Simple changes to content curation algorithms affect the beliefs people form in a collaborative filtering experiment

Jason W. Burton (jb.digi@cbs.dk)
Department of Digitalization, Copenhagen Business School, Howitzvej 60, 2000 Frederiksberg, DK
Center for Adaptive Rationality, Max Planck Institute for Human Development, Lentzeallee 94, 14195 Berlin, DE

Stefan M. Herzog (herzog@mpib-berlin.mpg.de)
Center for Adaptive Rationality, Max Planck Institute for Human Development, Lentzeallee 94, 14195 Berlin, DE

Philipp Lorenz-Spreen (lorenz-spreen@mpib-berlin.mpg.de)
Center for Adaptive Rationality, Max Planck Institute for Human Development, Lentzeallee 94, 14195 Berlin, DE

Abstract

Content-curating algorithms provide a crucial service for social media users by surfacing relevant content, but they can also bring about harms when their objectives are misaligned with user values and welfare. Yet, potential behavioral consequences of this alignment problem remain understudied in controlled experiments. In a preregistered, two-wave, collaborative filtering experiment, we demonstrate that small changes to the metrics used for sampling and ranking posts affect the beliefs people form. Our results show observable differences in two types of outcomes within statisticized groups: belief accuracy and consensus. We find partial support for hypotheses that the recently proposed approaches of "bridging-based ranking" and "intelligence-based ranking" promote consensus and belief accuracy, respectively. We also find that while personalized, engagement-based ranking promotes posts that participants perceive favorably, it simultaneously leads those participants to form more polarized and less accurate beliefs than any of the other algorithms considered.

Keywords: algorithmic curation; collaborative filtering; belief updating; engagement-based ranking; bridging-based ranking; intelligence-based ranking

Introduction

Social media has been cited as a vehicle of misinformation (Brady, Jackson, Lindström, & Crockett, 2023; McLoughlin & Brady, 2023), a facilitator of "filter bubbles" (Pariser, 2011), and a driver of ideological polarization (Van Bavel, Rathje, Harris, Robertson, & Sternisko, 2021; Levy, 2021). While there are competing views on the extent to which social media is responsible for such negative outcomes (e.g., Borgesius et al., 2016; Gentzkow & Shapiro, 2011; Guess et al., 2023; Bakshy, Messing, & Adamic, 2015), a popular line of reasoning raises concern with the way social media platforms algorithmically curate content for users. Since most platforms have commercial goals to retain their users and maximize revenue (e.g., through advertising), platforms are incentivized to design algorithms that curate content in ways that achieve these goals. Typically, this entails implementing some kind of engagement optimization, whereby algorithms automatically promote content to users that is predicted to garner clicks, shares, dwell time, and other forms of online attention (Narayanan, 2023b). Yet, studies suggest that the content that is likely to be engaged with online is content that is laden with negative emotion (Berger & Milkman, 2012; Robertson et al., 2023), divisiveness (Rathje, Van Bavel, & Van Der Linden, 2021; Hagey & Horwitz, 2021), and falsehoods (McLoughlin & Brady, 2023; Vosoughi, Roy, & Aral, 2018). This in turn gives rise to an alignment problem where the objectives of platforms' content-curating algorithms are at odds with the values and welfare of users (see e.g., Hadfield-Menell & Hadfield, 2019; Milli, Belli, & Hardt, 2021; Thorburn, Stray, & Bengani, 2022; Ekstrand & Willemsen, 2016).

In light of this problem, recent work has called for platforms' algorithms to be better aligned with high-level human values, such as well-being, community, and knowledge (Stray, Vendrov, Nixon, Adler, & Hadfield-Menell, 2021; Stray et al., 2022). While it is common for platforms to conduct some variation of engagement optimization (see e.g., Narayanan, 2023a; Covington, Adams, & Sargin, 2016; TikTok, 2021), this does not preclude the possibility of alternative approaches that seek to optimize for more prosocial aims. One such alternative is *bridging-based ranking* (Ovadya, 2022; Ovadya & Thorburn, 2023). As opposed to the commonly implemented *engagement-based ranking* on social media, bridging-based ranking entails promoting content "that helps bridge divides... leading to positive interactions across diverse audiences, even when the topic may be divisive" (Ovadya, 2022, p. 14). For example, a bridging-based ranking algorithm might promote content that receives bipartisan positive engagement (e.g., likes, upvotes) from users of different political identities to promote consensus. This approach is intuitively appealing given the many important but polarized topics, such as abortion rights or gun control in the United States, and bridging algorithms have already seen success in social media applications, such as Community Notes (formerly Birdwatch) on X (formerly Twitter) (Wojcik et al., 2022; X, n.d.).

However, uncritically striving to bridge divides to achieve consensus may not always be epistemically desirable. For instance, imagine a society tasked with predicting whether climate change will lead to a sea level rise of more than 0.3 meters (1 foot) by 2100 if additional preventive measures are not implemented. Here, consensus is not necessarily desirable given that there is an objective ground truth—the sea level will either rise more than 0.3 meters or not. Thus, public discourse should not be engineered to steer individuals towards some single, shared belief irrespective of the ground

In L. K. Samuelson, S. L. Frank, M. Toneva, A. Mackey, & E. Hazeltine (Eds.), *Proceedings of the 46th Annual Conference of the Cognitive Science Society.*

truth. What is desirable is "positive dissensus," whereby individuals may hold diverse beliefs in disagreement with one another, but the aggregation of those beliefs (e.g., through averaging or voting) provides an accurate collective judgement (Landemore & Page, 2015). In recognition of this, Burton (2023) proposes another approach to algorithmic content curation referred to as *intelligence-based ranking*, which entails promoting content to users that is likely to elicit belief updates that would benefit collective accuracy—most usefully when the ground truth is unknown. For example, such a ranking algorithm might promote content based solely on where the posting user is positioned in a distribution of beliefs in order to preserve diversity and guard against group biases (for proof of concept, see Burton, Almaatouq, Rahimian, & Hahn, 2024).

The recent proposals of bridging-based ranking and intelligence-based ranking demonstrate how it is feasible to identify high-level, prosocial objectives that could be operationalized and integrated into content-curating algorithms for social media. However, it remains to be seen as to whether these alternative approaches to algorithmic content curation have their intended effects, or whether their effects are any different from the status quo of engagement-based ranking in the first place.

In this paper, we set out a preregistered, two-wave experiment to compare the effects of engagement-, bridging-, and intelligence-based ranking on people's beliefs. The experiment is designed to simulate the collaborative filtering used by many social media platforms to generate personalized feeds for their users. In the first wave, Study 1, we create a content inventory by having participants engage with different argumentative posts pertaining to a range of topics. In the second wave, Study 2, participants complete a belief updating task in which they encounter "feeds" of relevant posts that have been sampled and ranked from the content inventory, following either an engagement-, bridging- or intelligence-based ranking approach[1]. All materials used and analyses reported were preregistered unless specified otherwise.

Study 1

The main purpose of Study 1 is to develop a content inventory. To enable personalized, algorithmic content curation in subsequent Study 2, we collect two types of data in Study 1. First, we survey participants' demographics and their prior beliefs on six topics (Table 1). Crucially, this survey includes a question on participants' political leaning (liberal vs. conservative), which will later be used to derive personalized feeds of content. Second, we collect engagement data by presenting participants with posts pertaining to each topic, which they must either upvote, downvote, or pass. In doing so, we are also able to analyze associations between engagement, participants' prior beliefs, and the language used in posts.

[1]Both studies received ethical approval from the IRB at the Max Planck Institute for Human Development (LIP-2023-01, LIP-2023-05). All registrations, materials, data, and analysis scripts are publicly available on OSF: https://osf.io/ep6gc/.

Method

We recruited 500 English-speaking residents of the U.S. via Prolific. Since political leaning along the liberal-conservative axis will provide the basis for personalization in Study 2, we recruited a balanced sample in Study 1 by using Prolific's prescreening tool to recruit 250 participants who self-identify as liberal and 250 participants who self-identify as conservative. Each participant first reported their age, gender, political leaning on a 10-point Likert scale from "Very liberal" (1) to "Very conservative" (10), and their prior beliefs on each topic (listed in Table 1). Then each participant proceeded through six pages, where each page pertained to one topic and contained 12 posts to either upvote, downvote, or pass. Participants were told that upvoting a post means "you agree with it and/or think more people should see it," downvoting a post means "you disagree with it and/or do not think more people should see it," and passing a post means "you feel unsure or indifferent about it." Participants were required to allocate at least two upvotes and two downvotes across the 12 posts for each topic to ensure that we collect sufficient engagement data for Study 2[2].

The posts in the content inventory were paraphrased from real social media posts found on X (formerly Twitter), Facebook, Reddit, and Kialo. Each post contains an argument either for ('pro') or against ('con') the claim made in the relevant topic; there are six pro and six con posts for each topic. In paraphrasing, we have manually manipulated the degree of 'toxicity' and 'certitude.' Toxicity is measured as a predicted probability (i.e., a score of 0.8 means 80% of readers would find the post toxic) using a machine learning classifier provided through Google's Perspective API (perspectiveapi.com). Certitude is measured as a count of words from the LIWC-22 certitude dictionary (Boyd, Ashokkumar, Seraj, & Pennebaker, 2022).The full list of posts used in Study 1 is publicly available on OSF.

Preregistered hypotheses

Although the primary purpose of Study 1 is instrumental—to create a content inventory to use for Study 2—we also took the opportunity to preregister and test three hypotheses:

H1 There is a significant association between concordance and upvoting, such that posts that are concordant with participants' prior beliefs are more likely to be upvoted.

H2 There is a significant association between discordance and downvoting, such that posts that are discordant with participants' prior beliefs are more likely to be downvoted.

H3 Posts' 'toxicity' and 'certitude' will be significantly associated with total engagement (summing over upvotes and downvotes), such that higher toxicity and higher certitude will predict greater overall engagement.

[2]The preregistration for Study 1 mistakenly states that participants must allocate four upvotes and four downvotes. Participants were only required to allocate two upvotes and two downvotes.

Table 1: Topics used in Study 1 and 2. For subjective topics, participants' beliefs were measured on a 0-100 slider where 0 indicated that they "completely disagree" with the topic and 100 indicated that they "completely agree." For objective topics, participants' beliefs were measured on a 0-100 slider where 0 indicated that they believe the topic "definitely will not happen" and 100 indicated that they believe the topic "definitely will happen." The objective topics are predictive in nature, meaning their true outcomes could neither be known by participants nor experimenters at the time of running the experiment. Data collection was completed on 10 May 2023 for Study 1 and 30 June 2023 for Study 2.

Topic prompt	ID	Type	Outcome
The S&P 500 Index will close at a lower value on 31 July 2023 than it did on 31 January 2023	sp500	Objective	False
July 2023 will be the hottest July on record according to NOAA's Global Surface Temperature Analysis	july	Objective	True
US President Joe Biden's approval rating will be above 44% on 31 July 2023 according to FiveThirtyEight	biden	Objective	False
Social media platforms should never ban or remove any content	moderation	Subjective	NA
All public bathrooms should be gender neutral	bathrooms	Subjective	NA
Religion has been a good thing for humanity	religion	Subjective	NA

We define a concordant post as one with a stance that aligns with the participant's prior belief [e.g., a 'pro' argument pertaining to one of those topics for which the participant indicates a 'pro' belief prior to being exposed to the arguments (i.e., a belief greater than 50 on our 0-100 scale), or a 'con' argument pertaining to one of the topics for which the participant indicates a 'con' belief (i.e., a belief less than 50 on our 0-100 scale)], and vice versa for discordant arguments.

Results

After excluding participants who either failed an attention check or requested for their data to be withdrawn (an option offered in the study debriefing), our final sample consisted of 497 participants aged from 18 to 85 years old ($M = 42.5$, $SD = 13.9$, 51% male, 48% female, 1% other). Some participants recorded as conservatives by Prolific self-identified as liberals in our survey, leading to 60% liberals in our sample.

To test H1, we use a mixed-effects logistic regression model with 'upvote' as the binary dependent variable, concordance as a fixed factor, and random slopes and intercepts for participants and posts. Trials where a participant's belief is exactly 50 are excluded in this analysis since concordance is undefined. The probability of *upvoting* a post that is *concordant* with one's prior (0.57; 95% CI [0.53, 0.61]) is higher than that of upvoting it when it is *discordant* (0.24, 95% CI [0.21, 0.27]; $\beta = 1.42$, $SE = 0.09$, $z = 15.63$, $p < .001$). This result corroborates H1. To test H2, we use the same model except using 'downvote' as the binary dependent variable and discordance as the fixed factor. The probability of *downvoting* a post that is *discordant* with one's prior (0.49, 95% CI [0.44, 0.54]) is higher than that of downvoting it when it is *concordant* (0.21, 95% CI [0.18, 0.23]; $\beta = 1.32$, $SE = 0.09$, $z = 15.06$, $p < .001$). This result corroborates H2. For H3, we use the same model except using 'engagement' as the binary dependent variable (i.e., both upvotes and downvotes are coded as engagement), 'toxicity,' 'certitude,' and their interaction as fixed factors, and random effects for participants and posts[3]. After removing degenerate random-effect parameters (Bates, Kliegl, Vasishth, & Baayen, 2015), the model only included random slopes and intercepts for 'toxicity' by participant. More toxic posts received more engagement ($\beta = 0.99$, $SE = 0.40$, $z = 2.51$, $p = .012$), while neither certitude ($\beta = -0.12$, $SE = 0.09$, $z = -1.31$, $p = .191$) nor the interaction between toxicity and certitude ($\beta = 0.14$, $SE = 0.32$, $z = 0.44$, $p = .657$) were associated with engagement. This result partially corroborates H3.

In sum, Study 1 suggests that participants are more likely to upvote posts that reinforce their prior beliefs, more likely to downvote posts that argue against their prior beliefs, and more likely to engage with toxic posts.

Study 2

In Study 2, we use the content inventory produced by Study 1 to create different, algorithmically ranked "feeds" and present them to participants in a standard belief updating paradigm. This entails having participants report a prior belief on a topic, engage with a "feed" with relevant posts, and then revise their belief. The experimental manipulation in this study is the way feeds of posts are generated, with the following five approaches used in a between-subjects design.

- **Random ranking**: Each participant views three arguments that have been randomly selected from the 12 possible ar-

[3]This analysis deviates from the preregistration, which stipulates the use of a mixed-effects linear regression model. Since engagement is a binary observation, a logistic model is more appropriate.

guments for each topic (re-sampled for each participant). This provides a baseline condition.

- **Engagement-based ranking**: Each participant views the three posts that received the greatest total engagement (i.e., the total sum of upvotes and downvotes) for each topic (ranked in descending order).

- **Engagement-based ranking (personalized)**: Liberal participants view the three posts that received the most liberal upvotes for each topic (ranked in descending order). Conservative participants view the three posts that received the most conservative upvotes for each topic (ranked in descending order). This condition mimics the engagement optimization commonly experienced on current social media platforms.

- **Bridging-based ranking**: Each participant views the three arguments that received the most balanced ratio of conservative upvotes to liberal upvotes for each topic (ranked from most balanced to least balanced). This sorting is referred to as *diverse approval* (Ovadya & Thorburn, 2023).

- **Intelligence-based ranking (personalized)**: Liberal participants view the three posts that received the largest sum of conservative upvotes and liberal passes for each topic (ranked in descending order). Conservative participants view the three posts that received the largest sum of liberal upvotes and conservative passes for each topic (ranked in descending order). This sorting aims to promote posts to participants that may be neglected by people with similar beliefs, despite those posts being highly appreciated by people with opposing beliefs, potentially exposing participants to "inconvenient facts."

Method

We recruited a new sample of 1,000 English-speaking residents of the U.S. via Prolific. To ensure a balanced sample of liberal and conservative participants, we used Prolific's prescreening tool to recruit 500 participants who self-identify as liberal and 500 participants who self-identify as conservative. First, each participant reported the same demographics as in Study 1. Then, each participant was randomly assigned to one of the five conditions, and then proceeded through the belief updating task. For each of the six topics, participants reported their prior belief on a 0–100 scale, and then viewed a "feed" of three posts about the topic (sampled and ranked from the content inventory created in Study 1). When viewing a feed, participants were required to either upvote, downvote, or pass each individual post. After viewing a feed, participants then revised their belief on the topic. Finally, upon completing the belief updating task for all six topics, participants indicated their level agreement with three general statements on a 5-point Likert scale from "Strongly disagree" (1) to "Strongly agree" (5): "The posts I saw were *civil*," "the posts I saw were *emotional*," and "the posts I saw were *insightful*."

The key dependent variables in Study 2 are the observed change in belief variance (ΔVar) for subjective topics, and the observed change in collective error (ΔCE) and average individual error (ΔIE) for objective topics. Each dependent variable is observed within statisticized groups on each relevant topic, where each participant i holds a belief $B \in \{0,1,2,\ldots,100\}$. To generate our statisticized groups, within each condition, we drew 10,000 unique samples of 100 participants (without replacement). Within each group, we calculated the dependent variables separately for each topic as follows:

$$\Delta Var_j = Var(B_{revised,j}) - Var(B_{initial,j}) \quad (1)$$

ΔVar_j denotes the change in variance for topic j. $Var(B_{revised,j})$ represents the variance, Var, of revised beliefs on topic j, and $Var(B_{initial,j})$ represents the variance of initial beliefs on topic j. A negative ΔVar indicates a reduction in belief variance — an increase in consensus.

$$\Delta CE_j = (T_j - \bar{B}_{revised,j})^2 - (T_j - \bar{B}_{initial,j})^2 \quad (2)$$

ΔCE_j denotes the change in collective squared error (i.e, Brier score) for a topic j. T_j represents the binary truth value for the j^{th} topic, where T_j is 0 if the event did not occur and 100 if it did occur. $\bar{B}_{revised,j}$ represents the mean revised belief on topic j, and $\bar{B}_{initial,j}$ represents the mean initial belief on topic j. A negative ΔCE thus indicates a reduction in collective squared error, that is, an increase in collective accuracy.

$$\Delta IE_j = \frac{1}{n}\sum_{i=1}^{n}(B_{revised,i,j} - T_j)^2 - \frac{1}{n}\sum_{i=1}^{n}(B_{initial,i,j} - T_j)^2 \quad (3)$$

ΔIE_j denotes the change in average individual squared error for a topic j. $B_{revised,i,j}$ represents the revised belief of participant i for topic j, and $B_{initial,i,j}$ represents the initial belief of participant i for topic j. A negative ΔIE_j thus indicates a reduction in average individual squared error, that is, an increase in average individual accuracy.

Preregistered hypotheses

We preregistered three main hypotheses for Study 2:

H4 For the subjective topics (i.e., with no ground truth) the bridging-based ranking condition will decrease variance (i.e., promote consensus) within statisticized groups more than the other conditions.

H5 For the objective topics (i.e., with a ground truth), the intelligence-based ranking condition will increase collective accuracy within statisticized groups more than the other conditions.

H6 For the objective topics, the intelligence-based ranking condition will increase average individual accuracy within statisticized groups more than the other conditions.

Since we examine our dependent variables within statisticized groups, it would be inappropriate to conduct standard inferential tests because the resulting p-values could be made arbitrarily small by simply increasing the number of statisticized groups. For this reason, we evaluate our hypotheses using a descriptive approach where, for each hypothesis, we calculate pairwise probabilities of superiority A between each condition for each topic. Here A indicates the probability that, for a specific topic, the value of the dependent variable of a randomly sampled group from one condition is more desirable (in this case, lower) than that of a randomly sampled group from another condition (Ruscio, 2008).

As preregistered, H4 is considered to be supported if the bridging-based ranking treatment's mean probability of superiority (across the three subjective topics) is greater than 51% against each of the other treatments, or partially supported if it is greater than 51% against any of the other treatments. H5 and H6 are supported if the intelligence-based ranking treatment's mean probability of superiority (across the three objective topics) is greater than 51% against each of the other treatments, or partially supported if it is greater than 51% against any of the other treatments.

Results

After excluding participants who either failed an attention check or requested their data be withdrawn, our final sample consisted of 995 participants aged from 18 to 79 years old ($M = 42.4$, $SD = 14.1$, 56% male, 43% female, 2% other), and 50% of the participants self-identified as liberal.

To test H4, we calculated pairwise probabilities of superiority, A, of ΔVar for the bridging-based ranking condition vs. all other conditions for each subjective topic. We find an average probability of superiority of 65% (corresponding to a medium effect size, Cohen's $d = .53$) against the personalized engagement-based ranking condition. However, the average probability of superiority for bridging-based ranking against the other conditions is less than 50%, suggesting that bridging-based ranking does not promote consensus more than intelligence-based ranking ($A = .33$, $d = -.61$), non-personalized engagement-based ranking ($A = .38$, $d = -.42$), or random ranking ($A = .35$, $d = -.56$). We thus only find partial support for H4.

To test H5, we calculated pairwise probabilities of superiority, A, of ΔCE for the intelligence-based ranking condition vs. all other conditions for each objective topic. This analysis returns an average probability of superiority of 67% ($d = .61$) against the random ranking condition, 60% ($d = .35$) against the personalized engagement-based ranking condition, and 57% ($d = .26$) against the bridging-based ranking condition. However, the average probability of superiority of ΔCE for intelligence-based ranking against non-personalized engagement-based ranking is ($A = .17$, $d = -1.35$), suggesting that intelligence-based ranking does not promote collective accuracy more than non-personalized engagement-based ranking. These results partially confirm H5.

To test H6, we do the same analysis as for H5, but substitute ΔIE for ΔCE. This returns the same pattern of results as observed for H5, but with different effect sizes. The average probability of superiority of intelligence-based ranking versus random-ranking ($A = .80$, $d = 1.20$), personalized engagement-based ranking ($A = .72$, $d = .82$), and bridging-based ranking ($A = .59$, $d = .32$) partially confirm H6. Yet, the average probability of superiority of ΔIE for intelligence-based ranking against non-personalized engagement-based ranking is 25% ($d = -.97$), suggesting that intelligence-based ranking does not promote individual accuracy more than non-personalized engagement-based ranking.

Exploratory analyses showed a dissociation between participants' perceptions of the posts they saw and the effect those posts had on consensus and belief accuracy. Between condition comparisons with Tukey's multiple comparison test show that participants in the non-personalized engagement-based ranking condition rated posts as generally less civil (d's range from .36 to 1.08; all p's $< .001$) and more emotional than participants in any other condition (d's range from .35 to .85; all p's $< .017$), and less insightful than participants in the intelligence-based ($d = .28$, $SE = .14$, $p = .022$) and bridging-based ranking conditions ($d = .42$, $SE = 0.14$, $p < .001$). Intriguingly, participants in the *personalized* engagement-based ranking condition—the condition mimicking the curation commonly experienced on current social media platforms—rated posts as more insightful than participants in any other condition (d's range from .30 to .71; all p's $< .05$), and personalized engagement-based feeds received significantly more upvotes and less downvotes than any other condition (Figure 1, B). Despite the positive reception, those same personalized engagement-based feeds led participants towards less consensus and less accurate beliefs than any other condition (i.e., the average probability of superiority versus all other conditions for each of the dependent variables is less than 50%).

In sum, the results of Study 2 partially confirm the hypotheses that bridging-based ranking and intelligence-based ranking can promote consensus and belief accuracy, respectively. Although results appear to be highly topic-specific (Figure 1, A), they show how small, simple changes to the metrics used to algorithmically curate content can have potentially large effects on belief accuracy and consensus.

Discussion

Our experiment clearly illustrates how variations in algorithmic content curation can lead groups to form different beliefs—for better or worse. Contrary to studies arguing that the importance ascribed to social media platforms' algorithms may be overblown (e.g., Borgesius et al., 2016; Bakshy et al., 2015; Guess et al., 2023), we observe large effect sizes in a controlled setting with only small, simplistic changes to the metrics used to sample and rank posts. Moreover, our results speak directly to the issue of misalignment between platform algorithms and user welfare: Personalized engagement-based feeds tended to receive more upvotes, less downvotes, and

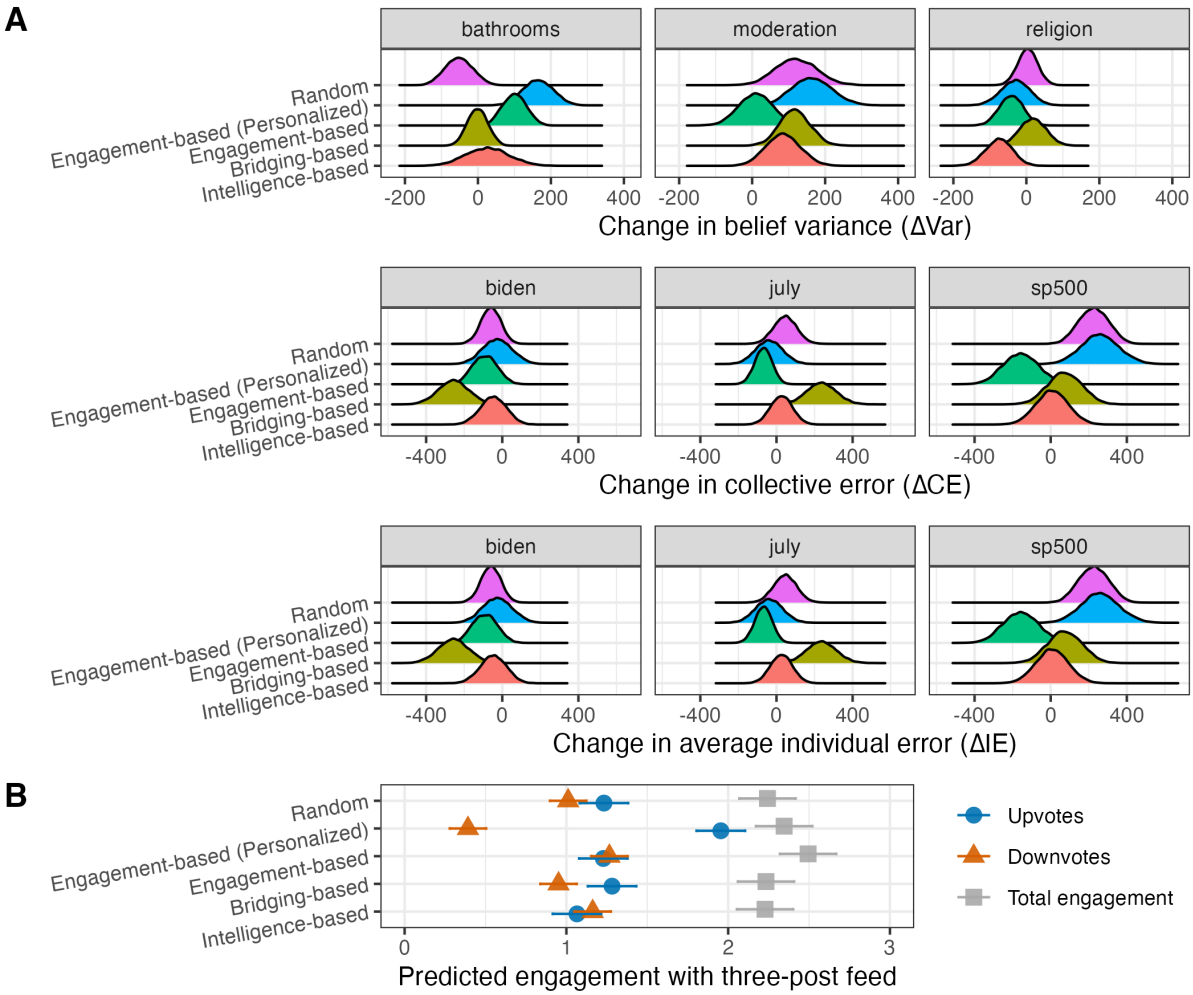

Figure 1: Study 2 results. **A.** Distributions of dependent variables within statisticized groups (i.e., each distribution consists of 10,000 observations). Negative or lower values are desirable. The top row pertains to H4 (i.e., reduction in variance = increased consensus), the middle row to H5, and the bottom row to H6 (i.e., reduction in individual or collective errors, respectively). **B.** Mixed-effects linear regression models' predictions of the number of upvotes, downvotes, and total engagement (upvotes plus downvotes) that feeds receive (i.e., separate models predicting upvotes, downvotes, and total engagement). Condition is specified as a fixed factor with random intercepts for participants and topics. Error bars display 95% confidence intervals.

be rated as more insightful, yet those same feeds led participants towards less consensus and less accurate beliefs than any other ranking approach. While our findings shed a negative light on the current use of personalized engagement-based ranking, encouragingly, we also show that it is possible to re-design algorithmic content curation for prosocial outcomes using, say, bridging- or intelligence-based ranking. However, given the variable effects we observe across topics (Figure 1, A), designing algorithms that work reliably across domains will be a challenge for future research. Yet, this challenge is worthwhile because such algorithms could amplify quality content online by tapping into the collective intelligence of users and their natural interactions, rather than relying on any third-party judgments on a post-by-post basis.

Beyond our substantive findings, our studies provide a methodological contribution. Existing studies of algorithmic content curation on social media have largely relied on observational data or social media field experiments, to which access is often gatekept by the platforms themselves. Furthermore, relying on observational data alone runs the risk of interpreting spurious correlation as cause-and-effect (Burton, Cruz, & Hahn, 2021), and social media field experiments are typically unable to rule out spillover effects, where a control group may be indirectly exposed to the same algorithmically curated content as the experimental group if their contacts are not also in the control group, and vice versa (Forastiere, Airoldi, & Mealli, 2021). Given these challenges, our studies offer a template for running relatively cheap, controlled experiments to test for effects of different approaches to content curation without the need for platform access.

References

Bakshy, E., Messing, S., & Adamic, L. A. (2015). Exposure to ideologically diverse news and opinion on facebook. *Science*, *348*(6239), 1130–1132.

Bates, D., Kliegl, R., Vasishth, S., & Baayen, H. (2015). Parsimonious mixed models. *arXiv preprint arXiv:1506.04967*.

Berger, J., & Milkman, K. L. (2012). What makes online content viral? *Journal of Marketing Research*, *49*(2), 192–205.

Borgesius, F. J. Z., Trilling, D., Möller, J., Bodó, B., De Vreese, C. H., & Helberger, N. (2016). Should we worry about filter bubbles? *Internet Policy Review*.

Boyd, R. L., Ashokkumar, A., Seraj, S., & Pennebaker, J. W. (2022). *The development and psychometric properties of LIWC-22*. University of Texas at Austin. Austin, TX. Retrieved from https://www.liwc.app

Brady, W., Jackson, J., Lindström, B., & Crockett, M. (2023, Oct). Algorithm-mediated social learning in online social networks. *Trends in Cognitive Sciences*, *27*(10), 947-960.

Burton, J. W. (2023). *Algorithmic amplification for collective intelligence.* Retrieved from https://knightcolumbia.org/content/algorithmic-amplification-for-collective-intelligence (Knight First Amendment Institute)

Burton, J. W., Almaatouq, A., Rahimian, M. A., & Hahn, U. (2024). Algorithmically mediating communication to enhance collective decision-making in online social networks. *Collective Intelligence*, *3*(2), 26339137241241307.

Burton, J. W., Cruz, N., & Hahn, U. (2021). Reconsidering evidence of moral contagion in online social networks. *Nature Human Behaviour*, *5*(12), 1629–1635.

Covington, P., Adams, J., & Sargin, E. (2016). Deep neural networks for YouTube recommendations. In *Proceedings of the 10th ACM Conference on Recommender Systems* (pp. 191–198).

Ekstrand, M. D., & Willemsen, M. C. (2016). Behaviorism is not enough: Better recommendations through listening to users. In *Proceedings of the 10th ACM Conference on Recommender Systems* (pp. 221–224).

Forastiere, L., Airoldi, E. M., & Mealli, F. (2021). Identification and estimation of treatment and interference effects in observational studies on networks. *Journal of the American Statistical Association*, *116*(534), 901–918.

Gentzkow, M., & Shapiro, J. M. (2011). Ideological segregation online and offline. *The Quarterly Journal of Economics*, *126*(4), 1799–1839.

Guess, A. M., Malhotra, N., Pan, J., Barberá, P., Allcott, H., Brown, T., ... others (2023). How do social media feed algorithms affect attitudes and behavior in an election campaign? *Science*, *381*(6656), 398–404.

Hadfield-Menell, D., & Hadfield, G. K. (2019). Incomplete contracting and AI alignment. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society* (pp. 417–422).

Hagey, K., & Horwitz, J. (2021). Facebook tried to make its platform a healthier place. It got angrier instead. *The Wall Street Journal*.

Landemore, H., & Page, S. E. (2015). Deliberation and disagreement: Problem solving, prediction, and positive dissensus. *Politics, Philosophy & Economics*, *14*(3), 229–254.

Levy, R. (2021). Social media, news consumption, and polarization: Evidence from a field experiment. *American Economic Review*, *111*(3), 831–870.

McLoughlin, K. L., & Brady, W. J. (2023). Human-algorithm interactions help explain the spread of misinformation. *Current Opinion in Psychology*, 101770.

Milli, S., Belli, L., & Hardt, M. (2021). From optimizing engagement to measuring value. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency* (pp. 714–722).

Narayanan, A. (2023a). *Twitter showed us its algorithm. what does it tell us?* Retrieved from https://knightcolumbia.org/blog/twitter-showed-us-its-algorithm-what-does-it-tell-us (Knight First Amendment Institute)

Narayanan, A. (2023b). *Understanding social media recommendation algorithms.* Retrieved from https://knightcolumbia.org/content/understanding-social-media-recommendation-algorithms (Knight First Amendment Institute)

Ovadya, A. (2022). *Bridging-based ranking.* Retrieved from https://www.belfercenter.org/publication/bridging-based-ranking (Belfer Center for Science and International Affairs, Harvard Kennedy School)

Ovadya, A., & Thorburn, L. (2023, October). *Bridging systems: Open problems for countering destructive divisiveness across ranking, recommenders, and governance.* Retrieved from https://knightcolumbia.org/content/bridging-systems (Knight First Amendment Institute)

Pariser, E. (2011). *The filter bubble: How the new personalized web is changing what we read and how we think*. Penguin.

Rathje, S., Van Bavel, J. J., & Van Der Linden, S. (2021). Out-group animosity drives engagement on social media. *Proceedings of the National Academy of Sciences*, *118*(26), e2024292118.

Robertson, C. E., Pröllochs, N., Schwarzenegger, K., Pärnamets, P., Van Bavel, J. J., & Feuerriegel, S. (2023). Negativity drives online news consumption. *Nature Human Behaviour*, *7*(5), 812–822.

Ruscio, J. (2008). A probability-based measure of effect size: robustness to base rates and other factors. *Psychological Methods*, *13*(1), 19.

Stray, J., Halevy, A., Assar, P., Hadfield-Menell, D., Boutilier, C., Ashar, A., ... others (2022). Building human values into recommender systems: An interdisciplinary synthesis. *ACM Transactions on Recommender Systems*.

Stray, J., Vendrov, I., Nixon, J., Adler, S., & Hadfield-Menell,

D. (2021). What are you optimizing for? Aligning recommender systems with human values. *arXiv preprint arXiv:2107.10939*.

Thorburn, L., Stray, J., & Bengani, P. (2022). *What does it mean to give someone what they want? the nature of preferences in recommender systems.* Retrieved from https://medium.com/understanding-recommenders/what-does-it-mean-to-give-someone-what-they-want-the-nature-of-preferences-in-recommender-systems-82b5a1559157 (Understanding Recommenders)

TikTok. (2021). *An update on our work to safeguard and diversify recommendations.* Retrieved from https://newsroom.tiktok.com/en-us/an-update-on-our-work-to-safeguard-and-diversify-recommendations

Van Bavel, J. J., Rathje, S., Harris, E., Robertson, C., & Sternisko, A. (2021). How social media shapes polarization. *Trends in Cognitive Sciences*, *25*(11), 913–916.

Vosoughi, S., Roy, D., & Aral, S. (2018). The spread of true and false news online. *Science*, *359*(6380), 1146–1151.

Wojcik, S., Hilgard, S., Judd, N., Mocanu, D., Ragain, S., Hunzaker, M., … Baxter, J. (2022). Birdwatch: Crowd wisdom and bridging algorithms can inform understanding and reduce the spread of misinformation. *arXiv preprint arXiv:2210.15723*.

X. (n.d.). *Community notes guide: Note ranking algorithm.* Retrieved from https://communitynotes.twitter.com/guide/en/under-the-hood/ranking-notes#complete-algorithm-steps