitations of the paper to include these considerations (p.24-25):

The first limitation is related to the differentiation of various mechanisms behind emotion regulation contagion. It is possible that what led to the reduction in ratings was both emotion contagion, where mere exposure to emotion led to non-treated reduction in emotions, and emotion regulation contagion, where the actual use of reappraisal spread to the non-treated participants. Our text analysis provides a partial response to this issue. Non-treated participants who were exposed to responses of treated participants were more likely to use texts that were similar to reappraisal in their own responses. This is especially striking given that participants' text responses were provided to novel pictures, before seeing others' responses. Future research should try to isolate emotion contagion and emotion regulation contagion in more controlled ways.

A second limitation is the concern that changes in ratings throughout the task may have been driven by compliance with other group members, rather than real changes in emotion. We directly addressed this concern with a supplementary study where participants gave

ratings privately showing changes in their ratings despite not being seen by other group members. Note that we do not think this study rules out the possibility that compliance may be playing some role in shaping participants' ratings. Instead, we believe that it provides strong support for the idea that compliance can't be the only factor driving performance. Additionally, participants' general sentiments towards Palestinians after the task, when participants knew that their ratings would not be observed by other participants, were also reduced with the increased proportion of treated participants, providing another evidence that compliance was not the only factor driving the results. Nevertheless, future studies should use other non-self-report methods to examine changes in emotion as a result of regulation contagion (see for example Oveis et al. (2020)).

R4.3 Relatedly, I unsure whether the authors would argue that what is happening in the intervention is the same or fundamentally different than the spread (or convergence?) of emotions in a group with different initial proportions of negative emotions (“naturally occurring” vs. the product of the intervention).

It would be great to clarify terminology here to make sure that we are properly interpreting the meaning you assign to the terms convergence and contagion here and throughout. We understand your interpretation of the term convergence as an umbrella term designed to capture different aspects of emotional influence (see Anderson et al., 2003; Fritsch et al., 2023; Parkinson, 2020; Wolf et al., 2023). When the term emotion convergence is used, contagion is often used to represent a more automatic process of influence mainly driven by mimicry. We agree that the term contagion was indeed traditionally designed to capture these more automatic processes, but over the years, the term contagion has been used as the umbrella term that captures all sorts of processes of influence, from mimicry to social appraisal. This decision to use contagion and not convergence as the umbrella term for a variety of processes is also reflected in our own previous research. The reasons we chose to do this are the following. First, contagion is a much more commonly used term than convergence (6 times as much according to Google Scholar) which makes it a better candidate to be the umbrella term going forward. Furthermore, as mentioned, more recent use of the term contagion captures influence as a result of multiple different mechanisms. In addition, the term convergence seems to suggest that all members of a group converge to each other to some middle value, whereas contagion can be applied to more asymmetrical processes of influence, such as the one described in the paper.

Regardless of the term used to describe the effect we fully agree that the previous version of the manuscript did not properly define emotion and emotion regulation contagion and explain the underlying mechanisms that are driving its occurrence. In line with comments from Reviewer 1 (see response R1.2), we now provide a clear definition of contagion and the mechanisms that may drive the findings (p.4-5):

In the current project, we examined the possibility of leveraging the process of emotion regulation contagion to change group emotions. We define emotion regulation contagion as the spread of emotion regulation strategies from people who have received emotion regulation interventions to those who are not treated. We see emotion regulation contagion of cognitive reappraisal, as tested in the current project, as driven by two processes. First,

regulation contagion may occur as a result of social appraisals, which are defined as the adoption of other people's appraisal of a certain situation in a way that impacts one's emotional response. In the current case, non-treated participants would adopt the reappraisals of the treated participants to conflict situations. Second, we believe that regulation contagion is driven by social learning, such that non-treated participants would learn how to generate reappraisals to similar stimuli and apply this knowledge to new stimuli.

We wish to distinguish the process of emotion regulation contagion from two similar processes (See Figure 1). The first is emotion contagion, where the mere expression of emotions of the treated participants is impacting the emotions expressed by the non-treated participants. Although emotion contagion is likely playing a role in impacting non-treated participants' emotions, we examined whether non-treated participants are also adopting the use of reappraisal, in addition to adjusting their emotion expressions to match those of other groups members. The second related process is compliance, where no real change in emotion experiences occurs, but merely a change in expressed emotions as a result of social pressure. Here again, although compliance may impact the results, we will show some evidence that non-treated participants are changing their emotions privately.

Thank you for pushing us for more clarity on this matter.

R4.4 The authors need to further develop, and consider possible tests for, the potential mechanisms by which emotions/emotional regulation spread in a group. I would have loved to see much more of the authors' thinking (which I believe appears in some of their other work) on whether and how the spread of emotions (or emotional regulation) differs from the spread of social behaviors (simple or complex), attitudes, perceptions in groups, etc. Are the models we already have for these types of diffusion processes adequate in this case or not? Should we think about this as an influence process (participants are learning what is acceptable text and emotional ratings in a group) or actually a "contagion" where seeing other people with negative emotions gives a participant more negative emotions? (I could see this argument holding for negative emotions, but the opposite, seeing more neutral emotions and "catching" those seems far less plausible, suggesting asymmetrical processes by proportion of treated participants in a group.) Also, should we characterize emotions/the use of emotional regulation as "spreading" as in diffusion, or something more like converging? I think the convergence language makes more sense here, but it is hard to tell from these analyses and without the group-level results, which I wish the authors had included. (I'd also like to know the N of the number of groups, which doesn't appear in the manuscript or SI.)

Thank you for your great comment, which touches on a few different issues. The first is whether the results in this study were driven solely by compliance or contagion. We agree that this was the biggest limitation of the study presented in our initial submission, and thus we have added an additional study to Supplementary Information to directly address this limitation. In this study, participants were merely exposed to other group members' texts but were provided their ratings in private. We find a reduction in participants' private

ratings when exposed to texts provided by treated participants. See a more detailed description of the study and the results in response R2.2.

As for your second question regarding the choice of using contagion and not convergence, we prefer using the term contagion and provided a reason for this choice in response R4.3. We do, however, agree that the term has to be properly defined and the mechanisms by which contagion occurs (explained in detail in response R1.2) should be clarified.

As for your third question on the number of groups, there were 829 distinct groups in total.

R4.5 I appreciated the testing of multiple functional forms of the relationship between the proportion of treated participants and the results are interesting, but they aren't a substitute for the elaboration of nature of spread/convergence in this case (especially given how minor the AIC differences were between them); e.g., should we think about spread/convergence as a threshold effect, and what inference processes in the group might give rise to that? As the authors know, there are many types of spread/contagion/spillover processes and I would have liked a more precise definition of which they believed they were observing and why.

Your question can be interpreted as both striving for more conceptual clarity at the onset as well as a better interpretation of the findings. In terms of the terminology, as stated in responses R1.2 and R4.3, we agree that further clarification is needed to explain the different terms. In terms of the type of contagion occurring, as indicated in response R3.2 we believe that the right way to evaluate this contagion process is using the term complex contagion, which is aimed to capture processes in which a certain critical mass is required to see some behavioral change. As for the specific model, we agree that the AIC differences are minor and the result should be taken with caution. We added a clarification of this small difference in the paper (p.13):

“It is worth mentioning that the AIC differences are minor and the model comparison result should be taken with caution.”

R4.6 Another significant concern I had was whether there was differential attrition from the experiments as a function of condition (proportion of treated individuals in a group) or as a function of having more negative emotions (on average). In the SI, the authors report starting with 5040 groups participants, and, after exclusions and attrition, ended with 2,830 people, indicating massive amounts of attrition. I would like to see the rate of attrition by condition (and by average emotion ratings) to ensure that the results are not significantly impacted by this attrition (e.g., are those with the most negative emotions at the outset more likely to drop out? Are those in low proportion conditions more likely to drop out?). (Also unclear was whether, in the analyses, the number of group members varied across trials. Whether they did or did not, the authors should discuss whether we should expect that the process of influence/spread/convergence differs at 50% treated group members when the group size is 4 vs. 6, for example.)

We recognize that the overall attrition rate (52.8%) may seem high, but if you look at the breakdown of the 5040 who finished the instructions, 1604 of them did the task twice in the same condition because we allowed these participants who did not start the actual task to run through the task again. To make it clear to readers, we moved the paragraph about attrition from the SI to the main manuscript (p.27-28):

Our aim was to recruit 2,800 participants, 400 in each proportion group. Participants were Jewish Israelis who were recruited through iPanel, an Israeli survey company in exchange for 25 NIS (~\$7). Participants could start the task only after being grouped with five other participants. If after 5 minutes of waiting we were not able to group participants, they were transferred to complete a different task. Therefore, our initial sample was much larger than the one we were able to group. Our initial recruited sample was 6,161 participants. Of those participants, 5040 were able to be grouped in the initial round and started the task, but may have dropped in many different stages. The list of 5040 included participants who started the task twice. We allowed participants who did not start the actual rating phase (1604) to run through the task again. As preregistered, our dropout criteria were (numbers may be overlapping): (1) Participants who failed the reading check in which they were asked to choose the option that best describes the strategy that they were assigned to before starting the actual task (146 participants). (2) Participants who did not provide text responses, provided nonsensical text responses, or provided the exact same text response (10 participants). (3) groups who reported that they did not see the images or who had technical issues (6 groups). Participants were also automatically dropped from the study if they did not complete 15 of the 20 trials. After all of these omissions we had 2,830 who completed the whole task. We removed all participants who were the only person remaining in the group, leaving us with 2,659 participants which is our final sample. (Gender: 1,158 males (43.6%), 1495 females (56.2%), 6 other or refused to say (0.2%); Age: $M = 42.07$, $SD = 14.618$). It is worth mentioning that the average attrition rate of non-treated participants (38.0%) is lower than that of the treated participants (50.2%). The difference in attrition rate might inflate the results because the treated participants may be generally better at reappraisal. Nonetheless, it is important to note that we are using the actual proportion of treated participants in our analysis at any time. This means whenever the dropout happened, it also influenced the proportion of the treated participants. We also conducted as-treated analyses with the original proportion values to ensure the robustness of our findings (See SI).

We also recognize that attrition may be impacting the results and below we provide the difference in attrition as a function of the proportion of regulators in the group. Starting with the attrition of non-treated participants, we see very little differences among proportion conditions.

Attrition Rate of Non-Treated Participants						
Number of treated participants in the group	0	1	2	3	4	5
Attrition rate (%)	39.3	37.1	35.9	39.3	40.0	36.1

Moving to the treated participants, we do see a difference in the attrition between the condition where there were only 1 or 2 treated individuals and the condition where there were 4 treated individuals. There is also a main effect such that the non-treated participants were less likely to drop, which might inflate the results because the treated participants may be generally better at reappraisal.

	Attrition Rate of Treated Participants					
Number of treated participants in the group	1	2	3	4	5	6
Attrition rate (%)	45.8	46.1	50.5	56.2	51.5	47.3

It is important to note that we are using the actual proportion in our analysis at any time. This means whenever the dropout happened, it also influenced the proportion of the treated participants. That is, the impact of attrition can be partially captured by the change of the actual proportion in the data, which helps to mitigate the confounding effect. In order to improve the clarity on this issue, we now added an as-treated analysis and we show that results stay the same (See SI), which means that attrition may not have impacted the results as much.

We added a discussion of the difference in attrition rate to the manuscript (p.):

It is worth mentioning that the average attrition rate of non-treated participants (38.0%) is lower than that of the treated participants (50.2%). The difference in attrition rate might inflate the results because the treated participants may be generally better at reappraisal. Nonetheless, it is important to note that we are using the actual proportion of treated participants in our analysis at any time. This means whenever the dropout happened, it also influenced the proportion of the treated participants. We also conducted as-treated analyses with the original proportion values to ensure the robustness of our findings (See SI).

R4.7 More minor comments:

The front end of the paper describing the procedures should be rewritten for clarity, e.g. the description of the control condition on p. 6, that the group size was not only 6 and what the distribution of group sizes were (related to above comment), etc. There were a number of key details about the procedures (the sequential randomization of group to condition and participant to condition; the existence of 20 experimental trials, etc.) that only appear in the Methods section that would have helped me understand the analytical choices better in the first reading.

Thank you for the suggestion for clarity in our writing of the experiment! We now detail the condition assignment as a sequential randomization process as you suggested, and add the total number of trials. We already explained on page 7-8 that the dropouts could change the actual proportion of non-treated participants, and we make it clear that the actual group size can be smaller than 6. We revised the experimental procedures at the front end of the paper according to your suggestions (p.7-8):

The conditions in this study were randomly assigned in two steps. We first randomly assign the number of treated participants, varying between zero and six, to each group. Within each group, we then randomly assigned the individual condition (treated or non-treated) to each participant. After being assigned to the group and to their condition (treated or non-treated), participants completed our emotional dynamics task in their group. During the task, participants saw pictures containing Palestinian resistance to the Israeli occupation and Palestinian violence against Israel. Pictures were mostly of terror attacks, or Palestinian demonstrations against Israel. These pictures were used in previous studies to elicit strong negative emotions among Jewish Israelis, mostly anger and sadness²⁶. After viewing each picture, participants were asked to produce a brief text that expressed their emotions (“What comes up for you when you see the picture?”, see further details in methods). Participants were then able to see the texts of all other participants in real time. They were then asked to rate their emotions in response to the picture on a 1-neutral to 10-very negative scale and again saw others’ real-time ratings (“Please rate the degree of negative emotions you feel in response to the picture”, see full description in Methods). Participants therefore had two points during each trial when they could impact others’ emotions, when observing others’ text and when observing others’ ratings (marked in red in Figure 1). In total, there were 20 trials in this task.

It is important to note that due to natural dropout in online studies, the size of the group, and the actual proportion of people that went through the reappraisal manipulation, sometimes changed during the task. Therefore, the actual group size could be smaller than six. To counter these variations during the task, we used the actual proportion of reappraisers within each group and in each trial rather than the assigned proportion (See the as-treated analyses with the proportions of each group as they were initially assigned in SI). We also controlled for group size in all of the following models.

R4.8 The authors claimed that treating “25-40% of the group with a reappraisal intervention results in significant and reliable emotion contagion at the group level” (p. 14). Based on Figure 2B, it looks like the effect only become “reliable” (which I believe to mean consistently significant and in the same direction) at 40%. Please clarify.

We now say (p.15): “treating at least 40% of the group with a reappraisal intervention results in significant and reliable emotion contagion at the group level”.

R4.9 The authors make a claim about this study testing the proportion of the network treated in a fully connected network (p. 5). Given that the authors do not use any information about the nature of the relationship among participants, the comparison to a network study is not appropriate here.

By saying “a fully connected network”, we referred to the fact that every member in the group can view all groupmates’ texts and emotions. We removed the term “fully connected network” to avoid confusion.

References

- Anderson, C., Keltner, D., & John, O. P. (2003). Emotional convergence between people over time. *Journal of Personality and Social Psychology*, 84(5), 1054–1068.
- Cheshin, A., Rafaeli, A., & Bos, N. (2011). Anger and happiness in virtual teams: Emotional influences of text and behavior on others' affect in the absence of non-verbal cues. *Organizational Behavior and Human Decision Processes*, 116(1), 2–16.
<https://doi.org/10.1016/j.obhdp.2011.06.002>
- Clément, F., & Dukes, D. (2017). Social appraisal and social referencing: Two components of affective social learning. *Emotion Review*, 9(3), 253–261.
<https://doi.org/10.1177/1754073916661634>
- Frattaroli, J. (2006). Experimental disclosure and its moderators: A meta-analysis. *Psychological Bulletin*, 132, 823–865. <https://doi.org/10.1037/0033-2909.132.6.823>
- Fritsch, J., Leis, O., Wohlfarth, D., Triller, N., & Jekauc, D. (2023). The role of emotional convergence in sport: A scoping review. *Sport, Exercise, and Performance Psychology*.
<https://doi.org/10.1037/spy0000330>
- Goldenberg, A., Saguy, T., & Halperin, E. (2014). How group-based emotions are shaped by collective emotions: Evidence for emotional transfer and emotional burden. *Journal of Personality and Social Psychology*, 107(4), 581–596. <https://doi.org/10.1037/a0037462>
- Goldenberg, A., Schöne, J., Huang, Z., Sweeny, T. D., Brady, T. F., Robinson, M. D., Levari, D. E., Zaki, J., & Gross, J. J. (2022). Amplification in the evaluation of multiple emotional expressions over time. *Nature Human Behavior*.

Guillory, J., Spiegel, J., Drislane, M., Weiss, B., Donner, W., & Hancock, J. (2011). Upset now?:

Emotion contagion in distributed groups. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, May 2011*, 745–748.

<https://doi.org/10.1145/1978942.1979049>

Hatfield, E., Cacioppo, J. T., & Rapson, R. L. (1994). *Emotional contagion*. Cambridge University Press.

Kim, D. A., Hwong, A. R., Stafford, D., Hughes, D. A., O'Malley, A. J., Fowler, J. H., & Christakis, N. A. (2015). A Randomised Controlled Trial of Social Network Targeting to Maximise Population Behaviour Change. *Lancet (London, England)*, 386(9989), 145–153. [https://doi.org/10.1016/S0140-6736\(15\)60095-2](https://doi.org/10.1016/S0140-6736(15)60095-2)

Kim, H.-M., Bock, G.-W., & Kim, H. S. (2021). A new perspective on online malicious comments: Effects of attention and neutralization. *Information Technology & People*, 34(1), 50–67. <https://doi.org/10.1108/ITP-04-2019-0179>

Kramer, A. D. I., Guillory, J. E., & Hancock, J. T. (2014). Experimental evidence of massive-scale emotional contagion through social networks. *Proceedings of the National Academy of Sciences*, 111(24), 8788–8790. <https://doi.org/10.1073/pnas.1320040111>

Lang, P. J., Bradley, M. M., & Cuthbert, B. N. (1997). International Affective Picture System (IAPS): Technical manual and affective ratings. *NIMH Center for the Study of Emotion and Attention*, 39–58. <https://doi.org/10.1027/0269-8803/a000147>

Lapidot-Lefler, N., & Barak, A. (2012). Effects of anonymity, invisibility, and lack of eye-contact on toxic online disinhibition. *Computers in Human Behavior*, 28(2), 434–443.

<https://doi.org/10.1016/j.chb.2011.10.014>

- Manstead, A., & Fischer, A. H. (2001). Social appraisal: The social world as object of and influence on appraisal processes. In K. R. Scherer, A. Schorr, & T. Johnstone (Eds.), *Emotion: Theory, methods, research* (pp. 221–232). Oxford University Press.
- Morris, A. S., Silk, J. S., Steinberg, L., Myers, S. S., & Robinson, L. R. (2007). The Role of the Family Context in the Development of Emotion Regulation. *Social Development*, 16(2), 361–388. <https://doi.org/10.1111/j.1467-9507.2007.00389.x>
- Niedenthal, P. M., Winkielman, P., Mondillon, L., & Vermeulen, N. (2009). Embodiment of emotion concepts. *Journal of Personality and Social Psychology*, 96(6), 1120–1136.
- Parkinson, B. (2020). Intragroup Emotion Convergence: Beyond Contagion and Social Appraisal. *Personality and Social Psychology Review*, 24(2), 121–140. <https://doi.org/10.1177/1088868319882596>
- Peters, K., & Kashima, Y. (2015). A multimodal theory of affect diffusion. *Psychological Bulletin*, 141(5), 966–992.
- Porat, R., Halperin, E., & Tamir, M. (2016). What we want is what we get: Group-based emotional preferences and conflict resolution. *Journal of Personality and Social Psychology*, 110(2), 167–190. <https://doi.org/10.1037/pspa0000043>
- Rösner, L., & Krämer, N. C. (2016). Verbal Venting in the Social Web: Effects of Anonymity and Group Norms on Aggressive Language Use in Online Comments. *Social Media + Society*, 2(3), 2056305116664220. <https://doi.org/10.1177/2056305116664220>
- Silge, J., & Robinson, D. (2016). tidytext: Text Mining and Analysis Using Tidy Data Principles in R. *Journal of Open Source Software*, 1(3), 37. <https://doi.org/10.21105/joss.00037>

Silvers, J. A. (2022). Adolescence as a pivotal period for emotion regulation development.

Current Opinion in Psychology, 44, 258–263.

<https://doi.org/10.1016/j.copsyc.2021.09.023>

Sorce, J. F., Emde, R. N., Campos, J., & Klinnert, M. D. (1985). Maternal emotional signaling:

Its effect on the visual cliff behavior of 1-year-olds. *Developmental Psychology*, 21, 195–200.

Walton, G. M., & Cohen, G. L. (2011). A Brief Social-Belonging Intervention Improves

Academic and Health Outcomes of Minority Students. *Science*, 331(6023), 1447–1451.

<https://doi.org/10.1126/science.1198364>

Wang, K., Goldenberg, A., Dorison, C., Miller, J., Uusberg, A., Lerner, J. S., & Gross, J. J.

(2020). A global test of brief reappraisal interventions on emotions during the COVID-19 pandemic. *Nature Human Behaviour*. <https://doi.org/10.31234/osf.io/m4gpq>

Wolf, S. A., Heerdink, M. W., & van Kleef, G. A. (2023). A Motivational Account of

Convergence in Emotion Expressions Within Groups: The Emotional Conformity Framework. *Emotion Review*, 15(4), 363–379.

<https://doi.org/10.1177/17540739231182241>

Yeager, D. S., Hanselman, P., Walton, G. M., Murray, J. S., Crosnoe, R., Muller, C., Tipton, E.,

Schneider, B., Hulleman, C. S., Hinojosa, C. P., Paunesku, D., Romero, C., Flint, K.,

Roberts, A., Trott, J., Iachan, R., Buontempo, J., Yang, S. M., Carvalho, C. M., ...

Dweck, C. S. (2019). A national experiment reveals where a growth mindset improves achievement. *Nature*, 573(7774), 364–369. <https://doi.org/10.1038/s41586-019-1466-y>