nature portfolio

Peer Review File

Redesigning algorithms to intervene on social norm misperceptions during a national election

Corresponding Author: Dr William Brady

Any redactions in this file are there to maintain patient confidentiality, the confidentiality of unpublished data, or to remove third-party material.

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

Attachments originally included by the reviewers as part of their assessment can be found at the end of this file.

Editorial Note: This manuscript was submitted to Nature as a Registered Report. The reviews and rebuttals for both the pre-study (Stage 1) and post-study (Stage 2) are included within this peer review file.

STAGE 1

Version 1:

Reviewer comments:

Referee #1

(Remarks to the Author)
The crux of the proposed experiment is to test the influence of standard engagement based social media algorithms on (a) the content that users see, (b) what forms of political dialog they perceive to be normative, and (c) their animosity toward members of other political parties.

I like the design. I find random assignment to one of three of custom ranking algorithms on the Bluesky platform to be an excellent method with which to manipulate the influence of social media algorithms. The purpose and value of the engagement based and reverse chronological algorithms is good and clear. I get the value of the “representative” feed but it might be better characterized as the feed intended to be depolarizing. It’s hard to say that it’s designed to be fully representative. One possibility is to admit that more plainly and move forward. Another would be to train it to more randomly sample from “civics” posts that appear on Bluesky (with the current superposter and toxicity filters) rather than constrain to a 75%/25% ingroup vs. outgroup split. It could be the case, which would be interesting, that the truly representative posting would be the most polarizing feed of the three. The 2018 PNAS cited explored the effect of one twitter bot on polarization metrics so this does seem like it could add quite a lot to the literature.

More generally, I think the dependent variables make sense. My one critique would be that the social norm perceptions are largely focused on the platform and do less to tap into broader perceptions about what is normative political discourse.

(Remarks on code availability)

Referee #2

(Remarks to the Author)
This pre-registered report proposes an experiment that will compare three social media feeds: a feed maximizing engagement (this is what most current algorithms prioritize), a reverse-chronological feed (a common benchmark), and a “representative diversification” feed the authors will create. The authors will recruit 1,600 participants to use the open-source BlueSky social media platform and randomize their feed.

The experiment focuses on the three main outcomes: the type of content displayed to users and whether users engage with it, how users perceive dialogue on and off social media, and partisan animosity.

The main contribution of this paper is to analyze a potential new algorithm for ranking posts. Scholars are worried that current social media feeds result in polarizing echo chambers. However, it is challenging to find a suitable alternative, especially since users will probably not want to use an algorithm that does not show them engaging content. The authors propose testing a specific method to optimize content, which might correct misperceptions, and decrease polarization, while still keeping users engaged. Since we have limited evidence on the effect of algorithms and have almost no evidence on alternatives for current algorithms, this is an important and novel research question. Furthermore, I was impressed by how thoroughly the authors thought about how they will design the algorithms.
Despite this potential contribution, I have two main concerns regarding the outcomes and interventions.

1. Outcomes: I am not sure how much we will learn from the current specific outcomes.
a. R1: The first set of outcomes tests what users are exposed to in their feed. This outcome is not related to users' behavior or their ideology. Instead, it is only measuring the results of the algorithm. In essence, the authors are testing whether the algorithm they designed successfully decreases the amount of intergroup, moralized, and emotional (IME) content. To me, this sounds almost like a manipulation check, not a primary outcome. If no effect is found, then the "representative diversification" should probably just be adjusted, and the experiment should be run with an alternative algorithm. The authors also compare the reverse-chronological feed to the engagement-based feed, but that comparison has already been made in the FB 2020 papers for large platforms and I am not sure that effect is contested.
b. R2: The set of outcomes focusing on perceptions is interesting. However, I am concerned with the first question measuring perceptions. The authors suggest asking participants what the base rate of IME content in their social networks is. I am not sure how the "base rate" is defined. However, if users see more IME content, it seems sensible that they would say that it is more common in their networks, in the sense that more people observe it. Perhaps the authors are referring to the base rate of content shared or produced (regardless of views), but I am not sure if that is a relevant metric, and in any case, it should be explicitly stated.
c. R2: I am interested in the outcome regarding the perception of polarizing political dialogue! I do not know if this outcome will be affected or not and I agree with the author's explanation that it is important to find out since these perceptions can eventually affect opinions and behavior.
d. RQ3 H1/H2: The next outcome focuses on engagement with content. Here the effect is no longer just the result of the algorithm as users decide which content to engage with. However, as long as users are affected by the algorithm, I still expect an effect to be found here. Based on previous studies, users respond to the algorithm and engage with the content that appears on their feeds. I am not sure how much it would advance the literature if we find that when the feed has less toxic content, users engage with emotional content less often. Relatedly, I recommend that the authors cite the working paper "Toxic content and user engagement on social media: Evidence from a field experiment" (Beknazar-Yuzbashev et al.) which also changes the amount of toxicity in the feed.
e. RQ H3-H4: Finally, partisan animosity is clearly important but based on previous studies (e.g., Guess et al. 2023), I am worried that a sample size of ~530 participants per treatment is not sufficient to find an effect.

2. Intervention – the authors are creating a "representative diversification" feed. However, I am not sure if this feed will be truly representative. It will show more posts from the outgroup, fewer posts that are toxic and from super sharers, and more posts that are constructive, will prioritizing engagement. I find this feed interesting, as it may balance the need for engagement while reducing potential harm. However, it is not clear that it achieves the authors' goals of exposing users to representative posts and correcting their misperceptions. If that is the goal, the authors could simply select random ingroup and outgroup posts and then they would be representative.

3. Specification
a. I am concerned that RQ3 H3-H4 condition on a post-treatment variable. This could be a "bad control" that would bias the results. Covariates that could be affected by the treatment should not be used as moderators. Estimating the effect of norms on partisan animosity directly is also problematic since norms are not randomly assigned. I think the authors should estimate the effect of the treatment (the feed) on partisan animosity. If an effect is found, they can make some assumptions to try to analyze whether the feed affects animosity through the social norms channel.
b. If I understand correctly the authors are pooling data from several surveys. That's a clever way to increase power. I think the authors should state that they will cluster standard errors at the participant level and take that into account in their power calculations.

4. Minor – what is the size of the baseline pool of content? This could matter for how much variation the treatment generates. For example, if the baseline pool includes 100 posts and the three treatments are reranking the order of these posts, then a user who views all 100 posts in a session will only see them in a different order but still be exposed to the same content across the treatment arms.

(Remarks on code availability)

Referee #3

(Remarks to the Author)
Summary

This registered report presents the design of a field experiment that seeks to understand the causal effects of social media

feed-ranking algorithms on social norm perceptions and partisan animosity. The main research questions are: 1) whether engagement-optimizing social media algorithms amplify intergroup, moralized, and emotional (IME) content, 2) whether the amplification of IME content causes people to misperceive social norms around political dialogue on social media, and 3) whether misperceptions of social norms caused by algorithmic amplification of IME impact people's social media behaviors and political attitudes.

The experimental design, to the best of my understanding, is as follows.

Sample: 1600 US citizens who use social media moderately and identify as Democrat or Republican.
Treatment: Random assignment to three ranking algorithms: Reverse-chronological (Control), Engagement-based (Treatment 1), Representative Diversification (Treatment 2). Treatment 2 is an algorithm that seeks to maintain a balance of partisan content (at most 75% ingroup-leaning and at least 25% out-group leaning), while downranking toxic content and content produced by very frequent posters.
Outcomes: The main outcomes of interest are: 1) measures of the amount of IME content displayed to users, 2) measures of perceptions of IME content in Bluesky, norms of polarizing political dialogue on Bluesky, and norms of polarizing political dialogue in the U.S., 3) measures of partisan animosity, and 4) measures of engagement on Bluesky.

Before giving my detailed comments, I want to note that the idea of using Bluesky to experimentally vary the users' exposure to algorithms is very creative. Moreover, the experiment is very ambitious and clearly will require a great deal of effort and resources. However, I have concerns regarding 1) the extent to which the proposed design can answer the research questions and 2) the internal validity of the experiment.

Main comments

In terms of the significance of the research questions for the field of study, it is of first order importance to understand the impact that the exposure to different types of content has on social norms. As the authors mention, some of the effects that the literature has documented of algorithms on individual behavior could be driven by individuals updating their beliefs about social norms. Hence, the research question directly addresses a potentially important mechanism. One thing that was not clear to me, however, was the relevance of conducting the experiment during the election (or, why is the national election mentioned in the title?). Is it because we expect users to behave differently during the election? Or because we expect the algorithms to play a bigger role at influencing the content that users see? Does it help with making the experiment more natural?

One of my main concerns regards the extent to which the proposed study can satisfactorily answer the research questions. The proposed analysis cannot really pin down the effect of algorithms on the misperception about social norms for three reasons. First, because it is unclear whether the proposed measures of perceived social norms capture the respondents' own acceptability of a certain behavior, or their belief about others' acceptability (what the literature typically considers to be the social norm). For instance, when a user sees the question: "How socially appropriate is it to post content that is likely to create conflict among political groups on social media," do they reply how acceptable it is for them, or do they reply their belief about how acceptable it is for others?

Second, given this issue, the design does not elicit the objective "ground truth" that is required to measure misperceptions (which I assume are defined as the difference between the beliefs about the social acceptability relative to the objective or "ground truth" acceptability). Suppose that the authors find that the engagement-based algorithm increases users' beliefs about the acceptability of posting IME content to 70/100 (vs. say, a level of 60 in the control group and 65 in T2). How do we know that this increase is not actually closing the gap (if the objective acceptability is 85?) In that case, the control group is starting from a misperceived level (potentially because these are active social media users who could be on an echo chamber), and T1 is actually reducing the misperception (and T2 causes a lower reduction in the misperception). To address this issue, the design would need to measure two objects (see Bursztyn et al., AER2023 "Misperceived social norms…"): the objective acceptability based on individual-level responses (i.e., based on the subjects' own acceptability) and the subjects' beliefs about others' acceptability.

Third, even if the experimental designed measures the pre-intervention objective/ground truth social acceptability, the intervention could affect not just the users' belief about this social acceptability, but also the objective social acceptability itself. In particular, if we define the misperception as the difference: belief about social acceptability – objective social acceptability, it is clear that the intervention could affect both components of this difference. For example, an increase in the exposure to toxicity could make people more toxic, which would increase the objective social acceptability of toxicity. If users' beliefs about this social acceptability increase by the same amount, the actual misperception would not change, but 1) the current design that conflates misperceptions and perceptions (beliefs) would (incorrectly) conclude that misperceptions increased and 2) even a design that elicits the objective social acceptability but does not measure the change in this social acceptability would conclude that misperceptions increased. To address this issue, the design would need to measure the change in the objective acceptability after the intervention (vs. before). An alternative would be to reframe the research questions around the effect on the perception of social norms, rather than around the misperception of these norms. However, just looking at the effect on perceptions leaves many questions unanswered (e.g., why is there an effect on perceptions? Is it because people are changing their own opinions about acceptability or is it because they are updating their beliefs about others?).

Regarding the soundness and feasibility of the methodology and analysis pipeline, my main concerns are that 1) I don't think that the design incentivizes participants to actively use the platform, and 2) I don't think that the experiment is well powered

to detect the effect sizes that one would expect in this literature. In terms of 1), participants will be told that “they must log in to Bluesky at least twice per day, and engage in some way (post, like, share) at least once per day.” Can’t participants simply log in two times, like the first post they see, and then log off? It is notoriously difficult for platforms to gain new users, particularly given strong network effects, so I doubt that these incentives will be enough to convince them to actively use Bluesky. Better designs (of course, less feasible) would either recruit currently active Bluesky users or would pay participants based on their total activity (or active time spent) on the platform.

More importantly, in terms of 2), the effect sizes in the literature that studies the effects of algorithms on attitudes (in the field) are much smaller than the 0.10 SD that the experiment is powered to detect and, because of this, typically have sample sizes at least an order of magnitude larger than the proposal (or exploit more granular repeated-measure longitudinal designs). For example, the paper by Levy, AER2021 “Social Media, News Consumption, and Polarization” cited in the proposal (which is one of the papers with larger effect sizes), estimates an effect size on affective polarization of 0.03-0.06 SD, with a sample size of 17,635. Note that these concerns could be alleviated by providing pilot data, but the pilot data that the current proposal reports only speaks to the “first stage” effect of the algorithms on content, not on the “second stage” outcomes of interest. As it stands, my prior after reading the proposal is that the experiment will not be very informative about the research question. However, I do not want to discourage the authors from running this ambitious experiment, so I would recommend to maybe emphasize a bit more the more granular outcomes (for which they have even daily data, which allows them to exploit better the longitudinal variation) such as engagement on Bluesky or interactions with other users, which are also interesting.

Lastly, the degree of methodological detail provided seems sufficient to replicate exactly the proposed experimental procedures and analysis pipeline and offers minimal possibility to deviate from the proposed analysis.

(Remarks on code availability)
I couldn't install some of the packages required to run the code analysis_template.R (lme4 and in consequence lmerTest and simr), so I couldn't review it in depth.

Referee #4

(Remarks to the Author)
This field experimental design proposal aims to investigate whether and how different social media feed ranking algorithms influence how polarizing users’ news feeds are, as well as users’ opinions and beliefs over how polarized political dialogue is more generally. The authors plan to recruit study participants via the study participant provider Prolific. They plan to pay participants to open an anonymous account on the social media platform Bluesky that allows the researchers to choose the feed ranking algorithm, subsequently randomizing participants into experiencing different feed ranking algorithms, and surveying them regularly on their beliefs and attitudes, as well as observing the content users see in their feeds.

Comments on the proposal with regards to detail and replicability:

The proposal clearly stated the recruitment procedure, experimental protocol (including survey design, algorithm construction, and randomization), recruitment method, hypotheses to be tested, and outcome variables to be used. The proposal is sufficiently detailed to allow for an exact replication of the proposed experimental procedures and analysis pipeline after the experiment (since all code and data will be made publicly available). Furthermore, I found the description of the methods is sufficiently clear and detailed to prevent undisclosed flexibility in the experimental procedures or analysis pipeline. I believe that the authors have considered sufficient outcome-neutral conditions for ensuring that the results obtained are able to test the stated hypotheses.

Comments on the contribution, originality and significance

The question of the extent to which social media algorithms contribute to a more politically polarized climate is very significant to the field of political science and political economy, as well as to society as a whole (as the answer may have important consequences for platform regulation).

There are already several studies on this topic, including Beknazar-Yuzbasheve et al. (2023), Jimenez-Duran et al. (2024), Levy (2021), as well as the academic-industry collaboration papers on Facebook that were published in Science last year (most relevant Guess et al. 2023).

Different from the existing field experiments, the proposal at hand plans to recruits study participants to open new, anonymous accounts on a social media platform that users are not familiar with yet (instead of running the field experiment with pre-existing users on Facebook/YouTube/Twitter); furthermore, different from existing field experiments, the proposal at hand will not compare alternative feed algorithms (such as reverse-chronological, or those weeding out toxic posts – both of which have been tested/evaluated in the prior work cited above, as well a “representativeness” algorithm that is supposed to how more even-handed content from across the politica sprectrum – closely related to the field study by Levy, 2021) against the “status quo” of the existing platform, but instead it will compare against an engagement-maximizing algorithm that the research team themselves will devise. It was unclear to me whether this engagement-maximizing algorithm is aimed at emulating as close as possible the algorithms used by existing platforms such as Facebook or Instagram, or whether it is supposed to capture engagement-maximization period (with no regard to whether it does so similarly to existing platforms’

algorithms).

The final difference to existing work is the focus on mechanisms, in particular, the proposal aims to investigate the impacts of feed ranking algorithms on second order believes – i.e. how the feed ranking algorithms influence what respondents believe other people believe.

As stated above, that there is already existing field experimental work that tests reverse-chronological feeds (Guess et al. 2023), feeds that down-rank toxic content (Beknazar-Yuzbashev, 2023), as well as feeds that promote content from across the political spectrum (Levy, 2021) against the status quo algorithms used by large platforms such as Facebook and YouTube. In my view, the proposal at hand stays rather close to the existing work in the treatment arms it aims to test. The investigation into the role of second order believes, while interesting, in my view provides a neat, but not ground-breaking contribution, in part because the study setting has lower external validity than the "native social media user" settings.

Methodological comments

By randomizing participants into different social media feeds generated by the research team in a tightly controlled setting, the authors can cleanly isolate the causal impact of these different feeds. So, the internal validity is very high. My concerns are more with the external validity:

The authors plan to run the field experiment on a small, not widely used social media platform to which they aim to recruit new members via Prolific, a provider of research participants for social science studies. However, the "native" social media environment differs from the one proposed in this field experiment in ways that render it difficult to generalize findings form the experiment to social media settings "in the wild": for example, the participants would use anonymized usernames, thereby turning off key social image and reputational concerns in their sharing behavior. Furthermore, the social media experience on the platform used for the experiment would not include the participants' social network (friends, family, colleagues, etc.), who are a key component of the ordinary social media experience. In sum, this approach to direct study participants to a new social network comes at the big cost of artificiality and limited generalizability to the more "natural" social media settings.

Relatedly, in order to induce participants to use the platform, the experimental protocol states that participants would be required to log into the social media platform twice a day, including sharing a post every day. I am worried that this is not enough to induce participants to use the platform. A suggestion to strengthen external validity and to ensure higher use of the platform, could be to look into recruiting pre-existing Bluesky users directly.

References

Beknazar-Yuzbashev et al. (2023): Toxic Content and User Engagement on Social Media: Evidence from a Field Experiment. Working Paper.

Guess et al. (2023): How do social media feed algorithms affect attitudes and behavior in an election campaign? Science, Vol. 381.

Jimenez-Duran et al. (2024): The effect of content moderation on online and offline hate: Evidence from Germany's NetzDG. Working Paper.

(Remarks on code availability)
not applicable

Version 2:

Reviewer comments:

Referee #1

(Remarks to the Author)
I found the authors' responses to be very reasonable. They satisfactorily answered my concerns about the representative algorithm's content curation in a way that is logically coherent and more generalizable. My only remaining concern is that the analyses declared in the paper do not look at the interplay between the role of perceived norms and downstream behavior. Perhaps I am missing that from not being fluent with the R code, but, if not, it would be good to look at relationships between IME content posted, norms, and behavior rather than the effects of the experiment on each of these pieces alone.

(Remarks on code availability)

Referee #2

(Remarks to the Author)

This pre-registered report improved substantially from the previous version. First, the authors will now recruit active Bluesky users to participate in the experiment. This improves the external validity of the research since these participants use the platform. This strategy also decreases the concern that participants will not engage with the platform during the treatment period. While external validity is still an issue since Bluesky is not as important a platform as Facebook, TikTok, Instagram, etc., this change is a big improvement. Second, the authors update their outcome variables. Most importantly, they replaced partisan animosity, which was underpowered, with perceived polarization, where they can make a bigger contribution. The authors also modified the "representative" algorithm.

I have several remaining comments:
1. I still do not understand what RQ1, H2 is testing. The authors design an algorithm that reduces toxicity. Wouldn't such an algorithm almost mechanically decrease the amount of IME content displayed to the user (I understand that IME and toxicity are not the same thing but the concepts are related)? It sounds like RQ1 is testing what the authors' algorithm shows. While this is important information that should be displayed in the paper in order to interpret the rest of the results, I do not think it is testing a theoretically important question.

2. The authors explain that the base rate of IME content is the "total proportion of IME content present in a sample of content collected from the Bluesky firehose (all messages posted to Bluesky after filtering) in a given time period". This is a great definition. It is based on objective data and takes into account all posts, regardless of the algorithm. It is also likely that people may suffer from misperceptions regarding this measure. They may see more emotional posts because an algorithm maximizes engagement, and thus think that posting such content is more common than it actually is. However, if I understand correctly this is not what the authors ask about in the survey. They ask: "What percentage of posts in your Bluesky social network are emotional in nature"
a. It is not clear if this question refers to the posts people see in their feeds or the posts created. If people interpret it as the posts they see in their feed, then it will be mechanically affected by the treatment.
b. This question asks about one's social network but not about Bluesky generally.
I think this question should be reworded so it is clear to the respondents what the authors are referring to and in a way that is consistent with their definition of the base rate of content.

3. I am still not sure why the "representative diversification" algorithm is called representative as it does not aim to target representativeness, instead, it aims to change the distribution of three dimensions of posts: number of posts by super posters, number of toxic posts, and number of constructive posts. The authors nicely show that the algorithm becomes more representative on one dimension by decreasing toxicity, but there are many other dimensions of representativeness. For example, the posts in the representative algorithm are probably not representative in terms of the topic discussed, how extreme they are, the left-right balance, and so forth. Based on Table S4 the representative algorithm is not even substantially more representative in terms of constructiveness (it shows ~39% more constructive posts than the baseline rate, while the engagement-based algorithm shows ~40 fewer posts than the baseline rate). As I wrote in my previous review, I understand the logic of using this algorithm. It neatly balances engagement with less polarizing content, and thus, it is also policy-relevant. Therefore, I do not think the authors necessarily have to change the algorithm, but I do think they should rename it so the readers can more easily understand what exactly is being tested. I actually like the name proposed by R3: "depolarizing algorithm", but that is just one suggestion.

4. I am worried that the authors will not be able to recruit enough participants. For example, the authors mention that Reddit has 15k active BlueSky users. If the authors design great ads, maybe 3% of those users will click the ad, 2% will meet the screening criteria, and 1% (150 users) will complete all surveys. Of course, this is just a back-of-the-envelope calculation. I do not think this is a reason to reject the design and the authors do not need to respond to this concern. I am only making this point to make sure the authors are aware of the recruitment risks (I am sure they already are).

(Remarks on code availability)

Referee #3

(Remarks to the Author)
I will keep my comments short; the authors clearly worked hard on my suggestions (and I believe that also those of the editor and the other referees) so I don't have much more to add.

The first remaining suggestion I have is to provide participants with a definition of what "toxic" means in the norm perception outcomes (2), (3a), (3b). You could potentially use the definition of Perspective API.

The second remaining suggestion is to add a test of differential attrition in the paper (and, ideally, a way to address differential attrition in case it occurs) and measures of attrition over time.

(Remarks on code availability)

Referee #4

(Remarks to the Author)
I had two main concerns with the previous version of the proposed study:

1. External validity
2. Contribution relative to the existing literature

Regarding 1.:
This concern is entirely resolved, since the authors now plan to recruit active Bluesky users. Now that the authors do recruit active Bluesky users, they could think about including a control arm that does business as usual – though I don't think this is critical.

It is a lot more difficult to recruit active users than to rely on simple Prolific convenience samples of non-users. Therefore, it was reassuring to read that the authors are aware of this issue and have a backup plan to screen Prolific participants based on their use of Bluesky.

Regarding 2.:

Relying on native users, thereby elevating external validity, also improves my perception of the contribution relative to the literature.

However, in the new version, the authors put much more weight on "soft" outcomes related to perceptions about others' beliefs and behaviors. These outcomes are "lower hanging fruit", because they are easy to measure and likely easier to move (hence the experiment can be done with a smaller sample), than the ultimately arguably more relevant outcomes of voting behavior, partisan attitudes, and behavior on the platform itself (i.e. what people post). They could also be more strongly driven by experimenter demand effects. So, I am not 100% on board with down-weighing these other outcomes. I have two considerations here: 1) So long as the other outcomes are included too, albeit not as primary outcomes, I suppose readers like me can scroll down to the relevant appendix section. 2) To deal with experimenter demand effects better: a) keep the treatment arm assignment as opaque as possible, i.e. don't be explicit in disclosing how exactly users' feeds will change; b) complement the outcomes with "harder" measures collected outside of the surveys: e.g., publish posts to users' feeds and measure whether they click on them/engage with them; publish poll questions to users' feeds over the course of the study period; send "friend" requests to participants, varying the party affiliation of the requestor.

(Remarks on code availability)

Version 3:

Reviewer comments:

Referee #1

(Remarks to the Author)
I appreciate the authors' response and agree that the proposed exploratory analyses are appropriate. I have not, however, found them in the manuscript or supplemental materials. Once they are included, I am happy to sign off on this registered report.

(Remarks on code availability)

Referee #2

(Remarks to the Author)
The authors made several additional improvements to this paper - they renamed the "representative" algorithm and reworded the perception questions to better capture the object they care about. They also detailed an impressive plan to recruit participants.

I appreciate their hard work revising the experiment plan and look forward to seeing their results!

(Remarks on code availability)

Referee #3

(Remarks to the Author)
The authors addressed my comments. I would only suggest to report Lee Bounds in the case of differential attrition, even if these bounds are likely to be wide. Otherwise, I am happy to sign off given the time-sensitive nature of the study.

(Remarks on code availability)

Referee #4

(Remarks to the Author)
All my concerns have been addressed and I have nothing to add other than to wish best of luck with the study. I am looking forward to reading the results.

(Remarks on code availability)

Referee #5

(Remarks to the Author)

(Remarks on code availability)

STAGE 2

Version 5:

Reviewer comments:

Referee #2

(Remarks to the Author)
This paper analyzes the effect of different Bluesky algorithms on the posts users see in their feeds, their perceptions, and the posts they engage with. Overall, the authors find that a reverse-chronological feed decreases exposure to moral, emotional, toxic, and political content, compared to an engagement-based feed. The diversified extremity algorithm they created mostly had the opposite effects. They also find that reverse-chronological feeds decrease perceptions of partisan animosity and users' own feelings of animosity (the effects on other perceptions were less clear). Finally, they do not find an effect on users' own engagement behavior.

Overall, the paper is executed well, mostly follows the pre-analysis plan, and makes an important contribution to the literature! Since this paper was accepted based on the plan, I will focus on the results, the exploratory analyses, the changes, and the writing, and less on the importance or broader contribution of the paper.

I. Changes from pre-registration
1. The authors exclude users who did not complete enough surveys or did not log in to Bluesky often enough. I do not think this was pre-registered, and it is important. Excluding users could lead to differential attrition. The authors argue in Table S38 that there is no differential attrition, but I am not sure that is the case. There is more attrition in the reverse-chronological condition compared to the two treatments that promote posts users are likely to engage with. This pattern is consistent with Guess et al. (2023a), who find that people spend dramatically less time on Facebook and Instagram when assigned to the reverse-chronological feed. Although the differential attrition is not large and likely not driving the results, the authors should check more closely. Specifically, they could relax the requirement of engaging with the platform for 5 of 8 weeks and examine how that affects the results.

2. Other changes in the plan—such as analyzing binary outcomes in RQ3—made sense to me. I also found the exploratory analyses sensible and the following particularly interesting
• Analyzing the effect on moral outrage content and constructive content (RQ1, H2).
• Analyzing the effect on enjoyment with the feed (RQ2, H1).
• Studying whether algorithmic amplification affects political polarization.
• Analyzing how skewed engagement with toxic political content is.

II. Comments on the analysis
3. I did not completely understand the paragraph arguing that adjusting for multiple hypotheses is not needed because the authors pre-registered directional hypotheses.
I do not see why directional predictions avoid the multiple-tests problem. However, this paragraph did appear in the pre-registered version, and I probably should have mentioned this then (the number of tests did increase substantially because of the new election-interaction terms).
More importantly, the paper does not seem to be analyzing one-sided tests. For example, it states that before the election, engagement-based feeds served less constructive content than reverse-chronological feeds and after the election, engagement-based feeds served more constructive content. The authors treat both directions as evidence of effects rather than simply confirming or rejecting a one-sided hypothesis. In my opinion, this is the correct approach; however, it is not clear that they can argue they are testing only one direction in their hypotheses.

4. I found the analysis of descriptive norms variables in RQ2 confusing, and I only understood it better after reading the discussion at the end of the paper.
It is not always fully clear what the outcome variable is in the text. The pre-registration discusses perceptions of IME and toxic content, but the text mostly discusses accuracy. I would analyze the effect on perceptions and the effect on accuracy as two separate outcomes and clearly distinguish between them. This analysis is provided in Table S59, but that table is hard to interpret (as I discuss below), and the main text does not discuss its results.
More importantly, I suggest replacing Figures 2–3 with a figure similar to Figure 1 (which is excellent), so the treatment effects of each treatment on each outcome are crystal clear. Of course, there could be other solutions, but currently this section should be rewritten to be clearer and more transparent.

5. The authors find some null effects on behavior and then include an interesting discussion at the end of the paper regarding these effects. I think it is important to discuss the precision of these effects. In other words, what effects can the authors rule out? For example, they could specify that they rule out a decrease in the likelihood of engaging with toxic content greater than X%. This will help the readers understand whether they do not find that the change in exposure did not map to a change in engagement because of selective behavior among Bluesky users or due to limited statistical power.

III. Comments on writing
6. Many appendix tables are hard to interpret because the reader must add coefficients and cannot easily tell, for example, whether each treatment is significant after the election. In my opinion, all of these tables should convey the same information that is conveyed in Table 1. They should show four numbers for each outcome: the effect (and standard error or confidence interval) of the two treatments compared to a reference group (reverse chronological in Figure 1, though engagement-based could also be the reference), before and after the election. The table could also note whether the difference between the two treatments is significant.

7. The discussion is well written and interesting. I also agree that this paper makes an important methodological contribution, as the authors state: “Our study also represents an example of an industry-independent model for studying algorithmic amplification at scale.”

8. The effect on constructive content is not clear. The authors first say “By contrast, diversified extremity feeds served less constructive content than engagement-based feeds both before and after the election, with a similar magnitude of difference across periods,” but later say “In contrast, our diversified extremity algorithm ... while preserving or even enhancing exposure to more positive and constructive dialogue.” Is the effect on constructive content positive or negative? Based on Figure 1, I think the former sentence is correct and the latter should be reworded.
More generally, it is surprising that diversified extremity feeds served less constructive content, since the algorithm was designed to promote exactly that content. Can the authors explain this result?

9. Literature
a. In the introduction, the authors write, regarding Guess et al. (2023a): “This study is foundational to the literature, as it was the first experimental manipulation of algorithmic amplification conducted with a broad swathe of the American populace on a real social media platform.” I am not sure this is accurate; for example, see Huszár et al. (2021).
b. They also say, “First, the researchers did not control how the content algorithms worked, making it unclear which specific features of the algorithms created content differences between conditions.” Two other experiments run at the same time did change features in the algorithm (Nyhan et al., 2023; Guess et al., 2023b). The text regarding these two commented appeared in the previous pre-registered version, and I apologize for only noticing and mentioning it now.
c. Finding that reducing negative content does not come at the expense of user satisfaction is interesting! To some extent, consistent with Braghieri et al. (2025).
d. In the discussion, the authors write: “At face value, these findings could be interpreted as evidence of limited downstream behavioral consequences of algorithmic amplification, consistent with recent experimental work suggesting limited effects of algorithmic feeds on user political attitudes and behaviors.” They cite Guess et al. (2023a), but that paper did find an effect on on-platform measures of political engagement. The broader point is that previous literature has found that consumers tend to click what appears in their feeds. It is completely fine that this paper did not find such an effect, and the authors discuss potential reasons in the discussion.

References

• Braghieri et al. (2025) – Frictions in News Consumption: Evidence from Social Media
• Guess et al. (2023a) – How do social media feed algorithms affect attitudes and behavior in an election campaign?
• Guess et al. (2023b) – Reshares on social media amplify political news but do not detectably affect beliefs or opinions
• Huszár et al. (2021) – Algorithmic amplification of politics on Twitter
• Nyhan et al. (2023) – Like-minded sources on Facebook are prevalent but not polarizing

(Remarks on code availability)

Referee #3

(Remarks to the Author)
Summary:

This paper presents findings from a field experiment that randomizes social media feed-ranking algorithms among U.S. Bluesky users during the 2024 presidential election. Participants were assigned to either a reverse-chronological feed, an engagement-based feed (similar to those used by major platforms), or a “diversified-extremity” feed that downweighted content from extreme, toxic, or overly active users. The main results are that engagement-based algorithms amplified intergroup, moralized, emotional, toxic, and political content relative to a chronological feed but that this amplification had limited effects on users’ engagement behaviors and somewhat mixed effects on perceived norms.

Evaluation:

I enjoyed reading the manuscript and this is a very impressive experiment. Below I give some comments that I think will help it improve.

- I would like to see more discussion about how the authors think of the election as a shock. Currently, they mention that “The preregistration did not anticipate modeling the election as an
exogenous shock. Because the election was likely to affect our politically relevant outcomes […]” but I didn’t understand 1) what is the election shocking (a shock to what)?, 2) why this shock was supposed to affect outcomes, and 3) why this shock would affect outcomes differentially across experimental conditions (to justify adding this non-preregistered dimension of heterogeneity).

- Relatedly, the paper sometimes presents results before/after the election without offering any explanation for why this is the case. For example, “Intergroup content moved in the opposite direction before the election (diversified extremity served more than engagement-based) and showed no difference after the election.” After reading this type of sentences, the reader is left wondering what drives these differences.

- Also relatedly, I didn’t understand the logic of the following rule: “if the majority of outcomes showed better fit (AIC/BIC criteria and Log-likelihood ratio test) when including a binary pre-/post-election indicator and its interaction with condition, then we report that specification in the main text”. Concretely, I don’t see why not just report the pre-registered specification in the main text and relegate to the appendix 1) specifications adding controls for before/after the election, and 2) heterogeneity analysis of the main treatment effects before and after the election.

- Other deviations from preregistration such as the different sampling scheme or time aggregation seem fine to me.

- Is there a single summary measure of IME content? I ask for three reasons. First, because the paper makes some statements that group IME as a single category (e.g., “these data support the idea that engagement-based algorithms specifically amplify IME content”). Second, because the prevalence of some forms of IME content is so small that it would be good to have a single summary statistic to evaluate the strength of the intervention (e.g., the diversified extremity algorithm reduced IME content by X standard deviations relative to the engagement-based algorithm). This is especially important since some of the null effects of RQ2 and RQ3 could be explained by a weak intervention. Third, because some forms of IME go in opposite directions (such as positive/negative emotional), so it would be interesting to have a discussion of how to interpret these changes.

- Relatedly, I would like to see more heterogeneity analysis of the social norm perception outcomes by the amount of IME content removed.

- I would like to see more information about how the intervention impacted user engagement (time spent on the platform, sessions, content consumed).

- Lastly, I find the following sentence inaccurate: "This study is foundational to the literature, as it was the first experimental manipulation of algorithmic amplification conducted with a broad swathe of the American populace on a real social media platform." There are several studies that conduct experimental manipulations of this type; for example, some of the Facebook/Instagram 2020 Election studies.

(Remarks on code availability)
Partly reviewed some of the analysis code (but did not install or run it).

Version 6:

Reviewer comments:

Referee #2

(Remarks to the Author)
This paper has improved substantially since the previous version. I look forward to seeing it published!

(Remarks on code availability)

Referee #3

(Remarks to the Author)
I read the responses to R2 and to my previous comments and I have nothing more to add -- I think the authors replied well to most of my outstanding concerns.

(Remarks on code availability)

Northwestern
Kellogg

School of Management
Northwestern University
2211 Campus Drive
Evanston, IL, 60208

Responses to Reviewer #1 comments

R1.1 Clarify details of the 'representative diversification' algorithm

Reviewer #1 questions whether the representative diversification feed is more focused on depolarization than representativeness. They write: *"I like the design. I find random assignment to one of three of custom ranking algorithms on the Bluesky platform to be an excellent method with which to manipulate the influence of social media algorithms. The purpose and value of the engagement based and reverse chronological algorithms is good and clear. I get the value of the "representative" feed but it might be better characterized as the feed intended to be depolarizing. It's hard to say that it's designed to be fully representative. One possibility is to admit that more plainly and move forward. Another would be to train it to more randomly sample from "civics" posts that appear on Bluesky (with the current superposter and toxicity filters) rather than constrain to a 75%/25% ingroup vs. outgroup split. It could be the case, which would be interesting, that the truly representative posting would be the most polarizing feed of the three. The 2018 PNAS cited explored the effect of one twitter bot on polarization metrics so this does seem like it could add quite a lot to the literature."*

We agree with the points raised by Reviewer #1, especially their recommendation to avoid constraining to a 75/25 ingroup vs. outgroup split. Reviewer #1's comments have also led us to reflect on what it means for a feed to be "representative." In one formulation, a feed can be representative of what is posted in aggregate on BlueSky, which means that it will still contain more posts from superposters. Participants in the reverse-chronological condition will see something similar to this sort of feed. In a second formulation, the feed can be representative of the platform's users (rather than the posts in aggregate). This algorithm will ensure similar numbers of posts from superposters and the less active users in order to ensure the more ideologically extreme posts of superposters are not overrepresented in people's feeds (as our theory and a growing body of evidence indicates that superposters' content is more ideologically extreme than the average user). Our intervention algorithm aims to mirror this latter form of representativeness.

We also highlight that the representative diversification algorithm applies its ranking on top of the engagement-based model, which possesses an important strength: The representative diversification algorithm should keep the feed as interesting as the engagement-based feed by rebalancing engaging content that is less toxic and more positive. In contrast, a random sample of content would be unlikely to draw as much engagement. Creating a highly engaging, representatively diverse feed is also important because it greatly increases the chances that social media platforms—who, in practice, tend to prioritize user engagement on their platforms for obvious reasons—would consider implementing the algorithm.

On pg. 8 of the revised manuscript, we updated the summary of conditions:

Ranking Algorithm	Description
Reverse-chronological (Control)	Ranks content that was posted most recently as the highest on participants' feeds
Engagement-based (Treatment 1)	Ranks content based on the probability that participants will engage with it in the future, estimated based on both a post's similarity to previous content with which a participant engaged and similarity to content that has high engagement in the broader community.
Representative Diversification (Treatment 2)	For content classified as political, the following ranking-criteria are applied, in order of descending priority, on top of the outcome of the engagement-based algorithm (high probability of **engagement in the future**). 1. Low probability of **super poster** 2. Low probability of **toxicity** 3. High probability of **constructiveness** For content not classified as political, it will rank content based on the probability of engagement in the future.

Table 3. Description of feed-ranking algorithms for each experimental condition. Participants are randomly assigned to one of three experimental conditions, where content they see on Bluesky Social is controlled by the respective feed-ranking algorithm.

In summary, our updated version of the representative diversification algorithm drops the 75/25% ingroup/outgroup filter. Rather, our current approach begins by drawing baseline posts from the output of the engagement-based model, and then applies three

ranking criteria, all three of which are meant to reduce the influence of the most extreme political users: downrank superposters, downrank toxicity, and uprank constructive dialogue. The combination of these properties underlies the representative diversification feed, which diversifies the political ideological distribution of a users' feed by making it more representative of the true distribution of political ideology of the platform, which is less extreme (political content in typical engagement-based feeds are dominated by the most politically extreme users and are more toxic and negative; Bail, 2021; Papakyriakopoulos et al., 2020; Smith, 2021).

Since the last submission, we also conducted a more extensive pilot that pooled a large baseline pool of posts from the Bluesky firehose (all the English posts to Bluesky in a one-week period). We then filtered these posts based on the criteria of our conditions to simulate our engagement-based and representative diversification conditions (see Table 3 above). In an important proof-of-concept, we confirmed that on Bluesky, the engagement-based feed increased toxicity by 48% *relative to the base rate* of toxicity on Bluesky (measured by the firehouse). 2.9% of content pulled from the firehose was classified as toxic compared to 4.3% of content in the simulated engagement-based feed. Importantly, our representative diversification feed did not simply remove toxicity, but made the levels more representative of the base rate at 2.8% of content. Full details of the new pilot are available in SI Appendix, Section 6.0 of the manuscript.

Below, we also paste the updated description of the representative diversification algorithm (Table 3 and pg. 9 of the manuscript):

> The 'representative diversification' algorithm (Treatment 2) is designed to diversify the participants' feed relative to engagement-based feeds by increasing the *social representativeness* of the feed, relative to the true distribution of ideological extremity present on the platform. Because it is intractable to estimate the true distribution of political extremity for all possible posts, we use three filters that should produce more representative posts, i.e., posts that do not represent the long tails of the distribution of political extremity. Specifically, for posts that are classified as political in nature (see SI appendix, 3.2.2), the algorithm (1) reduces probability of super posters, or those who post more than 5 times in a 24-hour window, (2) reduces probability of toxic posts, and (3) increases probability of constructive dialogue posts. We assume that these three filtering criteria in practice yield sampling that drops the long tails of the political post distribution because it is now well-established that the large majority of toxic political content is posted by extreme partisans who represent a small percentage of users[16–18]. In other words, long-tail users are the most extreme, the most toxic, and post the most content. These criteria expand upon previous work that simply removed toxic content in feeds, which was successful in reducing users' engagement with toxic content.[36]

Importantly, the representative diversification algorithm can be applied to any platform with varying underlying distributions of political ideology, whereas a bridging algorithm that aims to expose users to outgroup political posts (e.g. [20]) would be constrained to platforms that have enough outgroup content in the first place. The extent to which a user in the representative diversification feed is likely to see an "outgroup" post depends on the true distribution of users on the platform. For instance, many platforms including Bluesky tend to be very left-leaning, but the representative diversification algorithm will still diversify the feeds in terms of ideological extremity of left-leaning users. In fact, correcting perceptions of ingroup extremity may be more effective than exposing users to less extreme outgroup content. Exposing users to outgroup content has the possibility of backfiring and increasing outgroup animosity[37]. However, applied to a platform with a true mix of left-right ideology, the representative diversification algorithm would still limit extreme users appearing in posts from both sides of the political spectrum, which may ameliorate backfire effects of exposure to outgroup content.

Toxicity and constructive dialogue are classified using validated models built to classify social media data from Google's Perspective API [38]. Importantly, we demonstrate in pilot testing that filtering by removing toxicity and selecting for constructive dialogue still allows for strong moral opinions to be displayed. Rather than simply removing strong moral opinions, these filtering steps remove posts that contain strong moral opinions *and* have toxic properties such as personal attacks; in other words, toxic content is filtered independent of moral content (see SI Appendix, Section 3.0 for full technical details and validation of classifying toxicity and constructive dialogue). Furthermore, pilot testing described in SI Appendix, Section 6.0 suggests that these criteria do in fact reduce toxicity and increase constructiveness of dialogue relative to an engagement-based algorithm in simulated feeds tested on the Bluesky API firehose. Importantly, they also reduce toxicity and increase constructiveness in a direction that makes the composition of feeds more like the true base rate of English-language content as measured through Bluesky's firehose.

R1.2 Including perceptions of offline norms as an outcome

Reviewer #1 suggests that we add as measurement of perceptions of norms outside of Bluesky: *"More generally, I think the dependent variables make sense. My one critique would be that the social norm perceptions are largely focused on the platform and do less to tap into broader perceptions about what is normative political discourse."*

We agree that it is important to include perceptions of norms of political discourse outside of Bluesky. We actually include this as perceptions of partisan animosity in the U.S. Table 2 (pg. 22 of the manuscript) now indicates we will measure perceptions of polarizing discourse in the U.S. in the form of partisan animosity in the U.S., see also SI

Northwestern University
2211 Campus Drive
Evanston, IL, 60208

Appendix, Section 1.3. For convenience, we paste the relevant part of updated Table 2 and SI Appendix, Section 1.3 below:

Updated Table 2 section:

Intake survey	Experiment Phase Weeks 1-8
<u>Norm Perception DVs</u>	
(1) Perception of IME content frequency in participant's BlueSky social network (2) Perception of toxic political discourse in participant's Bluesky social network (descriptive) (3) Perception of toxic political discourse in participant's Bluesky social network (prescriptive) (4) Perception of ingroup's dislike of outgroup (perceived partisan animosity) (5) Perception of outgroup's dislike of ingroup (perceived partisan animosity)	(1) Perception of IME content frequency in participant's BlueSky social network (2) Perception of toxic political discourse in participant's Bluesky social network (descriptive) (3) Perception of toxic political discourse in participant's Bluesky social network (prescriptive) (4) Perception of ingroup's dislike of outgroup (perceived partisan animosity) (5) Perception of outgroup's dislike of ingroup (perceived partisan animosity)

Updated measures of perceptions of polarizing discourse in the U.S.:

(4) Perception of ingroup's dislike of outgroup (perceived partisan animosity) (0-100) (Brady et al., 2023)[1,2]

Please rate how your political group (Democrats[Republicans]) feels about Republicans [Democrats] on the scale below:

0	50	100
Strongly dislike	Completely neutral	Strongly like

(5) Perception of outgroup's dislike of ingroup (perceived partisan animosity) (0-100) (Brady et al., 2023)[1,2]

Please rate how (Republicans [Democrats]) feel about Democrats [Republicans] on the scale below:

0	50	100
Strongly dislike	Completely neutral	Strongly like

Responses to Reviewer #2 comments

R2.1 Concern over whether to include RQ1

Reviewer #2 is unsure whether the outcome tested in RQ1 (testing the difference in frequency of ingroup, moralized, and emotional (IME) content across each condition) is novel: "*I am not sure how much we will learn from the current specific outcomes. R1: The first set of outcomes tests what users are exposed to in their feed. This outcome is not related to users' behavior or their ideology. Instead, it is only measuring the results of the algorithm. In essence, the authors are testing whether the algorithm they designed successfully decreases the amount of intergroup, moralized, and emotional (IME) content. To me, this sounds almost like a manipulation check, not a primary outcome. If no effect is found, then the "representative diversification" should probably just be adjusted, and the experiment should be run with an alternative algorithm. The authors also compare the reverse-chronological feed to the engagement-based feed, but that comparison has already been made in the FB 2020 papers for large platforms and I am not sure that effect is contested.*

Although we believe Reviewer #2 raises valuable considerations, we respectfully disagree that engagement-based algorithms have been shown to increase IME content specifically in prior work (relative to reverse-chronological feeds), and therefore we believe that measuring IME content in feeds serves more than a manipulation check. The papers that Reviewer #2 cites have actually focused on more specific outcomes, namely: "toxicity", (Guess et al., 2023; Beknazar-Yuzbashev et al. 2023), "political content" (Guess et al., 2023) and "anger" (Milli et al., 2023). We view these dimensions as specific manifestations of IME content. The goal of our IME framework is to provide a

broader set of theory-driven content dimensions to organize these prior findings, and also to go beyond them. For example, our framework would suggest that moralized information may get promoted by typical content algorithms, even when it is not toxic or anger-driven. Shameful apologies or calls to collective action would be examples of this kind of content. To increase our understanding of the effects of engagement-based algorithms used on current social media platforms, we believe it is still worth measuring these specific outcomes, and to confirm whether our intervention reduces this content specifically. We also note that a recent review in *Nature* specifically called for more direct evidence of algorithmic amplification of "harmful" content (Budak et al., 2024), supporting the idea that measuring the specific content engagement-based algorithm amplify is important. *However, if the editor and reviewers disagree, we are willing to remove this outcome as a primary hypothesis (assessing it instead as an exploratory outcome)*.

R2.2 Clarifying measurement of perception of descriptive social norms

Reviewer #2 suggests clarifying how we define the base rate of IME content in order to measure descriptive social norms: *"The set of outcomes focusing on perceptions is interesting. However, I am concerned with the first question measuring perceptions. The authors suggest asking participants what the base rate of IME content in their social networks is. I am not sure how the "base rate" is defined. However, if users see more IME content, it seems sensible that they would say that it is more common in their networks, in the sense that more people observe it. Perhaps the authors are referring to the base rate of content shared or produced (regardless of views), but I am not sure if that is a relevant metric, and in any case, it should be explicitly stated."*

We apologize for not clearly defining base rate in the previous manuscript. We define base rate as the total proportion of IME content present in a sample of content collected from the Bluesky firehose (all messages posted to Bluesky after filtering) in a given time period. Filtering removes non-english messages, spam and NSFW content (which is also applied to all condition feeds). IME content is classified using our machine learning classification scheme developed and extensively validated for this project (described in SI Appendix, 3.3.3.

Thus, each week we can estimate the proportions of IME content posted across all of Bluesky as compared to perceptions participants form from viewing their feed determined by their experimental condition. Based on pilot testing, we estimate collecting ~7 million messages per week as a base rate comparison. This approach has the advantage of determining to what extent each condition feed creates perceptions

that depart from the actual base rate of content on Bluesky. The description is updated on pg. 10 of the manuscript:

> In order to collect the baseline pool of posts that will be filtered based on experimental condition, we pull from the Bluesky firehose (all posts in real-time), and immediately filter out posts that are non-English, spam, or contain vulgar or explicit content. We estimate that we will collect 60,000,000 posts throughout the 8-week experiment through the Bluesky firehose in our baseline pool of posts, based on our discussions with Bluesky developers in their Discord channel. Further filtering of the baseline posts are based on the criteria for each experimental condition as described above. Full technical details are presented in SI Appendix, Section 3.2. The Python code for our in-progress data pipeline, data storage, and browser extension is available here.
>
> The baseline pool of content also allows us to form base rate measurements of content that will be used to compare to participants' perceptions of content frequency in their feeds, see SI Appendix 1.3 for details on perceptions of content frequency, and SI Appendix 3.3.3 for details on classification of content.

R2.3 Agreement about perceptions of polarizing political dialogue

Reviewer #2 agrees that perceptions of polarizing political dialogue is an important outcome to measure: *"I am interested in the outcome regarding the perception of polarizing political dialogue! I do not know if this outcome will be affected or not and I agree with the author's explanation that it is important to find out since these perceptions can eventually affect opinions and behavior."*

We are glad Reviewer #2 finds this outcome interesting. Following their recommendation, we opted to make it the primary outcome of interest, along with perceived polarization, as described above in **E.1**.

R2.4 Clarifying user engagement outcome

Reviewer #2 wonders how novel it is to measure changes in user engagement-based feeds: *"The next outcome focuses on engagement with content. Here the effect is no longer just the result of the algorithm as users decide which content to engage with.*

However, as long as users are affected by the algorithm, I still expect an effect to be found here. Based on previous studies, users respond to the algorithm and engage with the content that appears on their feeds. I am not sure how much it would advance the literature if we find that when the feed has less toxic content, users engage with emotional content less often. Relatedly, I recommend that the authors cite the working paper "Toxic content and user engagement on social media: Evidence from a field experiment" (Beknazar-Yuzbashev et al.) which also changes the amount of toxicity in the feed."

We believe that measuring user behaviors in response to changes in feed content is important for two reasons. First, is it not a foregone conclusion that changes in feed content always lead to changes in user engagement, even though a few studies (which we now cite) are suggestive of this pattern. Even if a feed reduces emotional content, users may still engage with the emotional content more often when it does appear on the feed due to human attention biases. Furthermore, the question of whether a user actually *composes* less emotional content (rather than simply liking/sharing) as a result of viewing a feed with reduced emotional content is a direct measure of social learning which is the main focus of our revised manuscript.

Second, we also note that as discussed in **R2.1**, we are measuring more than engagement with posts and posting behaviors that contain toxicity. We are also measuring ingroup, moral, and emotional (IME) content more broadly, which we argued is theoretically important and has not been measured specifically in previous work. Nonetheless, we agree the Beknazar-Yuzbashev working paper is important and we cite it in our discussion of our representative diversification algorithm:

> The 'representative diversification' algorithm (Treatment 2) is designed to diversify the participants' feed relative to engagement-based feeds by increasing the *social representativeness* of the feed, relative to the true distribution of ideological extremity present on the platform. Because it is intractable to estimate the true distribution of political extremity for all possible posts, we use three filters that should produce more representative posts, i.e., posts that do not represent the long tails of the distribution of political extremity. Specifically, for posts that are classified as political in nature (see SI appendix, 3.3.2), the algorithm (1) reduces probability of super posters, or those who post more than 5 times in a 24-hour window, (2) reduces probability of toxic posts, and (3) increases probability of constructive dialogue posts. We assume that these three filtering criteria in practice yield sampling that drops the long tails of the political post distribution because it is now well-established that the large majority of toxic political content is posted by extreme partisans who represent a small percentage of users[16–18]. In other words, long-tail users are the most extreme,

the most toxic, and post the most content. These criteria expand upon previous work that simply removed toxic content in feeds, which was successful in reducing users' engagement with toxic content.[36]

R2.5 Questioning intervention effects on partisan animosity

Reviewer #2 is concerned that our intervention may not reduce partisan animosity in a detectable way: *"Finally, partisan animosity is clearly important but based on previous studies (e.g., Guess et al. 2023), I am worried that a sample size of ~530 participants per treatment is not sufficient to find an effect."*

As discussed in **E.1**, we removed partisan animosity as a primary outcome of interest and now focus on social norm perceptions and perceived polarization as suggested by Reviewer #2. However, we still measure partisan animosity as an exploratory outcome.

R2.6 Clarifying details about the intervention

Reviewer #2 questions whether the representative diversification intervention makes feeds more representative: *"Intervention – the authors are creating a "representative diversification" feed. However, I am not sure if this feed will be truly representative. It will show more posts from the outgroup, fewer posts that are toxic and from super sharers, and more posts that are constructive, will prioritizing engagement. I find this feed interesting, as it may balance the need for engagement while reducing potential harm. However, it is not clear that it achieves the authors' goals of exposing users to representative posts and correcting their misperceptions. If that is the goal, the authors could simply select random ingroup and outgroup posts and then they would be representative."*

Reviewer #1 and Reviewer #2 led us to reflect on what it means for a feed to be "representative." In one formulation, a feed can be representative of what is posted in aggregate on BlueSky, which means that it will still contain more posts from superposters. Participants in the reverse-chronological condition will see something similar to this sort of feed. In a second formulation, the feed can be representative of the platform's users (rather than the posts in aggregate). This algorithm will ensure similar numbers of posts from superposters and the less active users in order to ensure the more ideologically extreme posts of superposters are not overrepresented in people's feeds (as our theory and a growing body of evidence indicates that superposters' content is more ideologically extreme than the average user). Our intervention algorithm aims to mirror this latter form of representativeness.

We also highlight that the representative diversification algorithm applies its ranking on top of the engagement-based model, which possesses an important strength: The representative diversification algorithm should keep the feed as interesting as the engagement-based feed by rebalancing engaging content that is less toxic and more positive. In contrast, a random sample of content would be unlikely to draw as much engagement. Creating a highly engaging, representatively diverse feed is also important because it greatly increases the chances that social media platforms—who, in practice, tend to prioritize user engagement on their platforms for obvious reasons—would consider implementing the algorithm.

On pg. 8 of the revised manuscript, we updated the summary of conditions:

Ranking Algorithm	Description
Reverse-chronological (Control)	Ranks content that was posted most recently as the highest on participants' feeds
Engagement-based (Treatment 1)	Ranks content based on the probability that participants will engage with it in the future, estimated based on both a post's similarity to previous content with which a participant engaged and similarity to content that has high engagement in the broader community.
Representative Diversification (Treatment 2)	For content classified as political, the following ranking-criteria are applied, in order of descending priority, on top of the outcome of the engagement-based algorithm (high probability of **engagement in the future**). 1. Low probability of **super poster** 2. Low probability of **toxicity** 3. High probability of **constructiveness** For content not classified as political, it will rank content based on the probability of engagement in the future.

Table 3. Description of feed-ranking algorithms for each experimental condition. Participants are randomly assigned to one of three experimental conditions, where content they see on Bluesky Social is controlled by the respective feed-ranking algorithm.

In summary, our updated version of the representative diversification algorithm drops the 75/25% ingroup/outgroup filter. Rather, our current approach begins by drawing baseline posts from the output of the engagement-based model, and then applies three ranking criteria, all three of which are meant to reduce the influence of the most extreme

political users: downrank superposters, downrank toxicity, and uprank constructive dialogue. The combination of these properties underlies the representative diversification feed, which diversifies the political ideological distribution of a users' feed by making it more representative of the true distribution of political ideology of the platform, which is less extreme (political content in typical engagement-based feeds are dominated by the most politically extreme users and are more toxic and negative; Bail, 2021; Papakyriakopoulos et al., 2020; Smith, 2021).

Since the last submission, we also conducted a more extensive pilot that pooled a large baseline pool of posts from the Bluesky firehose (all the English posts to Bluesky in a one-week period). We then filtered these posts based on the criteria of our conditions to simulate our engagement-based and representative diversification conditions (see Table 3 above). In an important proof-of-concept, we confirmed that on Bluesky, the engagement-based feed increased toxicity by 48% *relative to the base rate* of toxicity on Bluesky (measured by the firehouse). 2.9% of content pulled from the firehose was classified as toxic compared to 4.3% of content in the simulated engagement-based feed. Importantly, our representative diversification feed did not simply remove toxicity, but made the levels more representative of the base rate at 2.4% of content. Full details of the new pilot are available in SI Appendix, Section 6.0 of the manuscript.

Below, we also paste the updated description of the representative diversification algorithm (Table 3 and pg. 9 of the manuscript):

> The 'representative diversification' algorithm (Treatment 2) is designed to diversify the participants' feed relative to engagement-based feeds by increasing the *social representativeness* of the feed, relative to the true distribution of ideological extremity present on the platform. Because it is intractable to estimate the true distribution of political extremity for all possible posts, we use three filters that should produce more representative posts, i.e., posts that do not represent the long tails of the distribution of political extremity. Specifically, for posts that are classified as political in nature (see SI appendix, 3.3.2), the algorithm (1) reduces probability of super posters, or those who post more than 5 times in a 24-hour window, (2) reduces probability of toxic posts, and (3) increases probability of constructive dialogue posts. We assume that these three filtering criteria in practice yield sampling that drops the long tails of the political post distribution because it is now well-established that the large majority of toxic political content is posted by extreme partisans who represent a small percentage of users[16–18]. In other words, long-tail users are the most extreme, the most toxic, and post the most content. These criteria expand upon previous work that simply removed toxic content in feeds, which was successful in reducing users' engagement with toxic content.[36]
>
> Importantly, the representative diversification algorithm can be applied to

any platform with varying underlying distributions of political ideology, whereas a bridging algorithm that aimed to expose users to outgroup political posts (e.g. [20]) would be constrained to platforms that have enough outgroup content in the first place. The extent to which a user in the representative diversification feed is likely to see an "outgroup" post depends on the true distribution of users on the platform. For instance, many platforms including Bluesky tend to be very left-leaning, but the representative diversification algorithm will still diversify the feeds in terms of ideological extremity of left-leaning users. In fact, correcting perceptions of ingroup extremity may be more effective than exposing users to less extreme outgroup content which has the possibility of backfiring in increase outgroup animosity (Bail). However, applied to a platform with a true mix of left-right ideology, the representative diversification algorithm would still limit extreme users appearing in posts from both sides of the political spectrum.

Toxicity and constructive dialogue are classified using validated models built to classify social media data from Google's Perspective API [36]. Importantly, we demonstrate in pilot testing that filtering by removing toxicity and selecting for constructive dialogue still allows for strong moral opinions to be displayed. Rather than simply removing strong moral opinions, these filtering steps remove posts that contain strong moral opinions *and* have toxic properties such as personal attacks; in other words toxic content is filtered independent of moral content (see SI Appendix, Section 3.0 for full technical details and validation of classifying toxicity and constructive dialogue). Furthermore, pilot testing described in SI Appendix, Section 6.0 suggests that these criteria do in fact reduce toxicity and increase constructiveness of dialogue relative to an engagement-based algorithm in simulated feeds tested on the Bluesky API firehose.

R2.7 Concern about previous analysis with partisan animosity

Reviewer #2 was concerned about the proposed moderation analysis when we included partisan animosity as one of the primary outcomes: *"Specification: I am concerned that RQ3 H3-H4 condition on a post-treatment variable. This could be a "bad control" that would bias the results. Covariates that could be affected by the treatment should not be used as moderators. Estimating the effect of norms on partisan animosity directly is also problematic since norms are not randomly assigned. I think the authors should estimate the effect of the treatment (the feed) on partisan animosity. If an effect is found, they can make some assumptions to try to analyze whether the feed affects animosity through the social norms channel."*

We first note that partisan animosity has been moved to an exploratory outcome. That said, we agree with Reviewer #2 that the moderation analysis may have specification issues, so in the exploratory analysis we will look at the main effect of condition on

partisan animosity, and we will explore a *mediation* analysis if there is a main effect. Since this analysis will be exploratory, we do not describe it in the registered report (following *Nature* guidelines), but we thank Reviewer #2 for this idea when we explore condition effects on partisan animosity.

R2.8 Highlighting pooling of standard errors

Reviewer #2 suggests we highlight the fact that we are pooling standard errors across several repeated measures to increase power: *"If I understand correctly the authors are pooling data from several surveys. That's a clever way to increase power. I think the authors should state that they will cluster standard errors at the participant level and take that into account in their power calculations."*

Reviewer #2 is correct and we now highlight this component of our model more in our discussion of the power analysis on pg. 11 of the manuscript:

> We plan to recruit 2000 participants to the study. We arrived at this number by conducting an *a priori* power analysis using the package 'simr' in *R* 4.3.1, aiming to detect effects as small as $d = .20$ with 99% power. We assumed an alpha level of .05 and 3 experimental conditions for our statistical test across all of our research questions that would have the lowest amount of power (a linear mixed model testing for between-group differences in norm perceptions; RQ2-3). In this model, we exploit the repeated measures nature of our data to cluster standard errors at the level of participants and increase power. We also assume a 20% attrition rate (participants who do not complete 75% of the survey, see below). The R script for the power analysis for 1600 participants (2000 - 20% attrition) is available here: https://osf.io/k2crm/?view_only=9a055df86e044d3f9a5aa8b118c0b4b0

R2.9 Question about baseline pool of content frequency

Reviewer #2 asks whether we know how content will form our baseline pool of content that is filtered to form the different feeds in our experiment: *"Minor – what is the size of the baseline pool of content? This could matter for how much variation the treatment generates. For example, if the baseline pool includes 100 posts and the three treatments are reranking the order of these posts, then a user who views all 100 posts in a session will only see them in a different order but still be exposed to the same content across the treatment arms."*

Based on our pilot testing and discussions with the Bluesky development team, we expect ~1 million posts per day (which may increase during the election season). In particular, the proportion of political posts will increase during the election season, and thus we will not have an issue of users running into the same content after re-ranking. We describe the numbers expected for our baseline pool of posts on pg. 10 of the manuscript:

> In order to collect the baseline pool of posts that will be filtered based on experimental condition, we pull from the Bluesky firehose (all posts in real-time) every 15 minutes, and immediately filter out posts that are non-English, spam, or contain vulgar or explicit content. We estimate that we will collect 60,000,000 posts throughout the 8-week experiment (~1,000,000 posts / day) through the Bluesky firehose in our baseline pool of posts, based on our discussions with Bluesky developers in their Discord channel. Further filtering of the baseline posts are based on the criteria for each experimental condition as described above. Full technical details are presented in SI Appendix, Section 3.2. The Python code for our in-progress data pipeline, data storage, and browser extension is available here.

Responses to Reviewer #3 comments

R3.1 Clarify decision to run experiment during U.S. election

Reviewer #3 asks us to clarify why we decided to run the experiment during the 2024 U.S. election: *"One thing that was not clear to me, however, was the relevance of conducting the experiment during the election (or, why is the national election mentioned in the title?). Is it because we expect users to behave differently during the election? Or because we expect the algorithms to play a bigger role at influencing the content that users see? Does it help with making the experiment more natural?"*

We chose to run the experiment during the U.S. election for two main reasons. First, it allows us to measure perceptions of social norms related to political discourse and polarization during a time where these outcomes are (arguably) the most meaningful (i.e., norms are most likely to impact actual political behavior such as voting). Second, the election time period will make political content represent a greater proportion of posts (Boulianne & Larsson, 2023; Neely, 2021), which is likely to make our representative diversification algorithm have a greater impact on content. We now reference these reasons on pg. 7:

Experiment procedures. The experiment will occur over a 2-month period from Monday, September 2nd, 2024 to Monday, November 4th, 2024 (the day before the 2024 U.S. Presidential election). We chose to run the experiment during this time because our research questions related to perception of norms around politics are most meaningful during this time as they may be related to political behavior such as voting. Furthermore, the election time period will make political content represent a greater proportion of posts, which is likely to make our representative diversification algorithm have a greater impact on content.

R3.2 Concerns about how to measure social norm misperceptions

Reviewer #3 brings up several concerns and suggestions for measuring social norm misperceptions: *"One of my main concerns regards the extent to which the proposed study can satisfactorily answer the research questions. The proposed analysis cannot really pin down the effect of algorithms on the misperception about social norms for three reasons. First, because it is unclear whether the proposed measures of perceived social norms capture the respondents' own acceptability of a certain behavior, or their belief about others' acceptability (what the literature typically considers to be the social norm). For instance, when a user sees the question: "How socially appropriate is it to post content that is likely to create conflict among political groups on social media," do they reply how acceptable it is for them, or do they reply their belief about how acceptable it is for others? Second, given this issue, the design does not elicit the objective "ground truth" that is required to measure misperceptions (which I assume are defined as the difference between the beliefs about the social acceptability relative to the objective or "ground truth" acceptability). Suppose that the authors find that the engagement-based algorithm increases users' beliefs about the acceptability of posting IME content to 70/100 (vs. say, a level of 60 in the control group and 65 in T2). How do we know that this increase is not actually closing the gap (if the objective acceptability is 85?) In that case, the control group is starting from a misperceived level (potentially because these are active social media users who could be on an echo chamber), and T1 is actually reducing the misperception (and T2 causes a lower reduction in the misperception). To address this issue, the design would need to measure two objects (see Bursztyn et al., AER2023 "Misperceived social norms…"): the objective acceptability based on individual-level responses (i.e., based on the subjects' own acceptability) and the subjects' beliefs about others' acceptability. Third, even if the experimental designed measures the pre-intervention objective/ground truth social acceptability, the intervention could affect not just the users' belief about this social acceptability, but also the objective social acceptability itself. In particular, if we define the misperception as the difference: belief about social acceptability – objective social*

acceptability, it is clear that the intervention could affect both components of this difference. For example, an increase in the exposure to toxicity could make people more toxic, which would increase the objective social acceptability of toxicity. If users' beliefs about this social acceptability increase by the same amount, the actual misperception would not change, but 1) the current design that conflates misperceptions and perceptions (beliefs) would (incorrectly) conclude that misperceptions increased and 2) even a design that elicits the objective social acceptability but does not measure the change in this social acceptability would conclude that misperceptions increased. To address this issue, the design would need to measure the change in the objective acceptability after the intervention (vs. before). An alternative would be to reframe the research questions around the effect on the perception of social norms, rather than around the misperception of these norms. However, just looking at the effect on perceptions leaves many questions unanswered (e.g., why is there an effect on perceptions? Is it because people are changing their own opinions about acceptability or is it because they are updating their beliefs about others?).

We found Reviewer #3's detailed discussion of the different ways to measure norm perceptions very insightful. It pushed us to clarify the ways in which we define misperceptions of social norms, and prompted us to include new measures of prescriptive norms to account for some of Reviewer #3's distinctions. We focus on perceptions of polarizing discourse on the Bluesky social network since it is our main outcome of interest in the revised version of the manuscript (but we also use the same procedure described below for ingroup, moral and emotional (IME) content).

<u>Measuring descriptive norms and misperceptions</u>

First, we consider perceptions of *descriptive norms*, or perceptions of how common polarizing discourse is in a user's Bluesky social network. We define misperceptions of descriptive norms of polarizing discourse using toxicity as a key metric. Thus, misperception is defined as the difference between (1) participants' perception of the percentage of toxic political content in their Bluesky social network and (2) the actual percentage of toxic political content posted to the Bluesky social network. We measure both of these outcomes in the revised version of the manuscript.

(1) Perceptions of polarizing discourse in a user's Bluesky social network is measured every week using the following revised item (pg. 29 of the manuscript).

Perception of toxic political discourse in participant's Bluesky social network (0-100) (descriptive) [1,2]

What percentage of content in your Bluesky social network is toxic political content that is likely to create conflict among political groups?

0%	50%	100%
I never see toxic political content on Bluesky		Toxic political content is all I see on social media

(2) As stated on pg. 10 of the manuscript, we are collecting all (english, non-spam) messages from the Bluesky firehose everyday (~7 million messages per week). To measure the actual percentage of toxic political content, we will classify all the posts as political and then as toxic using the classification tools described in SI Appendix, Section 3.3.2 (Google's Perspective API). This allows us to estimate the actual percentage of toxic political content on the Bluesky social network to compare the perceptions in each condition.

Thus, we can test whether the engagement-based algorithms increase perceptions of descriptive norms regarding polarizing discourse on social media, and we can measure to what extent this is a misperception compared to other conditions (all with reference to the true baseline frequency). We note that participants are given the definition of toxic content based on how the Perspective API defines it, to ensure participants' perceptions and the classification scheme are using the same working definition of toxicity.

<u>Measuring prescriptive norms and misperceptions</u>

We agree with Reviewer #3's distinction between a participant's perceptions of the social appropriateness of others posting toxic political content (Item A below), and their perception of others' judgment of social appropriateness of posting toxic political content (Item B below). We measure these constructs as follows (SI Appendix, Section 1.3, or pg. 29 of the manuscript):

Northwestern University
2211 Campus Drive
Evanston, IL, 60208

Item A (other):

Perception of toxic political discourse in participant's Bluesky social network (0-100) (prescriptive - other) [1,2]

How socially appropriate *do other people on Bluesky feel it is* for you to post toxic political content that is likely to create conflict among political groups?

-50	0	50
Very socially inappropriate	Completely neutral	Very socially appropriate

Item B (self):

Perception of toxic political discourse in participant's Bluesky social network (0-100) (prescriptive - self) [1,2]

How socially appropriate *do you feel it* is for other people on Bluesky to post toxic content that is likely to create conflict among political groups?

-50	0	50
Very socially inappropriate	Completely neutral	Very socially appropriate

Now, we can define misperception of prescriptive norms as the difference between Item A and the mean of Item B across participants.

Also to Reviewer #3's point, since we measure all descriptive and prescriptive norm items weekly for 8 weeks, we can explore the changes in magnitude of misperceptions created (or mitigated) by each condition over the course of the experiment. Although our main hypotheses described in the registered report are about overall misperceptions comparing conditions, we agree it is important to test changes in misperceptions over time as an exploratory analysis and plan to do so.

In summary, the revised version of the manuscript now details measurement of perceptions of descriptive and prescriptive norms of political dialogue, as well as their

counterpart “baseline truth” so that we can determine the magnitude of norm misperceptions. As Reviewer #3 suggests, we believe this is a considerable improvement to the previous version of the manuscript that only measured changes in perceptions of norms.

R3.3 Concern about participant incentives

Reviewer #3 is concerned about participant incentives: *“Regarding the soundness and feasibility of the methodology and analysis pipeline, my main concerns are that 1) I don’t think that the design incentivizes participants to actively use the platform. In terms of 1), participants will be told that “they must log in to Bluesky at least twice per day, and engage in some way (post, like, share) at least once per day.” Can’t participants simply log in two times, like the first post they see, and then log off? It is notoriously difficult for platforms to gain new users, particularly given strong network effects, so I doubt that these incentives will be enough to convince them to actively use Bluesky. Better designs (of course, less feasible) would either recruit currently active Bluesky users or would pay participants based on their total activity (or active time spent) on the platform.”*

As described above in **E.3**, we agreed with Reviewer #3 (and Reviewer #2) and opted to recruit active Bluesky users – rather than recruiting new users to Bluesky – to address concerns about participant incentives to use the platform. The relevant revision is in the main text (pg.7):

> *Experiment procedures*. The experiment will occur over an 8-week period from Monday, September 2nd, 2024 to Monday, November 4th, 2024 (the day before the 2024 U.S. Presidential election). We chose to run the experiment during this time because our research questions related to perception of norms around politics are most meaningful as they may be related to political behavior such as voting. Furthermore, the election time period will make political content represent a greater proportion of posts[27,28], which is likely to make our representative diversification algorithm have a greater impact on content. To provide time to pre-screen participants, starting in August, 2024 we will recruit participants for a study on “Testing New Social Media Feeds” via advertising on Reddit (e.g. on r/BlueSkySocial with 15.4k active users) and X, both of which contain active Bluesky users. We will aim to collect the email addresses of 3000 interested users to fill out a brief demographic survey with the aim of recruiting 2000 study participants (see Sampling Plan below). If advertising on Reddit and Twitter fail to secure 3000 interested users, as a backup we will work with Prolific consultancy services to recruit participants on Prolific who are active Bluesky users (correspondence with Prolific consultancy services confirmed that they can add Bluesky use as a screen to their current pool of participants).

Of the 3000 interested users, we will select users that log into Bluesky at least once per day, or once per week if not enough participants meet the daily use criteria. Once participants are invited to the experiment, we will test for comprehension of the experiment instructions, which requests continued use of the platform for the 2-month study period. Experiment instructions can be viewed in SI Appendix, Section 1.1.

Screening and replacement of participants who do not understand or consent to the experiment instructions will be replaced by other participants in the first recruitment group of 3000 participants. Once the participant pool is finalized, all participants will take an intake survey that measures demographics and provides starting measurements for our key dependent variables, see Table 2 for a summary, and SI Appendix, Section 1.2 for full details of measures.

R3.4 Concerns about effect size on partisan animosity outcome

Reviewer #3 is concerned that our experimental manipulation will not have a measurable effect on partisan animosity: *"2) I don't think that the experiment is well powered to detect the effect sizes that one would expect in this literature. More importantly, in terms of 2), the effect sizes in the literature that studies the effects of algorithms on attitudes (in the field) are much smaller than the 0.10 SD that the experiment is powered to detect and, because of this, typically have sample sizes at least an order of magnitude larger than the proposal (or exploit more granular repeated-measure longitudinal designs). For example, the paper by Levy, AER2021 "Social Media, News Consumption, and Polarization" cited in the proposal (which is one of the papers with larger effect sizes), estimates an effect size on affective polarization of 0.03-0.06 SD, with a sample size of 17,635. Note that these concerns could be alleviated by providing pilot data, but the pilot data that the current proposal reports only speaks to the "first stage" effect of the algorithms on content, not on the "second stage" outcomes of interest. As it stands, my prior after reading the proposal is that the experiment will not be very informative about the research question. However, I do not want to discourage the authors from running this ambitious experiment, so I would recommend to maybe emphasize a bit more the more granular outcomes (for which they have even daily data, which allows them to exploit better the longitudinal variation) such as engagement on Bluesky or interactions with other users, which are also interesting."*

As discussed in **E.1**, we removed partisan animosity as a primary outcome of interest and now focus on social norm (mis)perceptions and perceived polarization as suggested by Reviewers #2 and #3. However, we still measure partisan animosity as an exploratory outcome.

Northwestern University
2211 Campus Drive
Evanston, IL, 60208

R3.5 Comment on methodological detail

Reviewer #3 approves of the level of methodological detail: *"Lastly, the degree of methodological detail provided seems sufficient to replicate exactly the proposed experimental procedures and analysis pipeline and offers minimal possibility to deviate from the proposed analysis."*

We thank Reviewer #3 for recognizing the effort and detail we put into the registered report so that it could be replicated, and so that it commits our research team to a specific analysis plan.

R3.6 Remark on code availability

Reviewer #3 had trouble installing R packages included in our power analysis script: *"I couldn't install some of the packages required to run the code analysis_template.R (lme4 and in consequence lmerTest and simr), so I couldn't review it in depth."*

We apologize for the inconvenience. In preparing the revision, we tested our script on several machines and with different users, and all of these runs worked correctly. However, we now include detailed instructions for installing R studio and R packages as a user note in the Wiki on our OSF page that contains the analysis scripts.

Responses to Reviewer #4 comments

R4.1 Comment on methodological detail

Reviewer # 4 approves of our level of methodological detail: "*The proposal clearly stated the recruitment procedure, experimental protocol (including survey design, algorithm construction, and randomization), recruitment method, hypotheses to be tested, and outcome variables to be used. The proposal is sufficiently detailed to allow for an exact replication of the proposed experimental procedures and analysis pipeline after the experiment (since all code and data will be made publicly available). Furthermore, I found the description of the methods is sufficiently clear and detailed to prevent undisclosed flexibility in the experimental procedures or analysis pipeline. I believe that the authors have considered sufficient outcome-neutral conditions for ensuring that the results obtained are able to test the stated hypotheses."*

We thank Reviewer #4 for recognizing the effort and detail we put into the registered report so that it could be replicated, and so that it commits our research team to a specific analysis plan.

R4.2 Clarity on design of the engagement-based algorithm

Reviewer #4 wanted us to clarify what the engagement-based algorithm condition is designed to model: *"Different from the existing field experiments, the proposal at hand plans to recruits study participants to open new, anonymous accounts on a social media platform that users are not familiar with yet (instead of running the field experiment with pre-existing users on Facebook/YouTube/Twitter); furthermore, different from existing field experiments, the proposal at hand will not compare alternative feed algorithms (such as reverse-chronological, or those weeding out toxic posts – both of which have been tested/evaluated in the prior work cited above, as well a "representativeness" algorithm that is supposed to how more even-handed content from across the political sprectrum – closely related to the field study by Levy, 2021) against the "status quo" of the existing platform, but instead it will compare against an engagement-maximizing algorithm that the research team themselves will devise. It was unclear to me whether this engagement-maximizing algorithm is aimed at emulating as close as possible the algorithms used by existing platforms such as Facebook or Instagram, or whether it is supposed to capture engagement-maximization period (with no regard to whether it does so similarly to existing platforms' algorithms)."*

Our engagement-based algorithm is primarily designed to capture engagement-maximization, but we designed it specifically with reference to Twitter's open-source details of their engagement algorithm (see here). We reference this point on pg. 8 of the revised manuscript:

> The 'engagement-based' ranking algorithm (Treatment 1) represents the key properties of "personalized" feeds used to deliver content on major social media platforms such as X and Meta (see [30,31]). These algorithms rank content based on two main criteria: *similarity* to content the user engaged with previously, and similarity to content that has high engagement in the broader community (e.g., content that has high engagement on the platform and/or content that the users' social network engaged with). To measure similarity and predict the likelihood of future engagement, we will construct a machine learning model that matches the state-of-the-art in engagement prediction on social media platforms[32,33]. During inference, the model generates posts to suggest for a user

by using approximate nearest neighbor (ANN)[34,35] to return the top posts a user has not engaged with but is most likely to (see SI Appendix, Section 3.0 for full model details).

R4.3 Comparing our proposed work to previous experiments

Reviewer #4 pushes us to clarify how our work extends previous work: "*As stated above, that there is already existing field experimental work that tests reverse-chronological feeds (Guess et al. 2023), feeds that down-rank toxic content (Beknazar-Yuzbashev, 2023), as well as feeds that promote content from across the political spectrum (Levy, 2021) against the status quo algorithms used by large platforms such as Facebook and YouTube. In my view, the proposal at hand stays rather close to the existing work in the treatment arms it aims to test. The investigation into the role of second order believes, while interesting, in my view provides a neat, but not ground-breaking contribution, in part because the study setting has lower external validity than the "native social media user" settings.*"

Although our work is certainly related to the work cited by Reviewer #4, we nonetheless believe ours builds upon and extends previous research, offering a substantial contribution beyond the previous work cited by Reviewer #4.

(1) We first note that Reviewer #4 considered our study to have lower external validity than previous work because we were recruiting new users. As a point well taken, our revised proposal recruits active Bluesky users (see **E.2**) and thus resolves this concern.

(2) Along with Reviewer #2 and Reviewer #3, we believe that studying the social norm perception outcome is a significant theoretical and practical contribution to the literature. The study of norms has a long history of research in the behavioral sciences for their significant impact on people's behavior as well as what is voted into public policy. In our revised proposal, we also now have the ability to measure misperceptions by reference to true base rates of descriptive and prescriptive social norms. This is a significant contribution because - as we outline in the introduction - norm perception may be the key mechanism through which exposure to different social media feeds affects attitudes and behavior over time.

(3) Our representative diversification intervention departs significantly from prior interventions both in its theoretical goal and its implementation. This is especially

true in the modified version described in detail in **R1.1**. Our intervention is specifically aiming to increase the social representativeness of feeds, relative to the true political ideological distribution of users on Bluesky. This is a different goal than simply filtering out toxicity (Baknazar-Yuzbashev, 2023), or having users follow cross-partisan news sources (Levy, 2021). In the revised introduction, we cite these previous important papers and explain how our intervention goal is different. Furthermore, the implementation is different in that filter of sociopolitical + toxic content is only one of 3 key ranking criteria. The bridging criteria has been removed in the revised version of intervention. Pragmatically, our intervention is also applicable to platforms that do not have both politically left and right content since it draws on the existing baseline distribution of political content of a platform. Also, because it draws upon engaging content, it has a higher likelihood of being adopted by the platforms for where its impact could be publicly realized.

For a detailed discussion of the updated version of our representative diversification algorithm, see **R1.1.**

R4.4 Concerns with recruiting new users to Bluesky and incentives

Reviewer #4 is concerned that external validity is low if we were to recruit new users to Bluesky: *"By randomizing participants into different social media feeds generated by the research team in a tightly controlled setting, the authors can cleanly isolate the causal impact of these different feeds. So, the internal validity is very high. My concerns are more with the external validity: The authors plan to run the field experiment on a small, not widely used social media platform to which they aim to recruit new members via Prolific, a provider of research participants for social science studies. However, the "native" social media environment differs from the one proposed in this field experiment in ways that render it difficult to generalize findings form the experiment to social media settings "in the wild": for example, the participants would use anonymized usernames, thereby turning off key social image and reputational concerns in their sharing behavior. Furthermore, the social media experience on the platform used for the experiment would not include the participants' social network (friends, family, colleagues, etc.), who are a key component of the ordinary social media experience. In sum, this approach to direct study participants to a new social network comes at the big cost of artificiality and limited generalizability to the more "natural" social media settings. Relatedly, in order to induce participants to use the platform, the experimental protocol states that participants would be required to log into the social media platform twice a day, including sharing a post every day. I am worried that this is not enough to induce participants to use the*

platform. A suggestion to strengthen external validity and to ensure higher use of the platform, could be to look into recruiting pre-existing Bluesky users directly."

We were persuaded by Reviewer #4's (and also Reviewer #3) comments and opted to recruit active Bluesky users in the revised version of the proposal. We believe that this decision strongly improves the external validity concerns and assuages issues of participant incentives of platform use. We note that we will also pre-screen Bluesky users such that we recruit and enroll users who log in at least once a day. For convenience we paste our response from E.2 below:

We updated our participant recruitment plan to recruit 1600 active Bluesky participants (after attrition) in the Design section on pg. 7 of the manuscript:

> *Experiment procedures*. The experiment will occur over an 8-week period from Monday, September 2nd, 2024 to Monday, November 4th, 2024 (the day before the 2024 U.S. Presidential election). We chose to run the experiment during this time because our research questions related to perception of norms around politics are most meaningful as they may be related to political behavior such as voting. Furthermore, the election time period will make political content represent a greater proportion of posts[27,28], which is likely to make our representative diversification algorithm have a greater impact on content. To provide time to pre-screen participants, starting in August, 2024 we will recruit participants for a study on "Testing New Social Media Feeds" via advertising on Reddit (e.g. on r/BlueSkySocial with 15.4k active users) and X, both of which contain active Bluesky users. We will aim to collect the email addresses of 3000 interested users to fill out a brief demographic survey with the aim of recruiting 2000 study participants (see Sampling Plan below). If advertising on Reddit and Twitter fail to secure 3000 interested users, as a backup we will work with Prolific consultancy services to recruit participants on Prolific who are active Bluesky users (correspondence with Prolific consultancy services confirmed that they can add Bluesky use as a screen to their current pool of participants).
>
> Of the 3000 interested users, we will select users that log into Bluesky at least once per day, or once per week if not enough participants meet the daily use criteria. Once participants are invited to the experiment, we will test for comprehension of the experiment instructions, which requests continued use of the platform for the 2-month study period. Experiment instructions can be viewed in SI Appendix, Section 1.1.
>
> Screening and replacement of participants who do not understand or consent to the experiment instructions will be replaced by other participants in the first recruitment group of 3000 participants. Once the participant pool is finalized, all participants will take an intake survey that measures demographics and provides starting measurements for our key dependent variables, see Table 2 for a summary, and SI Appendix, Section 1.2 for full details of measures.

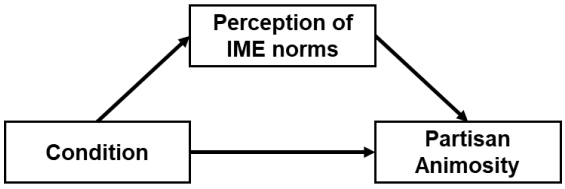
Northwestern
Kellogg
School of Management
Northwestern University
2211 Campus Drive
Evanston, IL, 60208

Responses to Reviewer #1

R1.1 Suggested new analysis

Reviewer #1 requests that we include analyses of the relationship between norm perception and user posting/engagement with Ingroup, Moral and Emotional (IME) content (behavior): *"I found the authors' responses to be very reasonable. They satisfactorily answered my concerns about the representative algorithm's content curation in a way that is logically coherent and more generalizable. My only remaining concern is that the analyses declared in the paper do not look at the interplay between the role of perceived norms and downstream behavior. Perhaps I am missing that from not being fluent with the R code, but, if not, it would be good to look at relationships between IME content posted, norms, and behavior rather than the effects of the experiment on each of these pieces alone."*

We agree that it is important to explore the relationship between changes in norm perception and platform behaviors and intentions, including engagement with IME content. We plan to test a mediation model on three behavior and intention outcomes in which condition impacts platform behaviors through changes in norm perceptions of IME content. The behaviors and intentions we will test in this mediation model include: (1) engagement with IME content on the platform, (2) partisan animosity and (3) endorsement of anti-democratic norms (see pg. 29 of the manuscript for measure details). For example, see the mediation figure below:



Following the advice of several reviewers in the last round of revisions, we will keep the mediation models exploratory since our sample size is 95% powered to detect our primary hypotheses only. Following Nature guidelines: (https://www.nature.com/nature/for-authors/registered-reports) we will put these mediation analyses in the Stage 2 report under a section entitled "Exploratory analyses", which all reviewers will have a chance to review.

Responses to Reviewer #2

R2.1 Clarification of RQ1, H2

Reviewer #2 requests further clarification about the importance of RQ1, H2: "*This registered report improved substantially from the previous version. First, the authors will now recruit active Bluesky users to participate in the experiment. This improves the*

external validity of the research since these participants use the platform. This strategy also decreases the concern that participants will not engage with the platform during the treatment period. While external validity is still an issue since Bluesky is not as important a platform as Facebook, TikTok, Instagram, etc., this change is a big improvement. Second, the authors update their outcome variables. Most importantly, they replaced partisan animosity, which was underpowered, with perceived polarization, where they can make a bigger contribution. The authors also modified the "representative" algorithm.I have several remaining comments. 1. I still do not understand what RQ1, H2 is testing. The authors design an algorithm that reduces toxicity. Wouldn't such an algorithm almost mechanically decrease the amount of IME content displayed to the user (I understand that IME and toxicity are not the same thing but the concepts are related)? It sounds like RQ1 is testing what the authors' algorithm shows. While this is important information that should be displayed in the paper in order to interpret the rest of the results, I do not think it is testing a theoretically important question."

We thank Reviewer #2 for pushing us to further clarify why RQ1 is a theoretically important contribution, rather than simply a manipulation check. We believe that the answer to the question, "*I understand that IME and toxicity are not the same thing but aren't the concepts are related?*" is "*yes*," but importantly the answer is not true by definition. It is possible to express moralized and emotional content without expressing toxicity (e.g., shame-filled apologies or calls to collective action), and thus there is no mechanical reason why RQ1, H2, should be true. As of now, the link between IME content and toxic content has not been established by prior literature; rather, it remains a theorized link in our model.

Testing whether reducing toxicity also reduces IME content (RQ1, H2) helps us understand how tightly coupled these variables really are. It is plausible that our intervention algorithm does not actually reduce IME content, or that it reduces one of the specific I-M-E dimensions, but not all of them. Given these considerations, we do not believe that the RQ1 hypotheses simply represent a manipulation check, and would prefer to keep them in as a component of our primary analyses. However, if the editor disagrees, we're certainly willing to treat RQ1 as exploratory.

R2.2 Improving precision of the norm perception measure

Reviewer #2 suggests that we reword our descriptive norm perception measure to be more comparable to how we measure IME information in the Bluesky firehose:
"The authors explain that the base rate of IME content is the 'total proportion of IME content present in a sample of content collected from the Bluesky firehose (all messages posted to Bluesky after filtering) in a given time period". This is a great definition. It is based on objective data and takes into account all posts, regardless of the algorithm. It is also likely that people may suffer from misperceptions regarding this

measure. They may see more emotional posts because an algorithm maximizes engagement, and thus think that posting such content is more common than it actually is. However, if I understand correctly this is not what the authors ask about in the survey. They ask: "What percentage of posts in your Bluesky social network are emotional in nature" (A) It is not clear if this question refers to the posts people see in their feeds or the posts created. If people interpret it as the posts they see in their feed, then it will be mechanically affected by the treatment. (B) This question asks about one's social network but not about Bluesky generally. I think this question should be reworded so it is clear to the respondents what the authors are referring to and in a way that is consistent with their definition of the base rate of content."

We found this comment very helpful. In response, we revised our measures of descriptive social norm perceptions to ask about Bluesky overall rather than "your Bluesky social network". This way, participants' responses to the descriptive norm perception measures are now better calibrated to the base rate measurement of IME content from the Bluesky firehose.

> Before: "What percentage of *posts in your Bluesky social network* are emotional in nature?"
>
> Revised: "What percentage of *posts on Bluesky* do you think are emotional in nature?"

All updated survey questions can be viewed on pgs. 31-32 of the manuscript.

R2.3 Suggestion to change name of the intervention algorithm

Reviewer #2 suggests that we may want use a more precise name for our intervention algorithm: *"I am still not sure why the "representative diversification" algorithm is called representative as it does not aim to target representativeness, instead, it aims to change the distribution of three dimensions of posts: number of posts by super posters, number of toxic posts, and number of constructive posts. The authors nicely show that the algorithm becomes more representative on one dimension by decreasing toxicity, but there are many other dimensions of representativeness. For example, the posts in the representative algorithm are probably not representative in terms of the topic discussed, how extreme they are, the left-right balance, and so forth. Based on Table S4 the representative algorithm is not even substantially more representative in terms of constructiveness (it shows ~39% more constructive posts than the baseline rate, while the engagement-based algorithm shows ~40 fewer posts than the baseline rate). As I wrote in my previous review, I understand the logic of using this algorithm. It neatly balances engagement with less polarizing content, and thus, it is also policy-relevant. Therefore, I do not think the authors necessarily have to change the algorithm, but I do think they should rename it so the readers can more easily understand what exactly is being tested. I actually like the name proposed by R3: "depolarizing algorithm", but that is just one suggestion."*

Northwestern University
2211 Campus Drive
Evanston, IL, 60208

We are persuaded that our algorithm isn't "representative" in the broadest meaning of that term, and that we need a different term. It's true that our intervention algorithm is aimed at improving the "social representativeness" of feeds on a specific dimension: feeds in the experimental condition will align more closely with the true distribution of political ideological extremity of a given platform. But the reviewer is correct that this definition of "representative" is probably too specific to align with general uses of the term. That said, our intervention is not simply a "depolarizing" algorithm, as we are not intending to make feeds full of agreeable political content (in fact, one strength of our feed is that it will still represent true disagreement and strong opinions if the platform has such content).

Given these considerations, we renamed our intervention algorithm as the "*diversified extremity algorithm*," a term capturing that our algorithm diversifies the political extremity of feeds. We believe the new term provides a clearer description of the criteria used in the intervention.

R2.4 Concerns over Bluesky user recruitment

Reviewer #4 raises the point that it may be difficult to recruit Bluesky users: *"I am worried that the authors will not be able to recruit enough participants. For example, the authors mention that Reddit has 15k active BlueSky users. If the authors design great ads, maybe 3% of those users will click the ad, 2% will meet the screening criteria, and 1% (150 users) will complete all surveys. Of course, this is just a back-of-the-envelope calculation. I do not think this is a reason to reject the design and the authors do not need to respond to this concern. I am only making this point to make sure the authors are aware of the recruitment risks (I am sure they already are)."*

We are very receptive to this observation. We are aware of the potential difficulties in recruiting active users, and in the revised manuscript we update recruitment efforts with a multi-stage strategy until we hit our target of 2000 participants. We describe this on pg. 7 of the manuscript:

> To provide time to pre-screen participants, starting in August, 2024 we will recruit participants for a study on "Testing New Bluesky Feeds" via advertising on Bluesky, and also on Reddit (e.g., on r/BlueSkySocial, which has 15.4k active users) and X, both of which contain active Bluesky users. From advertising we will aim to collect 3000 user responses on a brief demographic survey with the aim of recruiting 2000 study participants (see Sampling Plan below). If advertising on Reddit and Twitter fail to secure 3000 survey responses, as a backup we will work with Prolific and Connect consultancy services to recruit participants on who are active Bluesky users (correspondence with Prolific consultancy services confirmed that they can add Bluesky use as a screen to their current pool of participants, and Connect already has at least 900 active Bluesky users). To further boost the recruiting if required, we will also work with

Bluesky developers to post a link to our study on the official Bluesky developer accounts (accounts with 100k to 1M followers).

We note that Bluesky developers have verbally agreed to post our study advertisement on their official accounts (accounts with 1 million to 100k followers), and the advertisement would specify that our study does not entail any collaboration with Bluesky or its employees. With these plans in place, we are confident we can get the required sample size. However, this comment also made us think carefully about a related issue: participant attrition. To this effect, we developed a contingency plan in the event our attrition rate is notably greater than estimated after two weeks of the experiment, which involves recruiting more participants to start 2 weeks later using the same pool of participants from our recruitment efforts. Under this contingency plan, we would enter participant start date as a covariate in a sensitivity analysis for all main models. The updated recruitment procedure as well as attrition contingency plan is now described on pg. 10 of the manuscript:

> If our attrition rate is estimated to be higher than 20% after two weeks of data collection, we will recruit new participants to increase the sample size back up to 2000 (still expecting ~20% attrition by the end of the experiment). If this plan is required, we will run robustness tests on all main models described below that control for participant start date.

Responses to Reviewer #3

R3.1 Providing definition of "toxicity" to participants

Reviewer #3 requests that we provide participants with a definition of toxicity when we measure their perception of toxicity norms on Bluesky: *"I will keep my comments short; the authors clearly worked hard on my suggestions (and I believe that also those of the editor and the other referees) so I don't have much more to add. The first remaining suggestion I have is to provide participants with a definition of what "toxic" means in the norm perception outcomes (2), (3a), (3b). You could potentially use the definition of Perspective API."*

We agree that it is important to provide participants with a working definition of toxicity to streamline interpretations of our measure. We now include the Perspective API definition and provide it to participants when evaluating our norm perception outcomes (2), (3a) and (3b). Below, we paste the definition, and see also pg. 30 of the manuscript:

> *Note: For all measures of toxicity perceptions above, participants are provided with the following definition of toxicity from Google's Perspective API: *A rude,*

disrespectful, or unreasonable comment that is likely to make someone leave a discussion

R3.2 Testing differential attrition in sample

Reviewer #3 requests that we include a formal rest of differential attrition and a plan to handle this potential issue: *"The second remaining suggestion is to add a test of differential attrition in the paper (and, ideally, a way to address differential attrition in case it occurs) and measures of attrition over time."*

We expanded our discussion on pg. 10 of the manuscript to explain more precisely how we will treat missing data and test for differential attrition:

> Missing data will be handled depending on whether missing data vary significantly by condition ('differential attrition'). We will test for differential attrition using a chi-square test, where a positive sign of differential attrition is determined by Ns that are significantly different as a function of condition. If there is no evidence of differential attrition, we will handle missing data by imputing the mean response on the dependent variable based on the participants' condition. If there is evidence of differential attrition, we will conduct two sensitivity analyses to test whether results are consistent: (1) impute the overall mean of the DV across conditions (2) enter attrition rate as a covariate in the main models of our primary analyses (e.g., inverse probability weighting).

Responses to Reviewer #4

R4.1 Considering a new control condition

Reviewer #4 was generally satisfied with the revisions and mentions that it could be possible to introduce a new control condition that is represented by Bluesky's default "explore" feed: *"I had two main concerns with the previous version of the proposed study: 1. External validity 2. Contribution relative to the existing literature. Regarding 1. This concern is entirely resolved, since the authors now plan to recruit active Bluesky users. Now that the authors do recruit active Bluesky users, they could think about including a control arm that does business as usual – though I don't think this is critical."*

We agree that a "business as usual" control condition would be valuable. In a world in which recruitment were straightforward and robust, we would be inclined to include it. But given that there are no guarantees on recruitment for an entirely novel study of this sort, we are wary of adding an additional condition. And as we consider

which control condition is best, we favor the one we have proposed over the business as usual option (a preference that the reviewer may well share).

Our preferred control condition aligns with the one used in previous work: the reverse-chronological feed condition. If our Stage 1 manuscript is accepted, based on Reviewer #4's comment, we would plan to discuss the possibility of using different Bluesky default feeds in the discussion section of the Stage 2 manuscript.

R4.2 Recruiting Bluesky users

Reviewer #4 notes that it is more difficult to recruit Bluesky users: *"It is a lot more difficult to recruit active users than to rely on simple Prolific convenience samples of non-users. Therefore, it was reassuring to read that the authors are aware of this issue and have a backup plan to screen Prolific participants based on their use of Bluesky."*

We are glad Reviewer #4 appreciated our "back-up" plan, and as described in **R2.4**, we further updated the manuscript with other contingency plans because we are aware of the potential difficulties in recruiting active users. In the revised manuscript we update recruitment efforts with a multi-stage strategy until we hit our target of 2000 participants. We describe this on pg. 7 of the manuscript:

> To provide time to pre-screen participants, starting in August, 2024 we will recruit participants for a study on "Testing New Bluesky Feeds" via advertising on Bluesky, and also on Reddit (e.g., on r/BlueSkySocial, which has 15.4k active users) and X, both of which contain active Bluesky users. From advertising we will aim to collect 3000 user responses on a brief demographic survey with the aim of recruiting 2000 study participants (see Sampling Plan below). If advertising on Reddit and Twitter fail to secure 3000 survey responses, as a backup we will work with Prolific and Connect consultancy services to recruit participants on who are active Bluesky users (correspondence with Prolific consultancy services confirmed that they can add Bluesky use as a screen to their current pool of participants, and Connect already has at least 900 active Bluesky users). To further boost the recruiting if required, we will also work with Bluesky developers to post a link to our study on the official Bluesky developer accounts (accounts with 100k to 1M followers).

We note that Bluesky developers have verbally agreed to post our study advertisement on their official accounts (accounts with 1 million to 100k followers), and the advertisement would specify that our study does not entail any collaboration with Bluesky or its employees. With these plans in place, we are confident we can get the required sample size.

R4.3 Including "harder" measures collected outside surveys

Reviewer #4 suggests that we include behavioral measures outside the items measured in the weekly surveys: "*Regarding 2.:Relying on native users, thereby elevating external validity, also improves my perception of the contribution relative to the literature. However, in the new version, the authors put much more weight on "soft" outcomes related to perceptions about others' beliefs and behaviors. These outcomes are "lower hanging fruit", because they are easy to measure and likely easier to move (hence the experiment can be done with a smaller sample), than the ultimately arguably more relevant outcomes of voting behavior, partisan attitudes, and behavior on the platform itself (i.e. what people post). They could also be more strongly driven by experimenter demand effects. So, I am not 100% on board with down-weighing these other outcomes. I have two considerations here: 1) So long as the other outcomes are included too, albeit not as primary outcomes, I suppose readers like me can scroll down to the relevant appendix section*". Reviewer #4 goes on to suggest, "*b) complement the outcomes with "harder" measures collected outside of the surveys: e.g., publish posts to users' feeds and measure whether they click on them/engage with them; publish poll questions to users' feeds over the course of the study period; send "friend" requests to participants, varying the party affiliation of the requestor."*

We share the reviewer's enthusiasm for "harder," behavioral measures. Indeed, one of several of our primary analyses (RQ3) includes measuring the impact of condition on user platform behavior: engagement with IME content which includes liking, sharing, or posting IME content (which we measure with our in-house machine-learning classifiers described in SI Appendix, 3.3.3, or pg. 46 of the manuscript). We clarified this part of the procedure in the revised manuscript, in part by adding a "behavioral measures" section to our summary of measures on pg. 24 of the manuscript (Table 2).

We also like Reviewer #4's idea of tracking the political ideology of users whom participants choose to follow. We now include that behavioral measure as well by analyzing the content of the followed-user's feed with our LLM ideological lean classifier we developed for this project (see pg. 46 of the manuscript). We added this as an exploratory measure in Table 2 as well.

As Reviewer #4 suggests, we will also include analyses of condition on voting behavior and partisan animosity in the exploratory analyses section of the Stage 2 manuscript, all of which are listed as exploratory measures starting on pg. 33 of the manuscript.

Northwestern University
2211 Campus Drive
Evanston, IL, 60208

R4.4 Reducing experimenter demand effects

Reviewer #4 would like us to confirm that participants will not exactly know what condition feed is designed to do: *"To deal with experimenter demand effects better: a) keep the treatment arm assignment as opaque as possible, i.e. don't be explicit in disclosing how exactly users' feeds will change"*

We agree that an opacity of condition details is important to reduce demand. We edited our participant instructions with the following language to ensure opacity, while also giving them enough details for them to understand the context of the experiment (pg. 25):

> Welcome to our research study on Bluesky social media use. This study requires you to use Bluesky in certain ways for 8 weeks. We are interested in how different feeds impact your experience of the platform. You may notice your feed seems different than your previous default, or you may not. Please simply use Bluesky as you normally would. Your feed will be based partially on your own friends network and platform behaviors just like on other social media platforms (e.g. X, Instagram, Facebook). At the end of the experiment, we will provide you with full details of the feed and provide you with an option to continue using it, as well as open-source code if you are interested in the technical details.

Responses to Reviewer #2

R2.1 Clarifying decisions for user screening and attrition analyses

Reviewer #2 asks whether our decision to screen users who did not complete enough surveys or log into Bluesky enough was pre-registered: *"This paper analyzes the effect of different Bluesky algorithms on the posts users see in their feeds, their perceptions, and the posts they engage with. Overall, the authors find that a reverse-chronological feed decreases exposure to moral, emotional, toxic, and political content, compared to an engagement-based feed. The diversified extremity algorithm they created mostly had the opposite effects. They also find that reverse-chronological feeds decrease perceptions of partisan animosity and users' own feelings of animosity (the effects on other perceptions were less clear). Finally, they do not find an effect on users' own engagement behavior. Overall, the paper is executed well, mostly follows the pre-analysis plan, and makes an important contribution to the literature! Since this paper was accepted based on the plan, I will focus on the results, the exploratory analyses, the changes, and the writing, and less on the importance or broader contribution of the paper. I. Changes from pre-registration 1. The authors exclude users who did not complete enough surveys or did not log in to Bluesky often enough. I do not think this was pre-registered, and it is important."*

We appreciate your attention to potential deviations from the Stage 1 Registered Report protocol. In this case, the exclusion criteria for minimum survey completion and platform activity were preregistered in the Stage 1 protocol (to avoid any confusion, we now host the final approved Stage 1 protocol at the following link: https://osf.io/k2crm/files/c9a3m). Specifically, we stated that we would remove users who did not show survey or login activity for 75% of the study. From pg. 10 of the accepted Stage 1 manuscript:

> After the pre-screening, we will only exclude participants from data analysis who fail to complete 75% of the weekly surveys (i.e., miss more than 2 weeks of the surveys) or do not engage with the platform once / day for at least 45 of the 60 days of the experiment. For participants who are above the 75% activity threshold but miss surveys or engagement days, they will be kept in the study while treating their missing responses as missing data.

In Stage 2, we implemented these criteria with a protocol amendment: to preserve statistical power, we lowered the minimum completion/activity threshold from 75% to

63% (5 of 8 weeks) with editorial approval (Stage 2 manuscript, p. 11). As reported in the manuscript:

> We excluded participants from data analysis who failed to complete 63% of the weekly surveys (i.e., missed more than 3 weeks of the surveys (*n* = 359) or did not engage with the platform for at least 5 of the 8 weeks of the experiment (*n* = 166). For participants who were above the 63% activity threshold but missed surveys or engagement weeks, they were kept in the study while treating their missing responses as missing data. This criterion reflects an editor-approved deviation from our preregistered 75% threshold, adopted to improve statistical power. Importantly, all substantive conclusions are robust to alternative exclusion thresholds (see SI, Section 9.0).

Participants above the threshold were retained and missing responses were treated as missing data (as preregistered). Nonetheless, we agree that even with relaxed exclusion criteria it is important to test the robustness of results with fewer exclusions, see **R2.2** below.

R2.2 Further analyses of attrition

Reviewer #2 wonders whether there is differential attrition and suggests we conduct an analysis that is less strict for platform usage: *"Excluding users could lead to differential attrition. The authors argue in Table S38 that there is no differential attrition, but I am not sure that is the case. There is more attrition in the reverse-chronological condition compared to the two treatments that promote posts users are likely to engage with. This pattern is consistent with Guess et al. (2023a), who find that people spend dramatically less time on Facebook and Instagram when assigned to the reverse-chronological feed. Although the differential attrition is not large and likely not driving the results, the authors should check more closely. Specifically, they could relax the requirement of engaging with the platform for 5 of 8 weeks and examine how that affects the results."*

Thank you for raising this point. As preregistered in our Stage 1 protocol, we assessed differential attrition across experimental conditions using a chi-square test and found no statistically significant evidence that attrition differed by condition ($\chi^2 = 4.122$, df = 2, p = .127). We revised the Stage 2 manuscript language accordingly to state that there was no evidence of significant differential attrition, rather than implying that attrition was identical across conditions. From pg. 11 of the revised manuscript:

> Missing data were handled depending on whether missing data varied significantly by condition ('differential attrition'). We tested for differential

attrition using a chi-square test, where a positive sign of differential attrition was determined by Ns that are significantly different as a function of condition. There was no evidence of significant differential attrition ($\chi^2 = 4.122$, df = 2, $p = .127$), so we handled missing data by imputing the mean response on the dependent variable based on the participants' condition, which was a preregistered decision (see SI, Section 7.1 and 9.0 for details of differential attrition analysis and further robustness tests of all RQ analyses where exclusion criteria are relaxed).

Nevertheless, it is true that attrition is descriptively highest in the reverse-chronological condition; although the difference is small, it is consistent with prior evidence that reverse-chronological feeds reduce platform use. We therefore conducted additional robustness analyses as you suggested. Specifically, we re-estimated all analyses from the Stage 1 protocol under progressively relaxed survey and platform-activity thresholds (including thresholds below the preregistered ≥5/8-week criterion). These analyses are reported in a new SI section dedicated to differential attrition and compliance robustness checks (SI Section 9.0). Across these specifications, the substantive conclusions are unchanged: effect directions and magnitudes remain highly similar, and statistical inferences are consistent with those reported in the main text. As an example, we display the results of the robustness checks you suggested for RQ1, toxic content here:

Robustness across inclusion thresholds: Toxic content

Entries are β coefficients; bold indicates p < .05.

	Pre-election			Post-election		
Inclusion threshold	DE – RC	EB – RC	EB – DE	DE – RC	EB – RC	EB – DE
≥1 weeks (n=1824)	-0.00065	**0.00824**	**-0.00889**	**0.00658**	**0.01127**	**-0.00469**
≥2 weeks (n=1766)	-0.00063	**0.00824**	**-0.00888**	**0.00650**	**0.01125**	**-0.00476**
≥3 weeks (n=1728)	-0.00061	**0.00827**	**-0.00888**	**0.00647**	**0.01126**	**-0.00479**
≥4 weeks (n=1636)	-0.00056	**0.00829**	**-0.00885**	**0.00651**	**0.01127**	**-0.00476**
≥5 weeks (n=1475)	-0.00053	**0.00822**	**-0.00876**	**0.00641**	**0.01092**	**-0.00451**

Table S89. Effect of treatment condition on content using progressively relaxed exclusion criteria (intergroup content). The table displays the coefficients for each treatment (diversified extremity algorithm vs reverse chronological, DE - RC; engagement-based algorithm vs. reverse chronological, EB - RC) at different levels of exclusion criteria. Valid weeks were defined as the number of weeks participants logged in and filled out a survey. Bold values represent significant values at p <.05.

R2.3 Compelling exploratory analyses

Reviewer #2 found several exploratory analyses compelling: *"Other changes in the plan—such as analyzing binary outcomes in RQ3—made sense to me. I also found the exploratory analyses sensible and the following particularly interesting • Analyzing the effect on moral outrage content and constructive content (RQ1, H2). • Analyzing the effect on enjoyment with the feed (RQ2, H1). • Studying whether algorithmic amplification affects political polarization. • Analyzing how skewed engagement with toxic political content is."*

Thanks for the positive feedback, and we are glad you found our exploratory analyses interesting!

R2.4 Questions about one-sided hypothesis testing

Reviewer #2 asks for clarification about the pre-registered decision not to adjust for multiple hypotheses for direction predictions, and also requests clarification about whether one-sided tests were conducted: *"Comments on the analysis. I did not completely understand the paragraph arguing that adjusting for multiple hypotheses is not needed because the authors pre-registered directional hypotheses. I do not see why directional predictions avoid the multiple-tests problem. However, this paragraph did appear in the pre-registered version, and I probably should have mentioned this then (the number of tests did increase substantially because of the new election-interaction terms). More importantly, the paper does not seem to be analyzing one-sided tests. For example, it states that before the election, engagement-based feeds served less constructive content than reverse-chronological feeds and after the election, engagement-based feeds served more constructive content. The authors treat both directions as evidence of effects rather than simply confirming or rejecting a one-sided hypothesis. In my opinion, this is the correct approach; however, it is not clear that they can argue they are testing only one direction in their hypotheses."*

Thank you for raising this point. We agree that preregistered directionality does not by itself remove the multiple-testing issue. In the Stage 1 protocol, we argued based on prior literature that for confirmatory, preregistered hypotheses with *a priori* directional predictions, we reduce researcher degrees of freedom (e.g., post hoc selection across outcomes or directions) and prioritize power for these planned tests given the well-known Type I/Type II tradeoff in familywise adjustments (Matsunaga, 2007). This is particularly important in our case based on the structure of the research questions being evaluated. The confirmatory hypotheses concern patterns of results across sets of IME outcomes, rather than treating statistical significance on any single outcome as

sufficient evidence for each research question hypothesis. Under this evidentiary logic, transparently reporting all planned tests without multiplicity adjustment is appropriate, as the interpretation depends on consistency across outcomes rather than isolated findings.

Accordingly, in our Stage 2 we used two-sided statistical tests throughout and reported two-sided *p*-values. Following our preregistered directional predictions, we interpret results as confirmatory support only when effects fall in the predicted direction; statistically significant effects in the opposite direction are treated as disconfirming and are not counted as evidence for the hypothesis. This approach avoids post hoc selection of direction while remaining agnostic to unexpected effects.

By contrast, for outcomes designated as exploratory (including constructive content) we fully agree that effects should be interpreted two-sided and adjusted for multiple comparisons. Accordingly, we now apply Holm–Bonferroni corrections within each research question for all exploratory analyses. Our substantive conclusions are unchanged after this correction.

We now further clarify this issue on pg. 11 of the Stage 2 manuscript:

> Following our Stage 1 plan, we did not adjust alpha for multiple tests for confirmatory, preregistered hypotheses with a priori directional predictions. For these planned tests, our goal was to minimize researcher degrees of freedom (e.g., post hoc selection across outcomes or directions) while retaining power, acknowledging the Type I/Type II tradeoff inherent in familywise error corrections (for a review, see [47]). We used two-sided tests, but we interpret results as confirmatory support only when effects fall in the preregistered direction; statistically significant effects in the opposite direction are treated as disconfirming (and are not counted as evidence for the hypothesis). However, for outcomes designated as exploratory, we interpret effects two-sided and control for multiple testing using Holm–Bonferroni within each research question.

Regarding the exploratory outcome of constructive content specifically, see **R2.10** below.

R2.5 Reporting both accuracy and raw norm perception outcomes

Reviewer #2 suggests that we report both accuracy and raw norm perception variables because accuracy was not preregistered: *"I found the analysis of descriptive norms variables in RQ2 confusing, and I only understood it better after reading the discussion at the end of the paper. It is not always fully clear what the outcome variable is in the text. The pre-registration discusses perceptions of IME and toxic content, but the text mostly*

discusses accuracy. I would analyze the effect on perceptions and the effect on accuracy as two separate outcomes and clearly distinguish between them. This analysis is provided in Table S59, but that table is hard to interpret (as I discuss below), and the main text does not discuss its results. More importantly, I suggest replacing Figures 2–3 with a figure similar to Figure 1 (which is excellent), so the treatment effects of each treatment on each outcome are crystal clear. Of course, there could be other solutions, but currently this section should be rewritten to be clearer and more transparent."

Thank you for raising this point and providing suggestions for how we can be clearer about reporting the results of the accuracy analysis as well as visualizing those results. We preregistered examining accuracy in response to reviewer comments from the Stage 1 revision, and we believe that the ability to examine accuracy is a key contribution of our study design. From the approved Stage 1 protocol (pgs. 5-6):

> We will also test a related research question about the direct social consequences of algorithmic amplification: does the amplification of IME content cause people to misperceive social norms around political dialogue on Bluesky? We plan to test a specific version of this question that asks whether participants misperceive the base rate of IME content in their social network, and a general version of this question that involves whether participants in the algorithm-mediated conditions misperceive political dialogue on Bluesky to contain more toxicity than its true base rate of toxic political dialogue. We also test whether conditions impact perceptions of partisan animosity in the U.S.:
>
> *RQ2, H1*: Compared to a reverse-chronological newsfeed, an engagement -based algorithm will significantly (and inaccurately) increase participants' perceptions of the base rate of IME content in their online social network, their perceptions of toxic political dialogue on Bluesky, and their perceptions of partisan animosity in the U.S.
>
> *RQ2, H2*: Compared to an engagement-based algorithm, a diversified extremity algorithm will significantly decrease (toward accuracy) participants' perceptions of the base rate of IME content in their online social network, their perceptions of toxic political content on Bluesky, and their perceptions of partisan animosity in the U.S.

However, we noticed that we did not report how we would specifically compute accuracy in the analysis section of the Stage 1 registered report, which we assume is what led to the confusion. We apologize for this oversight. In response, we made four changes to improve the clarity in the reporting of RQ2:

1. *Main-text rewrite for clarity.* We rewrote the RQ2 Results section to clearly define descriptive and prescriptive norm accuracy and how they relate to participants' raw perceptions (pgs. 20–23).

2. *New figures mirroring Fig. 1***.** We replaced the prior presentation with two coefficient plots (like Fig. 1) that make treatment effects crystal clear: one figure for the accuracy outcomes and another for perceived network polarization (new Figs. 3–4). We paste Fig. 3 below for your convenience.

3. *Explicit reporting of raw perceptions.* We added a full set of models and a narrative write-up for the raw perception variables in the SI (Section 7.5), alongside a parallel coefficient plot for those outcomes.

4. *Improved SI tables.* We revised the relevant SI table(s) so they are easier to interpret based on your comment in **R2.7.**

We hope these revisions help further clarify the RQ2 accuracy results. If the editor wishes, we can also report the raw perception results write-up in the main text.

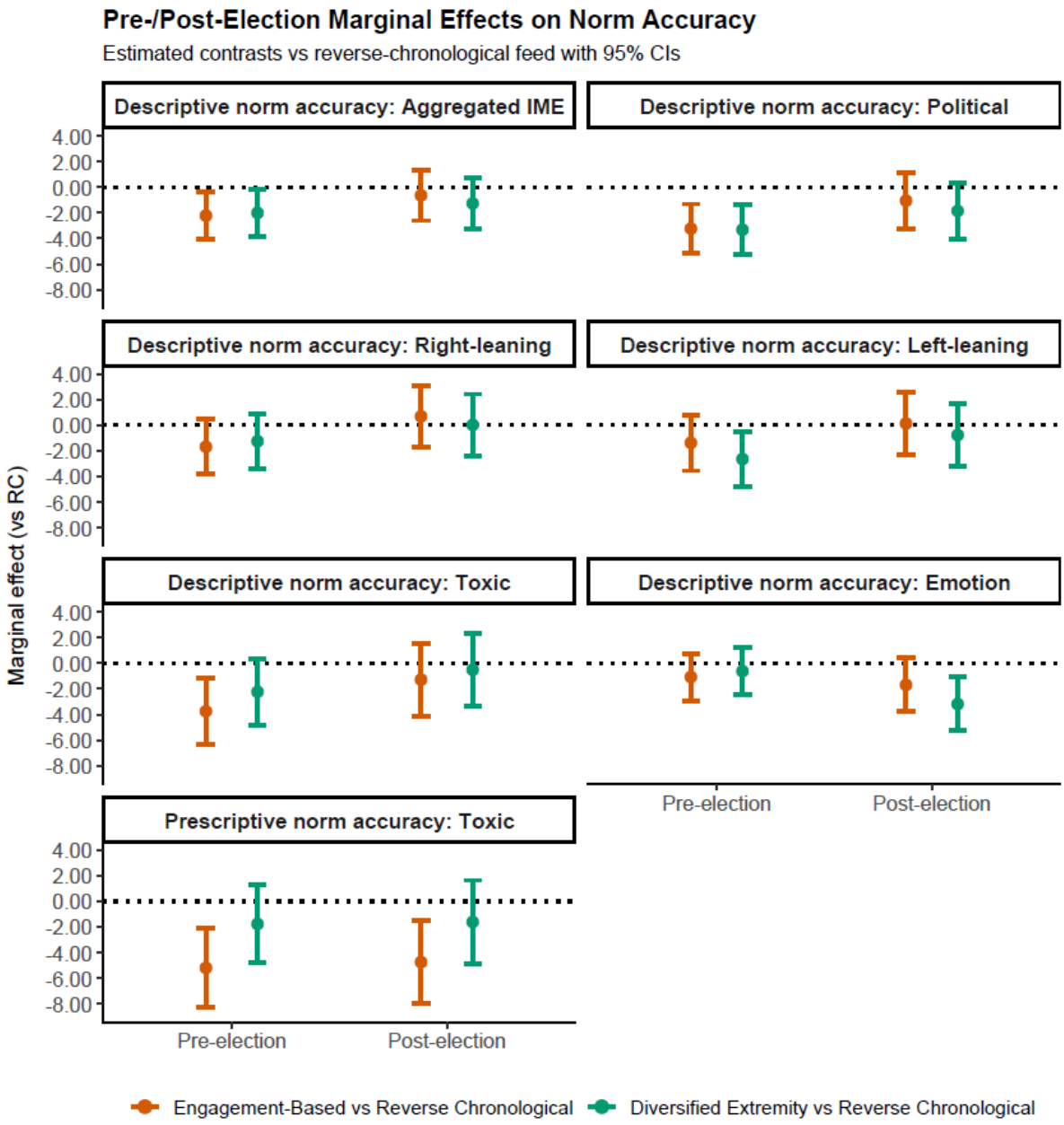


Fig. 3. Effects of engagement-based (EB) and diversified extremity (DE) algorithm feeds compared to a reverse-chronological (RC) feed on norm accuracy before and after the 2024 U.S. Presidential election. The plot displays marginal effects of EB vs. RC and DE vs. RC on descriptive norm accuracy (political, right-leaning, left-leaning, toxic, and emotional content), prescriptive norm accuracy, and perceived network polarization. Descriptive accuracy outcomes represent the difference between participants' perceived prevalence of content and observed platform baselines, and prescriptive accuracy represents the difference between perceived and actual norms of content appropriateness based on judgments from participants in our sample. Values greater than 0 indicate increased overestimation (and values less than 0 indicate increased underestimation) relative to baseline for accuracy measures, compared to the reverse-chronological feed.

R2.6 Reporting precision estimates for null effects

Reviewer #2 suggests that we discuss precision estimates on null effects reported in the manuscript: *"The authors find some null effects on behavior and then include an interesting discussion at the end of the paper regarding these effects. I think it is important to discuss the precision of these effects. In other words, what effects can the authors rule out? For example, they could specify that they rule out a decrease in the likelihood of engaging with toxic content greater than X%. This will help the readers understand whether they do not find that the change in exposure did not map to a change in engagement because of selective behavior among Bluesky users or due to limited statistical power."*

We appreciate your suggestion and agree that clarifying the precision of the null behavioral effects is important for interpreting whether these findings reflect limited statistical power or instead rule out substantively large changes in engagement.

To address this concern, we added a new subsection in the RQ3 section of the main text ("Contextualizing null effects on engagement") that goes beyond statistical significance and explicitly quantifies what magnitudes of engagement change are incompatible with the data. Specifically, we translate the 95% confidence intervals from our mixed-effects logistic models into bounds on proportional changes in engagement likelihood relative to baseline rates. This allows us to state which sizes of behavioral effects can be ruled out, rather than simply noting that effects are not statistically significant. We report a summary of this analysis on pg. 28-29 of the main text:

> As described above, engagement with IME content was rare overall and highly skewed across users. Given this rarity, the absence of statistically significant effects on engagement raises a key interpretive question: whether these null findings primarily reflect limited statistical power, or whether they indicate that changes in algorithmic exposure did not translate into substantively meaningful shifts in engagement behavior, an interpretation that depends on the extent to which large effects can be confidently ruled out.
>
> To address this question, we examined the precision of the estimated algorithm-condition effects by translating their confidence intervals into bounds on the magnitude of engagement changes that are compatible with the data. Rather than focusing solely on statistical significance, this approach allows us to assess what size behavioral effects can be ruled out. Across intergroup, moral, emotional (positive and negative), and sociopolitical content, the estimates were sufficiently precise to rule out moderate-to-large proportional changes in

engagement likelihood. Specifically, depending on content category and contrast, the data rule out changes on the order of approximately 50–70% decreases or 100–175% increases relative to baseline engagement rates (the one exception was toxic content, which had smaller bounds due to its rarity, see SI, Section 8.7 for more details and tabulated results). Thus, while smaller effects remain plausible, the observed null effects are inconsistent with very large immediate shifts in engagement behavior.

At the same time, these bounds should be interpreted with caution. Engagement outcomes were rare and observed over a relatively short time window, and many participants were early or infrequent users of the platform. Accordingly, while the present findings suggest that algorithmic changes to exposure did not produce large, immediate shifts in engagement behavior in this setting, different samples, platforms, or time horizons could plausibly yield different behavioral consequences (see Discussion).

We also describe the full computational details of this precision analysis in a new SI section (Section 8.9) and report the resulting bounds in a new table (Table S88). We believe these additions directly address your concern by clarifying what the null effects do and do not rule out, while remaining consistent with the broader interpretive framework of the paper.

R2.7 Improving appendix table presentation

Reviewer #2 suggests ways to improve the clarity of effects presented in our appendix tables: *"Comments on writing: Many appendix tables are hard to interpret because the reader must add coefficients and cannot easily tell, for example, whether each treatment is significant after the election. In my opinion, all of these tables should convey the same information that is conveyed in Table 1. They should show four numbers for each outcome: the effect (and standard error or confidence interval) of the two treatments compared to a reference group (reverse chronological in Figure 1, though engagement-based could also be the reference), before and after the election. The table could also note whether the difference between the two treatments is significant."*

Thanks for this suggestion. For all results reported in the main text, we updated the SI Tables based on your specifications. Each table presents the contrasts of interest for each election period, displaying the coefficient and standard error, as well as bolding to

designate where the effect is significant. We appreciate your suggestion and believe our tables are now highly interpretable.

RQ1: Toxic content — Pre/Post model Entries are β (SE). Bold indicates p < .05.			
	EB – RC	DE – RC	EB – DE
Pre-election	**0.008 (0.000)**	-0.001 (0.000)	**0.009 (0.000)**
Post-election	**0.011 (0.000)**	**0.006 (0.000)**	**0.005 (0.000)**

Table S16. Effects of feed algorithm condition on toxic content before and after the election. Entries report estimated differences in the proportion of toxic content between algorithm conditions from the pre/post-election model, expressed as regression coefficients (β) with standard errors in parentheses. Comparisons are shown for engagement-based (EB) versus reverse-chronological (RC) feeds, diversified extremity (DE) versus RC feeds, and EB versus DE feeds, separately for the pre-election and post-election periods. Bolded coefficients indicate statistically significant differences (p < .05).

R2.8 Well-written discussion

Reviewer #2 feels positively about our discussion: *"The discussion is well written and interesting. I also agree that this paper makes an important methodological contribution, as the authors state: "Our study also represents an example of an industry-independent model for studying algorithmic amplification at scale."*

Thanks for the praise!

R2.9 Clarifying effects on constructive content

Reviewer #2 pushes us to explain the potentially surprising effects found for constructive content: *"The effect on constructive content is not clear. The authors first*

say "By contrast, diversified extremity feeds served less constructive content than engagement-based feeds both before and after the election, with a similar magnitude of difference across periods," but later say "In contrast, our diversified extremity algorithm ... while preserving or even enhancing exposure to more positive and constructive dialogue." Is the effect on constructive content positive or negative? Based on Figure 1, I think the former sentence is correct and the latter should be reworded. More generally, it is surprising that diversified extremity feeds served less constructive content, since the algorithm was designed to promote exactly that content. Can the authors explain this result?"

We appreciate the important points you raise here about an imprecise sentence in the Discussion and about the seemingly counterintuitive result for the exploratory constructive content outcome.

You are correct that, as shown in Fig. 1, diversified extremity (DE) served slightly less constructive content than the engagement-based (EB) feed (both pre- and post-election). Our original Discussion phrasing inadvertently implied the opposite. In the revised manuscript, we rewrote that sentence to align precisely with the results. Notably, this sentence now refers to positive emotional content (where DE does show relative improvement post-election). It now reads (pg. 30):

> In contrast, our diversified extremity algorithm, which was designed to reduce the influence of non-representative, extreme users, in practice reduced exposure to toxic, moralized, and highly negative content while preserving or even enhancing exposure to more positive emotional content, particularly after the election.

However, why would DE show less constructive content if part of its scoring involves upranking constructive content? We agree this result is surprising at face value, and we investigated several mechanisms that help explain it.

1. *Overlap between "constructive" and "outrage" in the classifier labels.*
   First, our classifier reveals a meaningful overlap between posts labeled as "constructive" and those labeled as expressing moral outrage. Using Perspective API thresholds (probability > .50), approximately 10% of posts classified as constructive were also classified as moral outrage. Substantively, these posts are not misclassified noise, but rather content that combines reasoned argumentation with morally condemnatory language. For example, we observed posts that articulate a solution or policy proposal while simultaneously condemning perceived wrongdoing.

Because engagement-based (EB) ranking disproportionately amplifies outrage-related signals, it is more likely to elevate this hybrid "constructive-with-outrage" subset than either the diversified extremity or reverse-chronological feeds. In contrast, DE is designed to reduce the influence of highly extreme and moralized users, which in practice also dampens exposure to this hybrid content, even when it contains constructive elements. As a result, EB can appear slightly higher than DE on the exploratory constructive content measure, not because DE suppresses constructive discourse per se, but because EB selectively amplifies constructive content that is embedded within moral outrage posts.

Consistent with this interpretation, when we reran the constructiveness analysis after excluding posts that simultaneously met the thresholds for constructive content and moral outrage, the EB–DE difference in constructive content was strongly attenuated. Moreover, EB no longer showed any increase in constructive content relative to the reverse-chronological feed after the election (see SI, Section 6.10; Table S38; Fig. S42). This pattern suggests that the apparent advantage of EB on constructive content is largely driven by outrage-linked constructive posts rather than by broader exposure to solution-oriented or cooperative discourse.

2. *Downranking superposters removed some highly active accounts that post "constructive" content*. A key design feature of the diversified extremity (DE) algorithm is to reduce the influence of superposters, because highly active accounts disproportionately contribute toxic and extreme content. One side effect of this design is that it also downranks a subset of highly active accounts that contribute relatively large volumes of constructive content (for example, commentators who post frequent, reasoned political threads).

   Importantly, this effect does not require that the same accounts be responsible for both high toxicity and high constructiveness. Rather, reducing the influence of the most active accounts mechanically trims both ends of the activity distribution, removing highly prolific sources of content regardless of whether their output skews primarily toxic, primarily constructive, or a mixture of both. In exploratory analyses, we found that superposters—defined purely by posting volume—exhibited higher average levels of constructive content than non-superposters, even as they also contributed disproportionate amounts of toxic content overall. As a result, the superposter filter can reduce the aggregate prevalence of constructive content in DE relative to EB, even though DE explicitly up-ranks constructiveness conditional on posts remaining in the candidate pool.

Because these analyses help explain the surprising results, we added a new section (SI Appendix, Section 6.10) that contains: (i) this explanatory discussion, (ii) the overlap analysis of constructiveness and outrage including a table of illustrative examples, (iii) a new superposter analysis. We also added a brief pointer in the main text near the exploratory constructiveness results to contextualize why EB can look higher than DE on this measure, despite DE's constructiveness upranking (pg. 18):

> The constructive content findings however appear to be driven by unexpected overlap in outrage and constructive content dialogues driven by political arguments that use both reasoning and outrage expression (for a full exploration of this issue, see SI Appendix, Section 6.10).

Finally, we mention these outcomes in the discussion since they highlight a practical lesson for algorithm design regarding optimizing multiple objectives on pg. 32 of the discussion:

> More broadly, our exploratory constructive dialogue results highlight an important design limitation for future algorithmic interventions: optimizing multiple objectives can produce unintuitive interactions when content labels are not orthogonal (e.g., constructive dialogue can co-occur with moral outrage) and when upstream filtering decisions (such as superposter downranking) alter which "constructive" content remains eligible for reranking (See SI Appendix, Section 6.10).

R2.10 Changing "first" language when citing previous work

Reviewer #2 suggests that we reword our statement that Guess et al. (2023) was the "first" experimental manipulation of algorithmic amplification: *"Literature. In the introduction, the authors write, regarding Guess et al. (2023a): "This study is foundational to the literature, as it was the first experimental manipulation of algorithmic amplification conducted with a broad swathe of the American populace on a real social media platform." I am not sure this is accurate; for example, see Huszár et al. (2021)."*

Thanks for the suggestion. We changed the phrasing to remove the claim of primacy, and also cited Huszar (2021) on pg. 4 of the introduction:

> This study is foundational to the literature, as it served as a key experimental manipulation of algorithmic amplification (also see[23]) conducted with a broad

swathe of the American populace on a real social media platform. Despite its importance, the study had key limitations[24].

R2.11 Discussing treatment of algorithms in prior work

Reviewer #2 suggests we cite previous work that changed features in algorithms: *"They also say, "First, the researchers did not control how the content algorithms worked, making it unclear which specific features of the algorithms created content differences between conditions." Two other experiments run at the same time did change features in the algorithm (Nyhan et al., 2023; Guess et al., 2023b). The text regarding these two commented appeared in the previous pre-registered version, and I apologize for only noticing and mentioning it now."*

Thanks for the suggestions. We updated the introduction to include the two recommended citations, on pg. 4:

> First, the researchers did not control how the content algorithms worked, making it unclear which specific features of the algorithms created content differences between conditions (for other examples with more algorithm control see[25,26]).

R2.12 Citing Braghieri et al. (2025)

Reviewer #2 suggests that we cite a study that found similar results as our analysis of user satisfaction: *"Finding that reducing negative content does not come at the expense of user satisfaction is interesting! To some extent, consistent with Braghieri et al. (2025)."*

Thanks for this interesting citation. We agree it is relevant and added it to our discussion of interventions being aligned with user experience on pg. 30:

> It is typically assumed that reducing negative content that typically drives Engagement necessarily comes at the expense of user satisfaction, yet our data suggest that this trade-off may be overstated. Instead, the results suggest that algorithms designed to diminish the influence of non-representative, extreme users may represent a scalable and realistic path for platforms: they can curb the amplification of divisive content while maintaining, and in some cases even

enhancing, the user experience of the platform as a whole—a balance critical to both societal well-being and platform sustainability (see also[52]).

R2.13 Clarifying discussion about behavioral effects

Reviewer #2 suggests that Guess (2023) may not be the best citation when discussing the lack of algorithmic impact on engagement behaviors: *"In the discussion, the authors write: "At face value, these findings could be interpreted as evidence of limited downstream behavioral consequences of algorithmic amplification, consistent with recent experimental work suggesting limited effects of algorithmic feeds on user political attitudes and behaviors." They cite Guess et al. (2023a), but that paper did find an effect on on-platform measures of political engagement. The broader point is that previous literature has found that consumers tend to click what appears in their feeds. It is completely fine that this paper did not find such an effect, and the authors discuss potential reasons in the discussion."*

Thanks for the clarification on Guess et al. (2023). We removed the citation in that sentence so it only speaks to our findings, which are explained in the rest of the discussion paragraph on pg. 31:

> Regarding behavioral consequences, we found no evidence that algorithmic amplification directly influenced users' own engagement behaviors. Engagement-based feeds did not increase the likelihood of users engaging with intergroup, moral, emotional, political, or toxic content compared to reverse-chronological feeds, and diversified extremity feeds did not significantly reduce engagement relative to engagement-based feeds. At face value, these findings could be interpreted as evidence of limited downstream behavioral consequences of algorithmic amplification.

Responses to Reviewer #3

R3.1 Clarifying the election shock

Reviewer #3 would like us to clarify why the 2024 U.S. Presidential election should be considered a shock to political content in social media feeds: *"Summary: This paper presents findings from a field experiment that randomizes social media feed-ranking algorithms among U.S. Bluesky users during the 2024 presidential election. Participants were assigned to either a reverse-chronological feed, an engagement-based feed (similar to those used by major platforms), or a "diversified-extremity" feed that downweighted content from extreme, toxic, or overly active users. The main results are that engagement-based algorithms amplified intergroup, moralized, emotional, toxic, and political content relative to a chronological feed but that this amplification had limited effects on users' engagement behaviors and somewhat mixed effects on perceived norms. Evaluation: I enjoyed reading the manuscript and this is a very impressive experiment. Below I give some comments that I think will help it improve. I would like to see more discussion about how the authors think of the election as a shock. Currently, they mention that "The preregistration did not anticipate modeling the election as an exogenous shock. Because the election was likely to affect our politically relevant outcomes […]" but I didn't understand 1) what is the election shocking (a shock to what)?, 2) why this shock was supposed to affect outcomes, and 3) why this shock would affect outcomes differentially across experimental conditions (to justify adding this non-preregistered dimension of heterogeneity)."*

Thanks for pointing this out. First, we apologize for the confusion we created regarding what was preregistered in our approved Stage 1 protocol. The final, approved Stage 1 protocol *did* include the revised study dates (pre-/post-election), and the analysis plan explicitly noted the possibility of estimating pre- versus post-election models as a sensitivity analysis. You are correct that the previous Stage 2 manuscript inaccurately referred to an earlier draft of the Stage 1 protocol where we did not anticipate the pre/post election time period; we have corrected this error in the revised manuscript in the design section on pg. 7. To avoid any confusion, we now host the final approved Stage 1 protocol at the following link: https://osf.io/k2crm/files/c9a3m

The approved Stage 1 protocol also articulated why the election period should be treated as an exogenous shock in our experiment. In the design section (p. 7), we noted that the election would substantially increase the proportion of political content in the platform-wide post stream, which in turn was expected to make differences across

feed-ranking algorithms—particularly the diversified-extremity algorithm—more consequential for the content users were exposed to:

> *Furthermore, the election time period will make political content represent a greater proportion of posts[27,28], which is likely to make our diversified extremity algorithm have a greater impact on content.*

We hope that clarifying that this heterogeneity was preregistered addresses part of your concern. We also take your broader point seriously and therefore clarify why the election should be considered an exogenous shock in the first place.

In our design, the election constitutes an exogenous shock to the composition of the platform's content stream—most notably, a sharp increase in the base rate of political content and closely correlated intergroup, moralized, and emotional discourse. This shift is independent of our manipulation and changes the input that each feed-ranking algorithm operates on during the study window.

Our outcomes are directly linked to exposure and learning from that environment (e.g., perceived descriptive and prescriptive norms, partisan animosity). When the information environment becomes more political (as it does during an election) participants are exposed to more politically relevant material, and because political content is empirically intertwined with IME language during election periods, these outcomes are plausibly more responsive. Importantly, the "shock" is not to the algorithms themselves, but to the content they are ranking.

This matters because the three feed-ranking conditions are differentially sensitive to such a shift. Engagement-based ranking disproportionately amplifies engagement-relevant political and IME content (which we confirm in RQ1 analyses) when it becomes more prevalent; reverse-chronological feeds largely reflect the raw increase without additional amplification; and the diversified-extremity feed is designed to dampen the influence of highly active, extreme, or toxic sources, which is an intervention we expected to matter more precisely when political content comprises a larger share of the stream.

As a result, the same exogenous increase in political content can generate different exposure compositions and downstream effects across conditions. We now make this interaction between the election-driven shift in content and algorithmic ranking rules explicit in the revised manuscript (pg. 7):

> This time period still allowed us to run the experiment during a time when our research questions related to perception of norms around politics are most

meaningful to the extent they may be related to political behavior such as voting. Furthermore, the election time period made political content represent a greater proportion of posts[35,36], which we expected to make our diversified extremity algorithm have a greater impact on content. In other words, because the election induces a sharp, exogenous shift in the composition of the content stream—particularly the prevalence of political and IME content—we treat it as a shock that plausibly interacts with feed-ranking rules, which differ in how sensitively they amplify changes in the input distribution.

R3.2 Greater discussion of election effects

Reviewer #3 suggests that we offer greater explanation of pre/post election effects that were discovered in our results: *"-Relatedly, the paper sometimes presents results before/after the election without offering any explanation for why this is the case. For example, "Intergroup content moved in the opposite direction before the election (diversified extremity served more than engagement-based) and showed no difference after the election." After reading this type of sentences, the reader is left wondering what drives these differences."*

We appreciate this concern and agree that our original summary did not sufficiently explain the pre/post-election differences for RQ1, H2. We have revised the summary of these results to clarify which outcomes show stable directional effects versus which exhibit period-specific changes, and to offer a plausible explanation for the latter (pp. 15–16 of the main text). Specifically, we now write:

> *Summary (RQ1, H2)*. In line with *RQ1*, *H2*, for most outcomes, the direction of the diversified-extremity effect was stable across periods: diversified extremity consistently reduced aggregated IME content, as well as moral, negative emotional, toxic, and political content relative to engagement-based ranking, with the size of these reductions generally narrowing after the election (and slightly widening for political content). Two outcomes deviated from this general pattern. Intergroup content was higher under diversified extremity than engagement-based ranking before the election but did not differ after the election, and positive emotional content reversed across periods, with diversified extremity serving less than engagement-based ranking before the election but more after the election. A plausible explanation is that the election altered the supply and concentration of different content types across accounts. Before the election, intergroup language may have been more diffusely produced across many mid-engagement users, allowing diversification to increase intergroup

exposure relative to engagement-based ranking; after the election, intergroup content may have become rarer or more concentrated in the same high-engagement hubs, collapsing this difference. By contrast, post-election positive emotion may have shifted toward more community-dispersed expressions of solidarity, reassurance, and collective coping that diversification surfaced more effectively than engagement-based ranking.

R3.3 Explaining the use of model fit statistics as a decision rule

Reviewer #3 asks us to clarify our decision to use model fit statistics to determine whether the pre/post election indicator should be included in our model: *"Also relatedly, I didn't understand the logic of the following rule: "if the majority of outcomes showed better fit (AIC/BIC criteria and Log-likelihood ratio test) when including a binary pre-/post-election indicator and its interaction with condition, then we report that specification in the main text". Concretely, I don't see why not just report the pre-registered specification in the main text and relegate to the appendix 1) specifications adding controls for before/after the election, and 2) heterogeneity analysis of the main treatment effects before and after the election.*"

Thanks for bringing up this point. As clarified above (**R3.1**), the approved Stage 1 protocol anticipated the election as an exogenous shock to the content environment and explicitly mentions pre-/post-election analyses.

Our use of model fit statistics served a narrow and principled purpose: to confirm whether incorporating the election shock dimension anticipated in the Stage 1 protocol materially improved explanatory power across outcomes. Because the election induced a sharp shift in the content stream that plausibly interacts with feed-ranking rules (**R3.1**), excluding this dimension risks mis-specifying the data-generating process. Accordingly, we defaulted to estimating models that included a pre-/post-election indicator and its interaction with algorithm condition, and used model fit statistics to assess whether this theoretically motivated extension improved model specification. At the same time, we specified that if omitting the election-period terms resulted in a better-fitting model, we would report the base specification in the main text.

In practice, models including the pre-/post-election indicator and its interaction with condition consistently provided better fit (based on AIC, BIC, and likelihood ratio tests) for the majority of outcomes in RQs 1 and 2, indicating that election-period heterogeneity explained meaningful additional variance. By contrast, for RQ3

(engagement outcomes), election-period terms did not improve model fit, and we therefore report the base specification in the main text for those analyses.

For this reason, we foreground the better-fitting specification for each research question in the main text, while retaining full transparency and preregistration fidelity by (i) clearly flagging this decision in the manuscript and (ii) reporting the alternative specification in the SI. Importantly, substantive conclusions are consistent across specifications, and the SI allows readers to directly compare estimates with and without the election-period dimension.

We have revised the manuscript to clarify this logic and to make explicit that model selection was guided by Stage 1 expectations about the election as a shock (pg. 11-12 in the Analysis section):

> *Decision based on Stage 1 reported analysis plan.* As described in the Stage 1 protocol and reiterated in the Design section, the study spanned roughly 5–6 weeks before and 2–3 weeks after the 2024 U.S. presidential election (depending on wave). The approved Stage 1 protocol explicitly anticipated the election as an exogenous shock to the content environment and noted pre-/post-election modeling as a planned sensitivity analysis.
>
> Accordingly, we estimated models that included a binary pre-/post-election indicator and its interaction with algorithm condition for all outcomes. We then evaluated model fit using AIC, BIC, and likelihood ratio tests to assess whether incorporating election-period heterogeneity materially improved explanatory power for each research question.
>
> For RQs 1 and 2, models including election-period terms consistently fit the data better across the majority of outcomes, indicating that election-period heterogeneity captured meaningful additional variance. We therefore report these specifications in the main text and present the baseline models without election terms in the SI. By contrast, for RQ3 (engagement outcomes), election-period terms did not meaningfully improve model fit, and we therefore report the baseline specification in the main text and present election-period models in the SI. Substantive conclusions are consistent across specifications.

R3.4 Satisfactory decisions to deviate from preregistration plan

Reviewer #3 approves of the two deviations from the preregistration we reported: *"Other deviations from preregistration such as the different sampling scheme or time aggregation seem fine to me.*

We are glad you agree with our decisions.

R3.5 Using aggregate measure of IME information

Reviewer #3 suggests that we should analyze our research questions using an aggregate measure of IME information: *"Is there a single summary measure of IME content? I ask for three reasons. First, because the paper makes some statements that group IME as a single category (e.g., "these data support the idea that engagement-based algorithms specifically amplify IME content"). Second, because the prevalence of some forms of IME content is so small that it would be good to have a single summary statistic to evaluate the strength of the intervention (e.g., the diversified extremity algorithm reduced IME content by X standard deviations relative to the engagement-based algorithm). This is especially important since some of the null effects of RQ2 and RQ3 could be explained by a weak intervention. Third, because some forms of IME go in opposite directions (such as positive/negative emotional), so it would be interesting to have a discussion of how to interpret these changes."*

We appreciate your suggestion to analyze our research questions using an aggregate measure of IME content. In response, we now construct and analyze an aggregated IME measure across all three research questions (RQ1–RQ3), collapsing across our pre-registered and exploratory IME indicators.

First, this aggregate measure allows us to directly evaluate claims in the manuscript that refer to IME content as a unified category (e.g., statements that engagement-based algorithms amplify IME content). Second, it provides a clearer summary statistic for assessing the overall strength of the algorithmic interventions, particularly given the low base rates of some individual IME categories and the possibility that null effects in RQ2 and RQ3 could reflect weak aggregate effects rather than uniformly null effects across components. Third, because some IME subtypes move in opposing directions (e.g., positive vs. negative emotional content), the aggregate measure helps clarify how these offsetting shifts combine at the level of overall IME exposure, perceptions, and engagement.

Across RQ1–RQ3, results using the aggregated IME measure are consistent with the substantive conclusions drawn from the individual outcome analyses, but they provide a more concise and interpretable summary of algorithmic effects. To make this clearer to readers, we now include the IME aggregate alongside the individual outcomes in the main effects figures (Figs. 1 and 3), allowing readers to directly compare aggregated and disaggregated patterns within the same visual framework. We agree that this addition strengthens the paper by clarifying both the magnitude and interpretation of algorithmic effects on IME content.

R3.6 Further heterogeneity analyses

Reviewer #3 requests that we analyze the effect of condition on social norm perception variables while accounting for how these effects were impacted by the amount of IME content removed from feeds in the intervention condition: *"Relatedly, I would like to see more heterogeneity analysis of the social norm perception outcomes by the amount of IME content removed."*

Thanks for this suggestion. We agree that examining heterogeneity by the strength of the algorithmic intervention is important for interpreting null or weak effects on social norm perceptions. In response, we conducted two complementary robustness analyses using the aggregated IME measure (see **R3.5** above) that directly address this concern.

First, we tested whether the effects of algorithm condition on descriptive norm accuracy for IME content depended on how much IME content was actually removed from participants' feeds. We operationalized intervention strength as the amount of IME content removed relative to engagement-based feeds, computed at the participant-week level. Because this moderator is defined relative to engagement-based feeds, it is conceptually meaningful only for comparisons involving the diversified extremity (DE) condition; accordingly, we focused this heterogeneity analysis on DE versus EB contrasts. Across both pre- and post-election periods, and at both low (−1 SD) and high (+1 SD) levels of IME removal, DE feeds did not significantly differ from EB feeds in descriptive norm accuracy. Notably, the contrast between low and high IME removal corresponds to a difference of approximately 2.5 percentage points in IME exposure relative to EB. Thus, within the range of IME removal achieved by the diversified extremity algorithm in this study, reducing exposure to IME content did not measurably improve the accuracy of participants' perceptions of IME prevalence. At the same time, these results do not rule out the possibility that substantially larger reductions in IME exposure could yield detectable changes in descriptive norms.

Second, we addressed the possibility that condition differences in descriptive norm accuracy could reflect differences in how much a user was exposed to content (based on their frequency of logins), rather than algorithmic effects per se. We therefore re-estimated the descriptive norm accuracy model for the aggregated IME outcome while controlling for participant login frequency and baseline IME prevalence on the platform (derived from the Bluesky firehose on a weekly basis). The results closely mirrored the primary RQ2 findings: engagement-based feeds reduced descriptive norm accuracy relative to reverse-chronological feeds before the election but not after, and diversified extremity feeds did not significantly differ from engagement-based feeds in either period. Importantly, these patterns held even after accounting for both baseline IME prevalence and actual usage intensity.

Taken together, these two analyses suggest that (a) relatively greater removal of IME content by the diversified extremity algorithm did not yield significantly stronger effects on descriptive norm accuracy, and (b) the observed condition differences are not attributable to differential login behavior or baseline exposure or. As such, our substantive conclusions remain unchanged: changes in algorithmic exposure to IME content—whether modest or larger—did not reliably translate into more accurate social norm perceptions in this study.

We now call out these analyses in the results section for RQ2 on pg. 26 of the main text, and there are two new SI appendix sections (SI Appendix, Section 7.7.1 and 7.7.2) that detail these analyses and report results.

R3.7 Reporting intervention effects on user engagement

Reviewer #3 would like to see a report of how our intervention impacted user engagement: *"I would like to see more information about how the intervention impacted user engagement (time spent on the platform, sessions, content consumed)."*

Thanks for pushing us to clarify how the interventions affected user engagement. Engagement-related outcomes are the focus of RQ3, where we analyze participants' actual behavioral engagement with IME content, including binary indicators of whether users engaged with toxic, political, intergroup, moral, and emotional content. These analyses directly test whether algorithmic changes in feed composition translated into changes in content-specific engagement behavior, which is central to our theoretical claims about algorithmic amplification.

In addition, to address your interest in more general engagement metrics, we conducted a supplementary analysis examining weekly login frequency (session counts) by algorithm condition. Using mixed-effects models analogous to those in RQ2/3, we found no significant differences in weekly logins across conditions either before or after the election. Thus, while the interventions reliably altered exposure to IME content (RQ1), they did not produce significant changes in overall platform usage intensity. This finding is noteworthy because it suggests that modest reductions in toxic and IME content do not necessarily translate into decreased platform activity, indicating that reducing engagement-based amplification of IME content does not inevitably come at the expense of user activity (we added this point to the discussion on pg. 32 of the revised manuscript).

Together, these results indicate that the algorithmic interventions affected what content participants were shown without meaningfully changing how frequently they logged into our study feeds on a weekly basis. Full details of the login-frequency analysis are reported in SI Appendix, Section 8.8.

R3.8 Changing "first" language when citing previous work

Reviewer #3 suggests that we should not use the phrase "first" to describe the methods of Guess et al., (2023): "Lastly, I find the following sentence inaccurate: *"This study is foundational to the literature, as it was the first experimental manipulation of algorithmic amplification conducted with a broad swathe of the American populace on a real social media platform." There are several studies that conduct experimental manipulations of this type; for example, some of the Facebook/Instagram 2020 Election studies.*"

Thanks for the suggestion, which was also addressed in **R2.10** above. We changed the phrasing to remove the claim of primacy, and also cited Huszar (2021) on pg. 4 of the introduction:

> This study is foundational to the literature, as it served as a key experimental manipulation of algorithmic amplification (also see[23]) conducted with a broad swathe of the American populace on a real social media platform. Despite its importance, the study had key limitations[24].