How Does Empowering Users with Greater System Control Affect News Filter Bubbles?

Ping Liu[1], **Karthik Shivaram**[2], **Aron Culotta**[2], **Matthew Shapiro**[3], **Mustafa Bilgic**[4]

[1] LinkedIn Corporation, Sunnyvale, CA USA
[2] Department of Computer Science, Tulane University, New Orleans, LA USA
[3] Department of Social Sciences, Illinois Institute of Technology, Chicago, IL USA
[4] Department of Computer Science, Illinois Institute of Technology, Chicago, IL USA
piliu@linkedin.com, kshvaram@tulane.edu, aculotta@tulane.edu, shapiro@iit.edu, mbilgic@iit.edu

Abstract

While recommendation systems enable users to find articles of interest, they can also create "filter bubbles" by presenting content that reinforces users' pre-existing beliefs. Users are often unaware that the system placed them in a filter bubble and, even when aware, they often lack direct control over it. To address these issues, we first design a political news recommendation system augmented with an enhanced interface that exposes the political and topical interests the system inferred from user behavior. This allows the user to adjust the recommendation system to receive more articles on a particular topic or presenting a particular political stance. We then conduct a user study to compare our system to a traditional interface and found that the transparent approach helped users realize that they were in a filter bubble. Additionally, the enhanced system led to less extreme news for most users but also allowed others to move the system to more extremes. Similarly, while many users moved the system from extreme liberal/conservative to the center, this came at the expense of reducing political diversity of the articles shown. These findings suggest that, while the proposed system increased awareness of the filter bubbles, it had heterogeneous effects on news consumption depending on user preferences.

1 Introduction

Personalized recommendation systems help users find items of interest and foster new connections (Guo et al. 2017; Tang, Hu, and Liu 2013), but emerging research suggests that there are unintended side-effects. This is particularly the case for systems recommending political content, resulting in "filter bubbles" in which users are being pushed toward homogeneous rather than diverse political content (Pariser 2011; Bakshy, Messing, and Adamic 2015; Robertson et al. 2021; Liu et al. 2021).

Recent research has attempted to quantify filter bubbles and mitigate them algorithmically (Masrour et al. 2020; Liu et al. 2021; Shivaram et al. 2022); yet, there have been few attempts to investigate the effects of giving users greater control over the recommendation algorithm. We design and study a recommendation system with two enhancements: (1) *transparency:* the system exposes the current state of the recommendation system, revealing the political and topical interests inferred from user behavior; (2) *interaction:* the interface allows the user to adjust the recommendation system to receive more articles of a particular topic or political stance. Figure 1 shows a sample of the enhanced interface, where sliders can be adjusted by users in order to modify the news content that they receive.

The goal of this paper is to understand the impact of such an interface on system behavior and filter bubbles, as compared with the more limited types of interaction allowed by traditional recommendation systems. We are particularly interested in how the system affects both the diversity of recommended political news articles as well as user engagement with the system. To this end, we conducted a user study of 102 users (recruited from 850 users who completed a demographic survey and political qualification questionnaire), half of whom used the enhanced interface (the treatment group), and half of whom used a more conventional recommendation system where the only available actions were up-vote or down-vote an article (the control group). We compute measures over the attributes of the top recommended articles at the beginning and end of each session to study how this interface influences the types of articles shown to the user, how those articles change over time, and whether the system is accurate in its recommendations. By analyzing the results of over 3,000 user interactions with these systems, we make the following observations:

- **Extremeness:** Among users who were initially exposed to extreme, partisan articles, those in the treatment group were more likely to steer the system to less extreme articles. On the other hand, among users who were initially exposed to less extreme articles, those in the treatment group were somewhat more likely to steer the system to extreme articles. These results suggest a potential extremeness "sweet spot" that users seek.
- **Diversity:** For both treatment and control groups, users steered the system toward less politically diverse news. The largest difference between the groups was for users who were initially exposed to moderately diverse articles — such users in the control group steered the system to less diverse articles than users in the treatment group.
- **Up-vote Ratio:** For both treatment and control groups, the ratio of articles that are of up-voted by the users increased over time, particularly for users with low initial

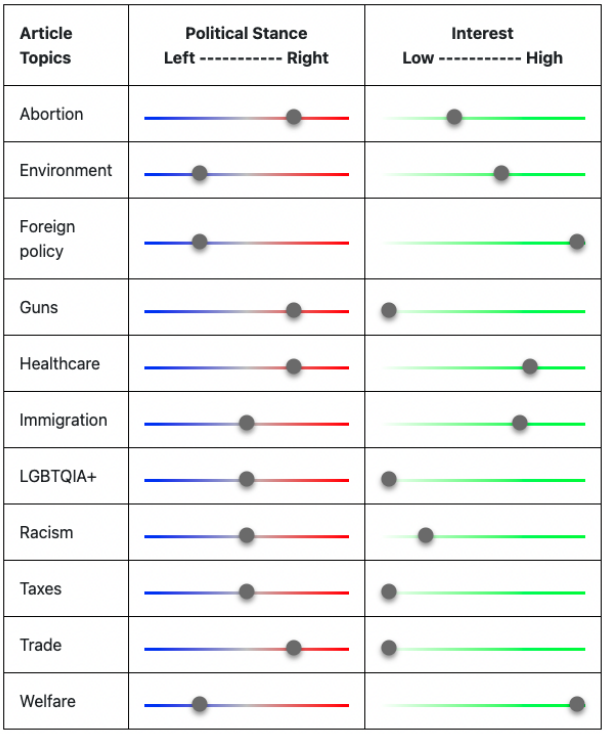


Figure 1: User interface to provide transparency and control over a political news recommendation system.

up-vote ratios. For users with moderate initial up-vote ratios, those in the treatment group were able to adjust the system to achieve greater system accuracy than those in the control group.

- **User awareness:** Through a post-study questionnaire to identify users' motivations and preferences toward our novel recommender system interface, we observed that the transparency of the enhanced interface raised user awareness regarding both the lack of diversity in their recommended articles as well as the inner-workings of news recommendation systems.

The rest of the paper is organized as follows. We discuss the relevant literature and our research questions in §2, we detail our approach in §3, followed by a presentation of the findings of the user studies in §4. We then discuss limitations of the present work as well as the potential for future research in §5, followed by a concluding section.

2 Related Work and Research Questions

Building on pioneering research on news recommender systems by Chesnais, Mucklo, and Sheena (1995), Kamba, Bharat, and Albers (1995), and Claypool et al. (1999), Mitova et al. (2023) provide a systematic survey and investigation of news recommender systems in terms of how they affect journalists/media outlets (delivery perspective) and news readers (acquisition perspective). A central challenge posed by recommender systems is the lack of diversity in the recommended items (Kunaver and Požrl 2017). Among the reasons for this diversity problem are: bias introduced by content (e.g., the model latching on to specific keywords), feedback loops when recommendations by the model end up in its training data, and popularity bias when popular rather than niche items are recommended. These are long-standing concerns and have been studied as early as by Smyth and McClave (2001) and Ziegler et al. (2005).

The lack of recommendation diversity can create a negative experience for users, as they may be exposed to similar content repeatedly while missing niche content about books, movies, and consumer goods. In the news recommendation domain, it can also lead to echo chambers and filter bubbles (Pariser 2011), where users are overrecommended news items — particularly political news — with which they are ideologically or otherwise aligned. When exposed to political content consistent with one's views, people typically prefer more of the same (Rodriguez et al. 2017). The feedback loops created by these kinds of recommendation models exacerbates echo chambers (Pariser 2011; Bakshy, Messing, and Adamic 2015; Liu et al. 2021), with varying effects across parties (Tewksbury and Riles 2015).

User feedback and engagement is critical for understanding the mechanisms underlying filter bubbles. Munson and Resnick (2010) conducted a user study where people were assigned either ideologically consistent or inconsistent recommendations and asked to rate them. When the recommendations were unaligned, user satisfaction was low; however, when the list contained a large percentage of agreeable items, responses were much varied: some users were more satisfied, while others were not, suggesting that some people can be "challenge-averse" while others are "diversity-seeking." In a similar vein, Liu et al. (2021) curated a political news dataset covering numerous topics and conducted simulations using content-based and collaborative-filtering recommender systems. Users who were initially presented extreme news were subsequently presented even more extreme news, users shown more extreme news had higher up-vote ratios, and the recommender system had the least recommendation accuracy for users with diverse views among the various news topics, often resulting in recommendations that were ideologically uniform across topics. This is relevant given that exposure to diverse news content can be effective at reducing filter bubbles in news recommendations (Ookalkar, Reddy, and Gilbert 2019).

The need to combat filter bubble formation is clear (Resnick et al. 2013), but approaches vary. Li et al. (2023) take an algorithmic approach and conduct simulations, defining bubbles as communities that have many inward but few outward connections in a bi-partite graph of users and items. They use a reinforcement learning algorithm to decide which community-connecting edges should be added to the graph to increase diversity. Others, such as Masrour et al. (2020), propose solutions based on algorithmic fairness criteria, while still others propose an attention-based modeling architecture to reduce the political homogenization effect in news recommendation (Shivaram et al. 2022).

In short, the extant research on this subject is temporally static, simulates user-controllable news recommender systems rather than examining the real world (Wang et al.

2022), lacks an appropriate interaction tool (Faridani et al. 2010; Munson, Lee, and Resnick 2013), is overly descriptive (Harambam et al. 2019), or focuses on user control in non-news (i.e., social media, movie/music recommendation, etc.) contexts (Bhargava et al. 2019; Taijala, Willemsen, and Konstan 2018; Jannach, Naveed, and Jugovac 2016).

Research Questions

We conduct a user study of a political news recommendation system, where the *control* group has access to the traditional interaction mechanism of up-voting/down-voting a news article. The *treatment* group, however, has also access to an enhanced user interface (UI) (see Fig. 1) where they can adjust which political topics they are most interested in as well as their political preference for articles on each topic. We investigate the following research questions:

- **RQ1**: How does a user's interaction with a political news recommender system affect the system's recommendation trajectory?
- **RQ2**: Do changes in the recommendation system's trajectory differ significantly for the control group versus the treatment group?

Both research questions are motivated by the extant literature in several ways. First, based on simulations of news recommender systems presented in Liu et al. (2021), we know that users are increasingly presented more extreme, less diverse, and more homogeneous news articles. Yet, there are distinctions between "challenge-averse" and "diversity-seeking" individuals (Munson and Resnick 2010), illustrating that personal characteristics and preferences affect the diversity of opinions to which people are exposed. Second, people are open to the possibility of manipulating the recommender system to increase exposure to diverse content (Harambam et al. 2019) — and in fact are able to increase diversity through a transparency tool (Munson, Lee, and Resnick 2013). Greater diversity of content results in the perceived value of the recommendations initially going up but then going down (Ziegler et al. 2005), i.e., the ability to self-navigate through online content does not increase its diversity (Faridani et al. 2010).

For the *control* group, we expect that they will be presented increasingly more extreme and less diverse news over time, in line with (Liu et al. 2021). Given that party affiliation is a key predictor of the online content with which people engage (Allen, Martel, and Rand 2022; Törnberg 2022), that partisans are more entrenched in their beliefs (Brewer 2005) and thus more likely to be motivated by their preexisting beliefs (Kahan 2015; Lodge and Taber 2000), and given the connections between partisanship and news extremeness (Tewksbury and Riles 2015; Levendusky and Malhotra 2016), we expect to see the largest shifts for users who consume more moderate news content. Regarding model accuracy, measured through up-vote ratio, we expect our models to improve as additional training data is fed into them, increasing the up-vote ratio over time.

Even though there is evidence in the literature for both "challenge-averse" users who prefer to see agreeable news and "diversity-seeking" users who are more amenable to diverse opinions (Munson and Resnick 2010), we do not expect the simple mechanism of up-voting/down-voting articles to be sufficient for the diversity-seeking users to steer the system to less extreme and more diverse articles.

For the *treatment* group, we expect mixed results. We expect the interaction and transparency tool to enable "challenge-averse" individuals in the *treatment* group to steer the system to greater extremes than those in the *control* group, and enable those who are "diversity-seeking" in the *treatment* to steer it to less extreme and more diverse articles than those in the *control*. We expect mixed results for the up-vote ratio as well: the system should be able to learn user preferences better over time and hence lead to higher up-vote ratios; however, some users might use the transparency and interaction tool to drastically change the system and thus experience a lower up-vote ratio than those in the *control* group, as providing users with more control does not guarantee its effective utilization (Mitova et al. 2023).

3 Our Approach

Dataset We focus on the political news domain in our user study, which not only has implications for policy making and electoral outcomes but is also likely to contain the type of ideologically polarizing content that can be most impactful — and potentially harmful — for society.

We used the U.S. political news dataset from Liu et al. (2021), collected from September 2019 to August 2020. It contains articles from 41 news sources. Each article is annotated with a political stance rating in $\{-2, -1, 0, +1, +2\}$ by *www.allsides.com*, where -2 indicates extreme liberal and $+2$ indicates extreme conservative. Each article is also annotated with one or more political topics.

We sampled 8,000 articles from each of the five political stances, which resulted in a total of 40,000 articles, summarized by topic in Table 1a. Given that an article can contain more than one topic label (e.g., "immigration" together with "racism"), there are a total of 44,033 topic labels represented by the 40,000 articles. 5,000 articles were used to bootstrap the recommender system (explained in detail below), and the remaining 35,000 articles were retained as potential candidates for recommendation.

User recruitment We recruited participants from Amazon Mechanical Turk (AMT; https://www.mturk.com/) from September to December 2021. Users were required to reside in U.S., to have voted in the 2020 election, and to possess a sufficient level of political literacy as determined via a screening survey (see Appendix). Of the 850 users who completed this survey, 595 (70%) answered at least two of the three qualification questions correctly and were subsequently invited to participate in the full recommendation system user study. Of the 595 invitees, 146 users participated in the final study, of which 44 were dropped from the study due to failed attention checks (see Appendix). The final study consists of 102 users divided randomly into control and treatment groups, the demographic information of which is presented in Table 1b.

Topic	# articles	Topic	# articles
abortion	1,988	environment	2,854
foreign policy	5,759	guns	2,781
healthcare	5,999	immigration	5,771
LGBTQIA	1,611	racism	5,550
taxes	4,639	trade	3,794
welfare	3,287		
# articles	40,000		
# labels	44,033		

(a)

	Control	Treatment	Total
Male	26	30	56
Female	25	21	46
Democrat	17	18	35
Republican	23	24	47
Independent	10	9	19
Total	51	51	102

(b)

Table 1: Topic distribution for articles used in the user study (a); and demographics of the user study participants (b).

Profile Questionnaire	
abortion	Abortion should be legal in most cases.
environment	Stricter environmental regulations and laws are worth the costs.
foreign policy	Good diplomacy is the best way for the U.S. to ensure peace.
guns	Gun laws should be stricter than they are today.
healthcare	Providing healthcare to Americans is the federal government's responsibility.
immigration	Immigrants strengthen the United States in many different ways.
LGBTQIA	Members of the LGBTIA+ community should have the right to marry.
racism	Changes are needed in American society to improve racial equality.
taxes	The U.S. economic system unfairly favors powerful interests.
trade	U.S. involvement in the global economy is good for the country.
welfare	Poor people have hard lives because government programs do not do enough for them.

Table 2: Political stance pre-questionnaire. On a five-point Likert scale, users agreed or disagreed with each statement.

Recommender study pre-questionnaire Before users were presented with news articles, we required them to complete a pre-questionnaire in which they revealed their ideological positions and personal interest in each of the 11 predominant political topics in the dataset (Table 1a).

We built the pre-questionnaire based on the Pew surveys of U.S. political typologies (Doherty, Kiley, and Johnson 2017), which asked how users agree or disagree (five-point Likert response) with statements about each of the topics in the political news dataset. For example, a user's response to the statement, "Abortion should be legal in most cases," is used to estimate their stance on abortion. Agreement or disagreement for all eleven topics is based on the statements presented in Table 2. A user's personal interest in each of the issues was determined by asking them about the extent to which they are interested in each of the eleven topics (five-point Likert response). This political stance and interest-related information was used for bootstrapping the recommendation algorithm.

Recommender system We built a two-stage recommender algorithm. A personalized content-based recommender scores each potential news article. These scores were then adjusted based on the match of the user's political stance and interest in each topic to the political stance and topic of the candidate article.

Content-based recommender We trained a personalized content-based recommender separately for each user. While recommendation systems are an active research area with many proposed deep learning solutions (Covington, Adams, and Sargin 2016; Ying et al. 2018), our goal is not to develop a new recommendation algorithm but to generate a simpler system both to enhance interpretability and to avoid overfitting in the smaller data regime of the study.

We implemented a standard text-based recommendation system as follows. Each news article content was transformed into a tf-idf vector with a vocabulary size of $3,000$ words. A logistic regression classification model was trained to predict whether a user would like (up-vote) or dislike (down-vote) an article based on its content. After each user interaction (an up-vote or down-vote of an article), a stochastic gradient descent update was made to update the model and make new recommendations.

To address the "cold-start" challenge of recommender systems, we used users' responses to the pre-questionnaire (described above) as follows. We reserved $5,000$ of the $40,000$ articles for bootstrap purposes, and we bootstrapped the content-based recommender for each user based on their political stance and interest-related responses to the pre-questionnaire. For each user, we created a political stance vector u_s with 11 entries, each of which corresponded to their stances on the 11 topics from the questionnaire. Then, for each topic, we sampled "positive/up-vote" articles that matched the user's stance, and "negative/down-vote" articles that were furthest away from the user's stance. For example, if a user's stance was -2 on abortion, the algorithm drew 25 articles from the abortion topic with a partisan score of -2 as positive/up-vote, and then drew 25 articles from the same topic with a partisan score of $+2$ as negative/down-vote. Similar pairings were done for -1, $+1$, and $+2$. If the user's stance was 0, however, the algorithm drew positive ex-

amples from stance 0 and an equal number of negative examples from both $+2$ and -2, assigning them sample weights of 0.5. A logistic regression classifier was trained for each user using their personalized bootstrap dataset.

Recalibration of the scores To operationalize the interaction mechanism presented in the enhanced UI (Figure 1), we designed a two-component recommendation system that allows the users to directly influence the final ranking of the recommendations (Steck 2018). The first component is the content-based model described above. The second component captures how well the user's stance and interest in each topic (initially collected through the pre-questionnaire and then potentially modified through the interaction mechanism) match the stance and topic of the article.

Let u_s be a user's political stance vector, initialized using their answers to the political questions in the pre-questionnaire (Table 2). This vector has one entry per topic (i.e., 11 entries), each of which ranges from -2 to $+2$. Let u_t be user's interest vector on each topic (11 entries), indicating their interest in each topic, initialized based on their pre-questionnaire. Similarly, let a_s be an article's political stance on each topic, and let a_t be the article's binary vector indicating its topic(s).

Let the recommendation score of the content-based recommender for an article be s_r (i.e., the probability of the "up-vote" class in the binary classifier). The final recommendation score for that article combines s_r with the match between the user's interest and political stance vectors and those of the article:

$$s = \lambda s_r + (1 - \lambda)(Sim(u_t, a_t) + Sim(u_s, a_s))/2 \quad (1)$$

where Sim is the cosine similarity of the two vectors. The transparency and interaction tool (Figure 1) allowed users to adjust the average political stance and the proportion of the top-ranked articles, which was implemented algorithmically by adjusting u_t and u_s vectors. We set $\lambda = 0.4$ in our experiments, as our preliminary analysis showed that it provides the best balance between the content-based classifier preference and the user's stance and interest profile.

When users were presented with an article recommended by the system, they could up-vote the article, down-vote the article, or skip it. If the users up-voted or down-voted, the article was added to the training set, the recommender was retrained using stochastic gradient descent, and a new recommendation was presented on the next appearing page. Skipped articles generated new recommendations using the existing system. Users were presented with one article at a time until they up/down-voted a total of 30 articles.

Transparency and interaction The transparency and interaction tool (Figure 1) in the treatment group exposed users to information about the recommender system, allowing them to adjust the recommender system's ranking for each topic via political stance and interest sliders. This interface reflected the current state of the recommender system for that user via the average political stance and the topic distribution of the top K ranked articles ($K = 200$ in our experiments below). The left panel of Figure 1 ("Political Stance") showed the average political stance of the top-ranked articles on each topic; the right panel ("Interest") reflected the proportions of the topics in the top-ranked articles. Any change to a slider's position was recorded immediately by updating u_t and u_s in Eq.1 through a binary search (trying possible u_t and u_s values and re-ranking the news articles) to quickly find the new u_t and u_s values reflecting the desired topic interest and political stance among the top-ranked articles. The recommender system, the top-ranked articles, and the locations of the sliders were then updated. For example, if dissatisfied with the system's recommendations, the user reflected in Figure 1 could use the political stance slider to move towards the center on the abortion topic or to the right on taxes. With the interest slider, the same process occurs to affect the proportion of articles presented on each topic. Users also had the option of resetting sliders back to the last system recommendation update.

Access to the transparency and interaction tool was provided to users in the treatment group via a link at the top of each recommendation page. They could thus opt to view their system-based profile at any time; yet, to account for users who might not use the link, users were also automatically directed to this page after reading and scoring every five articles. To account for users who were overly-dissatisfied with the content they were reading, they would be automatically directed to the transparency and interaction tool page after three consecutive down-vote actions. Instructions regarding how the user could adjust the system and details about what the sliders' positions meant were provided at the top of the interaction page. Users could spend as much time as they liked on this page before continuing to the next news article in the recommendation process.

Recommender study post-questionnaire Given that users' perceived usefulness and perceived ease of use of a particular information technology is essential for their adoption of such technology (Davis 1989; Venkatesh et al. 2003; Venkatesh and Bala 2008), after users up-voted/down-voted a total of 30 articles (approximately a one-hour session), they were then presented with a post-questionnaire to gauge their perceptions about filter bubbles and the recommender system. Users in both the control and treatment groups answered questions (five-point Likert response) regarding the extent to which they enjoyed the system (*Qa*), the extent to which they were presented with diverse articles (*Qb*), and the extent to which the study helped them learn more about how news recommender systems work (*Qd*). This single-question approach is consistent with research that bases users' opinions of recommender systems based on a single question (Faridani et al. 2010; Tsai and Brusilovsky 2019).

While the control group lacked access to the transparency tool, we could assess whether their initial responses (*Qb*) remained unchanged after being presented with information about the articles they were shown. After answering *Qb*, users in the control group were presented with a histogram of the political stances of the articles they were shown (Figure 5 in the Appendix shows an example). Users in the control group responded once again to the diversity-of-news question (*Qc*), allowing us to use the difference between *Qb* and *Qc* as a measure of how clarity about the news the

user consumed impacted their beliefs about the diversity of news, particularly for users having no access to the interaction tool.[1]

Measures Inspired in part by prior work on diversity in recommendation systems (Castells, Hurley, and Vargas 2021), we identified four key measures to evaluate the system and compare the control and treatment groups.

(1) Average political stance: This measure captures the overall ideological stance of the recommender system for user u for a given time period. The recommender system political stance score for user u is the average of the political stance of the top K-ranked articles from the recommender system. Let a_j^u be the article ranked at position j for user u and $s(a)$ be the political stance of the article a. The recommender system political stance score is:

$$\textit{Political Stance}(u) = \frac{1}{K}\sum_{j=1}^{K} s(a_j^u) \tag{2}$$

(2) Average extremeness: This measure captures the overall extremeness of the recommender system for user u for a given time period. Because political stance ranges from -2 (extreme liberal) to $+2$ (extreme conservative), where 0 represents neutral, we use the absolute value of political stance for the extremeness measure. Hence, the recommender system extremeness score for user u is the average of the absolute value of the political stance of the top K ranked articles by the recommender system. The recommender system extremeness score is:

$$\textit{Extremeness}(u) = \frac{1}{K}\sum_{j=1}^{K} |s(a_j^u)| \tag{3}$$

(3) Diversity: This measure captures the diversity of political stances of the recommendations. Let p_l^u be the proportion of articles having political stance l in the top K recommendations for user u. We measure diversity using the normalized political stance entropy as follows:

$$\textit{Diversity}(u) = \frac{\sum_{l=-2}^{+2} -p_l \times \log(p_l^u)}{\log(5)} \tag{4}$$

where $\log(5)$ is a normalization constant to ensure the entropy of a 5-category distribution is between 0 and 1, with 1 representing maximum diversity.

(4) Up-vote ratio: Our measure of system accuracy is the proportion of the recommended articles liked by the user. Let r_i^u be 1 if the user u up-voted the i^{th} article shown to them, and 0 otherwise. *Up-vote Ratio* is defined as:

$$\textit{Up-vote Ratio}(u) = \frac{1}{N}\sum_{i=1}^{N} r_i^u \tag{5}$$

[1]Code and user survey and interaction data are available at: https://github.com/IIT-ML/icwsm-2024-filter-bubbles.

4 Evaluation Methodology and Findings

These four measures — political stance, extremeness, diversity, and up-vote ratio — are given our entire focus as we answer the research questions presented in §2.

RQ1 — *How does a user's interaction with a political news recommender system affect the system's recommendation trajectory?* — considers whether users are presented with progressively more extreme/moderate, liberal/center/conservative, diverse/homogeneous, and enjoyable/less-desirable articles while interacting with the system. We answer RQ1 twice, once for the control group and once for the treatment group.

We first quantify the *initial* extremeness, political stance, and diversity states of the recommender system for a user (the `begin` value) using the top K-ranked articles for that user prior to the first article being shown to the user.[2] This initial top-K ranking is based on the model bootstrapped from the pre-questionnaire responses (see Table 2), i.e., the user's political stance and interest for the 11 topics. The *final* measures (`end`) are similarly computed after the last article has been presented. Whereas extremeness, political stance, and diversity are captured at a large-scale using the top-K-ranked articles for the user, the up-vote ratio is based on the up/down votes the user assigned to each of the presented articles, and hence the `begin` up-vote ratio is captured via the first ten articles that the user up/down-voted, and the `end` value is computed based on the last ten articles. More formally, for measure m let m_b be the `begin` value and m_e be the `end` value. For users in the control group and treatment group, separately, RQ1 considers whether $\delta_m = m_e - m_b$ is significantly different from zero.

In addition to these main effects, we also expect there to be heterogeneous effects based on user attributes. For example, Munson, Lee, and Resnick (2013) find that conservative and liberal users respond differently to a recommendation interface that shows users summaries of the partisan lean of their reading habits. In order to operationalize this, we consider the initial state of the recommendation system, which, recall, is seeded based on the user survey. To balance our ability to identify heterogeneous effects with ensuring a sufficient sample size, for each measure we assign users into one of three bins, based on the initial state of the system for that user. That is, for each measure, we rank all 102 users based on their `begin` values m_b. We then assign each user to `low`, `medium`, and `high` subgroups based on their rank,[3] and let $g \in \{$`all, low, medium, high`$\}$. For each measure, m, we compute $\delta_m^g = m_e^g - m_b^g$ and test for whether it is significantly different from zero.

RQ2 — *Do changes in the recommendation system's trajectory differ significantly for the control group versus the treatment group?* — employs a between-group comparison

[2]We used $K = 200$ in our experiments to give all topic and political stance combinations (11 topics, 5 political stances = $11 \times 5 = 55$ possibilities) a reasonable chance to be included in the top K articles.

[3]Each subgroup had $102/3 = 34$ users, but they do not always divide evenly into control and treatment groups. See §5 for a discussion of sample sizes.

to examine whether the begin to end differences are larger/smaller for users assigned to the control group relative to users assigned to the treatment group. More formally, let $\Delta_m^g = \delta_m^g(\text{Treatment}) - \delta_m^g(\text{Control})$. We test for whether Δ_m^g is significantly larger/smaller than zero. Such tests are conducted in the context of the all group as well as the low, medium, and high subgroups.

Findings

Figure 2 presents the results for each measure. The first graph in each row shows the begin (m_b) and end (m_e) values for all users within the control and treatment groups. The second, third, and fourth graphs in each row present the same information based on, respectively, the low, medium, and high subgroups. The final, rightmost graph in each row plots the changes in these data for each group (δ_m^g). The first four columns of Figure 2 are primarily used to answer RQ1, while the rightmost column of Figure 2 is primarily used to answer RQ2. Error bars in each plot are bootstrapped 95% confidence intervals (1000 bootstrap samples).

Table 3 presents t-tests to assess within-group comparisons (i.e., whether m_e^g and m_b^g are significantly different) as well as between-group comparisons (i.e., whether δ_m^g differs significantly between the control and treatment groups). A response to RQ1 is provided in the "Begin vs End" columns of each measure (1, 2, 4, 5, 7, 8, 10, 11), and p values are computed using two-tailed paired t-tests based on a comparison of the values of m_b^g and m_e^g.[4] A response to RQ2 is provided in the "Change" column of each measure (3, 6, 9, 12), where p values are computed using one-tailed non-paired t-tests.[5] Each measure is considered in turn below.

Extremeness Based on Liu et al. (2021), we expect extremeness to go up for the control group, as the feedback loops to exacerbate the filter bubbles and the simple interaction mechanism of up-voting/down-voting will be inadequate for course correction. For the treatment group, we expect mixed results, where the interaction tool to satisfy the requests of both challenge-averse and diversity-seeking groups (Munson and Resnick 2010).

RQ1: Within-group comparisons in Figure 2a show that extremeness decreased for the entire sample (all) for both the treatment ($\delta = -.20, p = .025$) and control groups ($\delta = -.08$, $p = .065$). For the low, medium, and high subgroups, the within-group comparison results are mixed. For users starting with either medium or high extremeness, those in the treatment group generated significant decreases in extremeness ($\delta = -.43, p = .005$ and $\delta = -.40, p = .041$, respectively), while the those in the control group experienced very little change. For the low subgroup, however, extremeness increased slightly for those in the treatment group ($\delta = .15, p = .196$), while it decreased slightly in the control group ($\delta = -.12, p = .200$). In sum, for those in the treatment group, extremeness decreased at statistically significant levels for the all group and the medium and high subgroups. For the control group, changes from beginning to end were insignificant.

RQ2: We observe between-group comparisons in the rightmost graph of Figure 2a and column 3 of Table 3, showing that the change in extremeness was consistently larger for the treatment group relative to the control group. These differences are statistically significant for the low, medium, and high subgroups ($\Delta = .28, p = .034$, $\Delta = -.31, p = .023$, and $\Delta = -0.39, p = .024$, respectively). That is, when extremeness decreased for the treatment group (medium and high subgroups), it decreased significantly more than the control group. When extremeness increased for the treatment group (low subgroup), it increased significantly more than the control group.

These results point to heterogeneous effects of the proposed system — empowered with the transparency and interaction tool, the treatment group was able to make bigger and more significant changes to the system. However, this did not result in reducing extremeness for everyone: users whose system was initially medium or high reduced extremeness, while users whose system was initially low increased extremeness.

Political Stance Our expectations for changes in political stance are the same as those for extremeness: the control group solidifies their initial political assignments, and the treatment group is able to make bigger changes (positive or negative). Figure 2b shows the average political stance of the top K-ranked articles for each subgroup. To provide a more intuitive explanation of this measure, we swap the low, medium, and high subgroup labels with their respective ideologies. Given that the average begin values for these subgroups were, respectively, -1.9, -1.0, and 0.9, we label the three subgroups "strong liberal," "liberal," and "conservative."[6]

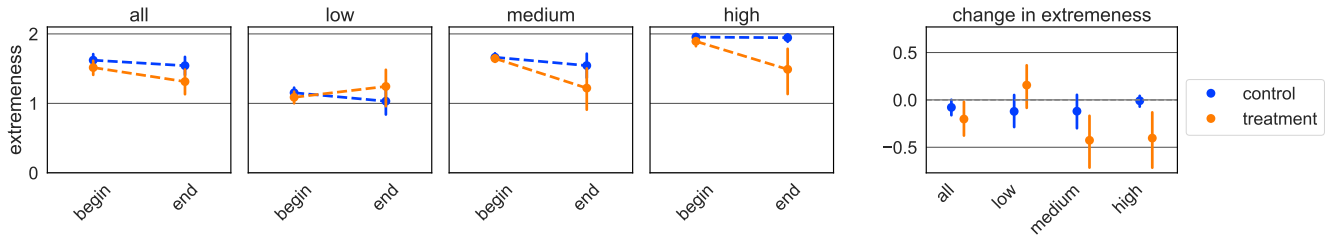
RQ1: Considering the entire sample (all), we observe that the political stance of users' news content shifted slightly towards the center for the treatment group ($\delta = .29, p = .084$) but changed little for the control group ($\delta = -.10, p = .296$). By subgroup, the most noticeable changes among users in the treatment group are for strong liberals ($\delta = .44, p = .050$) and liberals ($\delta = .62, p = .036$), who consumed significantly less partisan articles. For the control group, the most noticeable change is for conservative users, who also consumed less partisan articles ($\delta = -.43, p = .035$). These results are largely consistent with those of the extremeness measure: users generally moved the recommender towards less partisan news content.

RQ2: For the between-group comparisons (rightmost graph of Figure 2b and column 6 of Table 3), we observe

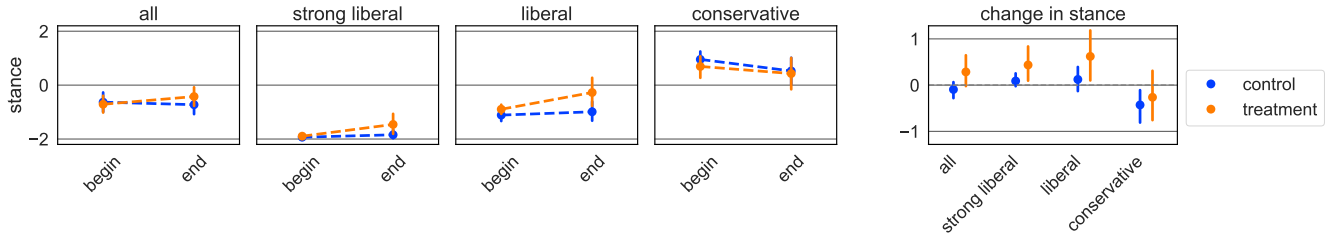
[4] Two-tailed t-tests examine whether the begin and end values are significantly different from each other; paired are used because these are the begin and end values for the same users.

[5] One-tailed t-tests examine whether the difference is significantly larger/smaller for the treatment group; non-paired are used because these are different users (i.e., control versus treatment).

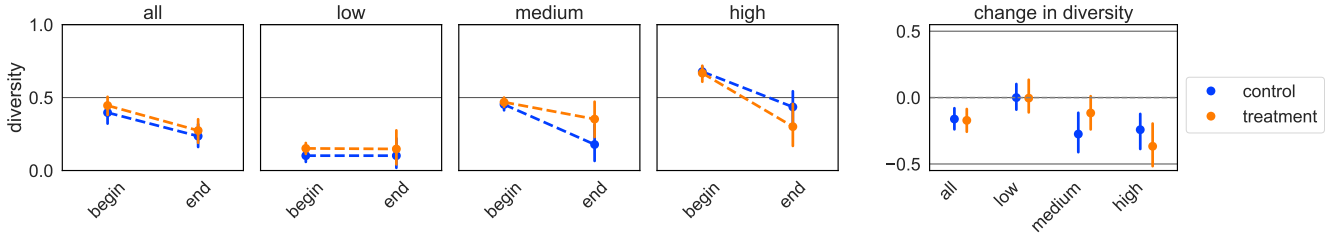
[6] Even though more people self-identified as Republican (47 users) than Democrat (35 users) in the pre-questionnaire (Table 1b), the initial state of the recommender was slanted slightly liberal. Recall that this initial state was based on user-provided responses to select Pew survey questions (Table 2); perhaps Republican users in our sample held more liberal views, or perhaps political ideology had shifted leftward since the Pew survey was published in 2017.



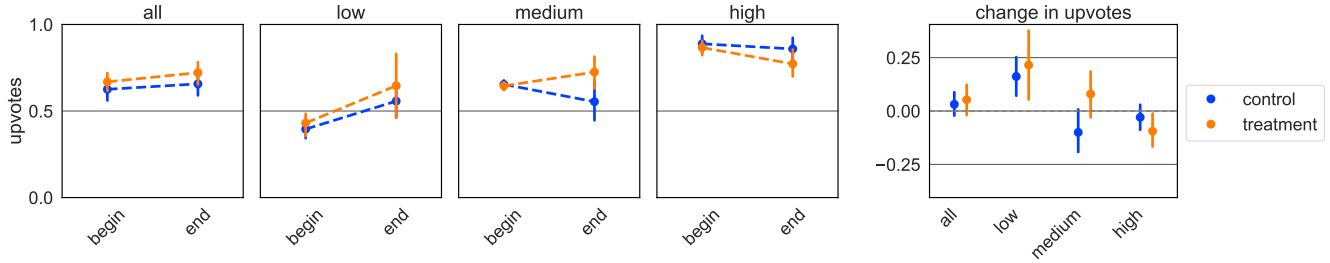
(a) Extremeness



(b) Political Stance



(c) Diversity



(d) Up-vote Ratio

Figure 2: Beginning (m_b^g) and ending (m_e^g) values, with bootstrapped 95% confidence intervals (1000 bootstrap samples), for each measure and subgroup, where subgroups are determined by the initial value of the system for each user. Rightmost graphs present changes over time (i.e., $\delta_m^g = m_e^g - m_b^g$).

that the treatment group experienced larger shifts towards the center than the control group when considering the `all` group ($\Delta = .38, p = .022$), as well as for strong liberals ($\Delta = .35, p = .064$) and liberals ($\Delta = .50, p = .055$). For conservatives, those in both the control and treatment groups directed the system towards the center, but the between-group difference was not statistically significant.

Diversity We expect the system to take the control group to less diverse articles (Liu et al. 2021);For the treatment group, we expect the interaction tool to either increase or decrease diversity depending on users' preexisting characteristics (Munson and Resnick 2010).

While extremeness and average political stance are certainly informative, two users with the same extremeness score might view very different types of news. For example, one user with an extremeness score of 1 might have read a diverse set of articles (e.g., from both $+1$ and -1 sources), while another user with an extremeness score of 1 might have viewed articles representing only a single political stance (e.g., only -1). To assess this, we analyze the political stance diversity of the top K-ranked articles, measured through normalized entropy of the political stance distribution of the articles (Eq. 4).

	Extremeness			Political Stance			Diversity			Up-vote Ratio		
	Begin vs End		Change	Begin vs End		Change	Begin vs End		Change	Begin vs End		Change
	C vs C	T vs T	C vs T	C vs C	T vs T	C vs T	C vs C	T vs T	C vs T	C vs C	T vs T	C vs T
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
All	.065	**.025**	.103	.296	.084	**.022**	**.000**	**.001**	.435	.292	.157	.324
Low	.200	.196	**.034**	.245	**.050**	.064	.999	.954	.482	**.003**	**.045**	.416
Medium	.209	**.005**	**.023**	.381	**.036**	.055	**.002**	.091	**.047**	.302	.118	**.033**
High	.774	**.041**	**.024**	**.035**	.225	.416	**.003**	**.000**	.091	.083	**.049**	.220

Table 3: p-values from t-tests of significance, $p \leq .05$ in bold. p-values smaller than 0.0125 (i.e., accounting for the Bonferroni correction for testing four hypotheses per measure) are in bold and underlined. Rows represent the aggregated group ("All") and all subgroups (*Political Stance* subgroups correspond with "strong liberal," "liberal," and "conservative").

RQ1: Figure 2c shows that, for the (`all`) group, diversity decreased significantly for both the control ($\delta = -.16, p < .001$) and treatment ($\delta = -.17, p = .001$) groups (c.f., Table 3, columns 7 and 8). This was not unexpected given the tendency for filter bubble-like conditions to arise. Specifically, users may find that the initial personalization of results by the system are excessively broad and not reflective of users' self-perceptions.

Considering subgroups separately, the most noticeable changes among the treatment group was the `high` subgroup, which decreased significantly in diversity ($\delta = -.37, p < .001$), and to a lesser extent the `medium` subgroup, which decreased slightly ($\delta = -.12, p = .091$). For the control group, users in both the `medium` ($\delta = -.27, p = .002$) and `high` ($\delta = -.24, p = .003$) subgroups showed significant decreases in diversity. There was no discernible change in diversity of news content for users in the `low` subgroup for either the control or treatment groups.

RQ2: Regarding between-group comparisons (rightmost graph of Figure 2c and column 9 of Table 3), there are no noticeable differences between the treatment and control groups when considering the diversity of news viewed by `all` users. However, for those in the `medium` subgroup, the control group exhibited a sharper decrease in diversity than the treatment group ($\Delta = .15, p = .047$). We suspect that user feedback led to homogenization of the content recommender in the control group, while the enhanced UI available to users in the treatment group allowed them to maintain higher diversity.

Up-vote Ratio Based on the large body of research in machine learning and recommender systems, we expect the up-vote ratio to increase for users in both the control and treatment groups, as the system collects more training data and learn the user preferences better. However, for the treatment group, users can effect the state of the recommender directly, and hence can lead to unexpected results if the users change the system drastically by significantly overpowering the underlying content-based recommender system.

RQ1: Regarding within-group comparisons of the up-vote ratio, presented in Figure 2d and columns 10 and 11 of Table 3, we observe for `all` users only a slight increase in up-vote ratios for both the controland treatmentgroups. However, when considering subgroups separately, we observe

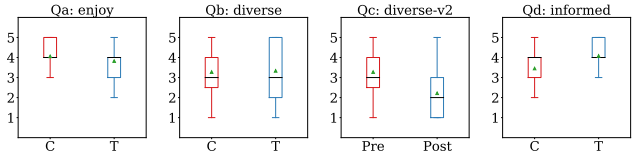


Figure 3: Post-questionnaire results. 1: strongly disagree, 5: strongly agree. Triangles represent means.

significant increases for the `low` subgroup for both control ($\delta = .17, p = .003$) and treatment ($\delta = .19, p = .045$) groups. Conversely, for the `high` subgroup, the treatment group exhibited a significant decrease in its up-vote ratio ($\delta = -.09, p = .049$); the control group also exhibited a decrease to a lesser extent ($\delta = -.06, p = .083$). We will revisit this finding when considering the results from the post-questionnaire analysis.

RQ2: For the between-group comparisons (rightmost graph of Figure 2d and column 12 of Table 3), the most notable difference between control and treatment groups is for the `medium` subgroup ($\Delta = .14, p = .033$), where up-vote ratios increased for the treatment group but decreased for the control group.

Summary of results Our analysis showed that the proposed interaction mechanism allowed users to exert greater control over the recommended articles, albeit in varying ways. The treatment group reduced *extremeness* significantly when the initial state was `medium` or `high` extreme, whereas the control group did not shift significantly. Yet, for *political stance*, all subgroups moved towards the center for both the control and treatment groups. Reductions in extremeness — both formally and in terms of political stance — may have been reduced, but this was accompanied with a decrease in the *diversity* of users' news content. The results was also mixed for the *up-vote ratio*: users who were initially less satisfied with news content (`low` up-vote ratio) could significantly improve their news consumption experience; those initially more satisfied (`high` up-vote ratio) became more dissatisfied.

Post-questionnaire analysis Figure 3 presents the post-questionnaire results. We find that users in the treatment group, though tending to have higher up-vote ratios in general, expressed less preference for reading articles based on

the presentation of the recommender system (*Qa: enjoy*). It is possible that the added cognitive load of the enhanced interface hampered the user experience, bu it is also possible that it encouraged users to explore articles that were less aligned with their stances and interests, leading to unpleasant cognitive dissonance, particularly for those experiencing moderate enjoyment, i.e., the `medium` subgroup in Figure 2d. Future work should integrate composite measures of enjoyment as well as conduct extensive follow-up interviews to understand the reasoning behind these differences.

There were no differences between groups in terms of the perceived level of exposure to the political diversity of the news articles to which users were presented (*Qb: diverse*); users from both the control and treatment groups agreed that they had been presented with diverse political perspectives. However, after answering *Qb*, when individuals in the control group were presented with the political stance distribution of the articles they were actually shown (see Figure 5 in Appendix as an example), and when they were asked a second time whether they thought they had been exposed to a diverse set of news articles, we observe a substantial drop in agreement. Presented in *Qc: diverse-v2*, the first response ("pre") is significantly higher than the second response ("post"). This indicates that when provided with transparency about the nature of the recommender system, users from the control group eventually realized that they were presented with less diverse articles than they thought. Finally, users in the treatment believed much more strongly that, having participated in this study, they were more informed about how news recommender systems work (*Qd: informed*). Taken together, the post-questionnaire results indicate that the enhanced UI keeps the user better informed, but it may come at a cost of greater UI complexity.

5 Discussion, Social Impacts, Limitations

Our efforts to understand user engagement and news content delivery in a real-world setting have shown that, with the right information and interaction tools, people can counter filter bubbles in some aspects (extremeness) while reinforcing them in other ways (diversity). If user interaction tools were made available at scale, one might observe a decrease in polarizing content. This would facilitate opportunities for discourse among politically disparate groups — conditions fundamental for healthy democratic institutions and for the development of policies reflecting the public's opinions and preferences (Habermas 1989; Lewandowsky et al. 2012).

Of course, many other factors are at play here, including trust in the media (Guess et al. 2021), personal experiences (Druckman et al. 2021), and online ideological segregation (Bail 2021; Iandoli, Primario, and Zollo 2021; Mosleh et al. 2021). Users' political preferences are particularly important, as existing research shows that individuals, especially those with firm beliefs, prefer to receive ideologically consistent (Stier et al. 2020; A. G. Ekström and Tsapos 2023), sometimes extreme content (Zhang, Zhu, and Caverlee 2023). This has prompted research focusing on how to "nudge" people out of their filter bubbles.[7]

In terms of limitations to our work, as in any controlled study, the behavior of users in a short-horizon study may differ somewhat from their long-term use of commercial recommendation systems. Furthermore, while our results are based on $\sim$100 total hours of real system interactions from 102 users in a controlled environment, future work that considers larger sample sizes in natural environments would enhance the external validity of this study and also enable a closer investigation of how filter bubble behaviors vary by topic. To explore the impact of the small sample size on the main results, a post-hoc power analysis revealed that our significant findings in Table 3 have moderate to high power ($\sim$.7-.99). However, some of the comparisons not found to be statistically significant may be in part influenced by these results' lower power.

Finally, we have in many ways provided users with a more nuanced control interface (Figure 1) than prior work, where the users are able to express a variety of political lean and interest across 11 topics (for e.g., liberal on abortion with low interest, conservative on taxes with high interest, etc.), rather than a single political identity (e.g., Democrat, Republican). While this implementation can allow the users to have diversity across topics, it may also lead to a lack of diversity on a given topic. A more comprehensive approach could be implemented where users can express, for example, that they would like to see only liberal views on abortion while simultaneously preferring both conservative *and* liberal content on taxes. Such an approach would of course make the interface much more complicated and potentially overwhelming for users, and future work can determine whether the added complexity of such interfaces would be justified.

6 Conclusions

News recommendation systems can affect civic discourse. Thus, any variants in such systems have the potential to exacerbate hyperpartisanship and misinformation. The goal of this study was to understand the effect that transparency and interaction mechanisms have on these systems. Our results suggest that giving users greater control over news recommendation systems can substantially change the type of news they see. Users who are initially shown extreme content can more efficiently move towards less extreme content if desired; likewise, users who are initially shown less extreme content can move towards more extreme content if desired. We find heterogeneous effects based on the initial state of the recommender system, and we call for future work that investigates further how user attributes interact with new interaction mechanisms for recommendation systems.

Acknowledgements

This work was supported by the NSF Award #1927407. AC was funded in part by the Tulane's Jurist Center for Artificial Intelligence and by Tulane's Center for Community-Engaged Artificial Intelligence.

[7] See, for example, Srba et al. (2023); Masrour et al. (2020); Pennycook et al. (2021), although the evidence is mixed on whether these methods help (Aslett et al. 2022) or foster a "boomerang effect" (Casas, Menchen-Trevino, and Wojcieszak 2023).

References

A. G. Ekström, E. J. O., G. Madison; and Tsapos, M. 2023. The search query filter bubble: effect of user ideology on political leaning of search results through query selection. *Information, Communication & Society*, 0(0): 1–17.

Allen, J.; Martel, C.; and Rand, D. G. 2022. Birds of a feather don't fact-check each other: Partisanship and the evaluation of news in Twitter's Birdwatch crowdsourced fact-checking program. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*, 1–19.

Aslett, K.; Guess, A. M.; Bonneau, R.; Nagler, J.; and Tucker, J. A. 2022. News credibility labels have limited average effects on news diet quality and fail to reduce misperceptions. *Science Advances*, 8(18): eabl3844.

Bail, C. 2021. *Breaking the Social Media Prism: How to Make Our Platforms Less Polarizing*. Princeton University Press.

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

Bhargava, R.; Chung, A.; Gaikwad, N. S.; Hope, A.; Jen, D.; Rubinovitz, J.; Saldías-Fuentes, B.; and Zuckerman, E. 2019. Gobo: A System for Exploring User Control of Invisible Algorithms in Social Media. In *Conference Companion Publication of the 2019 on Computer Supported Cooperative Work and Social Computing*, 151–155.

Brewer, M. D. 2005. The rise of partisanship and the expansion of partisan conflict within the American electorate. *Political Research Quarterly*, 58(2): 219–229.

Casas, A.; Menchen-Trevino, E.; and Wojcieszak, M. 2023. Exposure to extremely partisan news from the other political side shows scarce boomerang effects. *Political Behavior*, 45(4): 1491–1530.

Castells, P.; Hurley, N.; and Vargas, S. 2021. Novelty and diversity in recommender systems. In *Recommender systems handbook*, 603–646. Springer.

Chesnais, P.; Mucklo, M.; and Sheena, J. 1995. The Fishwrap personalized news system. In *Proceedings of the Second International Workshop on Community Networking 'Integrated Multimedia Services to the Home'*, 275–282.

Claypool, M.; Gokhale, A.; Miranda, T.; Murnikov, P.; Netes, D.; and Sartin, M. 1999. Combing content-based and collaborative filters in an online newspaper. In *Workshop on Recommender Systems-Implementation and Evaluation*.

Covington, P.; Adams, J.; and Sargin, E. 2016. Deep neural networks for youtube recommendations. In *ACM conference on recommender systems*, 191–198.

Davis, F. D. 1989. Perceived usefulness, perceived ease of use, and user acceptance of information technology. *MIS quarterly*, 319–340.

Doherty, C.; Kiley, J.; and Johnson, B. 2017. Political typology reveals deep fissures on the right and left: Conservative Republican groups divided on immigration,'openness'. *Pew Research Center*.

Druckman, J. N.; Klar, S.; Krupnikov, Y.; Levendusky, M.; and Ryan, J. B. 2021. Affective polarization, local contexts and public opinion in America. *Nature human behaviour*, 5(1): 28–38.

Faridani, S.; Bitton, E.; Ryokai, K.; and Goldberg, K. 2010. Opinion space: a scalable tool for browsing online comments. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 1175–1184.

Guess, A. M.; Barberá, P.; Munzert, S.; and Yang, J. 2021. The consequences of online partisan media. *Proceedings of the National Academy of Sciences*, 118(14).

Guo, H.; Tang, R.; Ye, Y.; Li, Z.; and He, X. 2017. DeepFM: a factorization-machine based neural network for CTR prediction. *arXiv preprint arXiv:1703.04247*.

Habermas, J. 1989. *The structural transformation of the public sphere: An inquiry into a category of a bourgeois society*. Cambridge: MIT Press.

Harambam, J.; Bountouridis, D.; Makhortykh, M.; and Van Hoboken, J. 2019. Designing for the better by taking users into account: A qualitative evaluation of user control mechanisms in (news) recommender systems. In *ACM Conference on Recommender Systems*, 69–77.

Iandoli, L.; Primario, S.; and Zollo, G. 2021. The impact of group polarization on the quality of online debate in social media: A systematic literature review. *Technological Forecasting and Social Change*, 170: 120924.

Jannach, D.; Naveed, S.; and Jugovac, M. 2016. User control in recommender systems: Overview and interaction challenges. In *International Conference on Electronic Commerce and Web Technologies*, 21–33. Springer.

Kahan, D. M. 2015. The politically motivated reasoning paradigm, part 1: What politically motivated reasoning is and how to measure it. *Emerging trends in the social and behavioral sciences: An interdisciplinary, searchable, and linkable resource*, 1–16.

Kamba, T.; Bharat, K.; and Albers, M. C. 1995. The Krakatoa Chronicle-an interactive, personalized newspaper on the Web. In *In Proc. 4th Intl. WWW Conf.*

Kunaver, M.; and Požrl, T. 2017. Diversity in recommender systems–A survey. *Knowledge-based systems*, 123: 154–162.

Levendusky, M.; and Malhotra, N. 2016. Does media coverage of partisan polarization affect political attitudes? *Political Communication*, 33(2): 283–301.

Lewandowsky, S.; Ecker, U. K. H.; Seifert, C. M.; Schwarz, N.; and Cook, J. 2012. Misinformation and its correction: continued influence and successful debiasing. *Psychological Science in the Public Interest*, 13(3): 106–131.

Li, Z.; Dong, Y.; Gao, C.; Zhao, Y.; Li, D.; Hao, J.; Zhang, K.; Li, Y.; and Wang, Z. 2023. Breaking Filter Bubble: A Reinforcement Learning Framework of Controllable Recommender System. In *Proceedings of the ACM Web Conference 2023*, 4041–4049.

Liu, P.; Shivaram, K.; Culotta, A.; Shapiro, M. A.; and Bilgic, M. 2021. The Interaction between Political Typology and Filter Bubbles in News Recommendation Algorithms. In *Proceedings of the Web Conference 2021*, 3791–3801.

Lodge, M.; and Taber, C. S. 2000. Three steps toward a theory of motivated political reasoning. In Lupia, A.; McCubbins, M. D.; and Popkin, S. L., eds., *Elements of Reason: Cognition, choice, and the bounds of rationality*, 183–213. Cambridge: Cambridge University Press.

Masrour, F.; Wilson, T.; Yan, H.; Tan, P.-N.; and Esfahanian, A. 2020. Bursting the Filter Bubble: Fairness-Aware Network Link Prediction. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 841–848.

Mitova, E.; Blassnig, S.; Strikovic, E.; Urman, A.; Hannak, A.; de Vreese, C. H.; and Esser, F. 2023. News recommender systems: A programmatic research review. *Annals of the International Communication Association*, 47(1): 84–113.

Mosleh, M.; Martel, C.; Eckles, D.; and Rand, D. G. 2021. Shared partisanship dramatically increases social tie formation in a Twitter field experiment. *Proceedings of the National Academy of Sciences*, 118(7).

Munson, S. A.; Lee, S. Y.; and Resnick, P. 2013. Encouraging reading of diverse political viewpoints with a browser widget. In *AAAI conference on weblogs and social media*.

Munson, S. A.; and Resnick, P. 2010. Presenting Diverse Political Opinions: How and How Much. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 1457–1466.

Ookalkar, R.; Reddy, K. V.; and Gilbert, E. 2019. Pop: Bursting News Filter Bubbles on Twitter Through Diverse Exposure. In *Conference Companion Publication of Computer Supported Cooperative Work and Social Computing*, 18–22.

Pariser, E. 2011. *The filter bubble: What the Internet is hiding from you*. Penguin UK.

Pennycook, G.; Epstein, Z.; Mosleh, M.; Arechar, A. A.; Eckles, D.; and Rand, D. G. 2021. Shifting attention to accuracy can reduce misinformation online. *Nature*, 592(7855): 590–595.

Resnick, P.; Garrett, R. K.; Kriplean, T.; Munson, S. A.; and Stroud, N. J. 2013. Bursting Your (Filter) Bubble: Strategies for Promoting Diverse Exposure. In *Proceedings of the 2013 Conference on Computer Supported Cooperative Work Companion*, 95–100.

Robertson, R. E.; Green, J.; Ruck, D.; Ognyanova, K.; Wilson, C.; and Lazer, D. 2021. Engagement Outweighs Exposure to Partisan and Unreliable News within Google Search. *arXiv preprint arXiv:2201.00074*.

Rodriguez, C. G.; Moskowitz, J. P.; Salem, R. M.; and Ditto, P. H. 2017. Partisan selective exposure: The role of party, ideology and ideological extremity over time. *Translational Issues in Psychological Science*, 3(3): 254.

Shivaram, K.; Liu, P.; Shapiro, M. A.; Bilgic, M.; and Culotta, A. 2022. Reducing Cross-Topic Political Homogenization in Content-Based News Recommendation. In *Proceedings of the 16th ACM conference on Recommender Systems*.

Smyth, B.; and McClave, P. 2001. Similarity vs. Diversity. In Aha, D. W.; and Watson, I., eds., *Case-Based Reasoning Research and Development*, 347–361. Berlin, Heidelberg: Springer Berlin Heidelberg.

Srba, I.; Moro, R.; Tomlein, M.; Pecher, B.; Simko, J.; Stefancova, E.; Kompan, M.; Hrckova, A.; Podrouzek, J.; Gavornik, A.; and Bielikova, M. 2023. Auditing YouTube's Recommendation Algorithm for Misinformation Filter Bubbles. *ACM Trans. Recomm. Syst.*, 1(1).

Steck, H. 2018. Calibrated recommendations. In *ACM conference on recommender systems*, 154–162.

Stier, S.; Kirkizh, N.; Froio, C.; and Schroeder, R. 2020. Populist attitudes and selective exposure to online news: A cross-country analysis combining web tracking and surveys. *The International Journal of Press/Politics*, 25(3): 426–446.

Taijala, T. T.; Willemsen, M. C.; and Konstan, J. A. 2018. Movieexplorer: building an interactive exploration tool from ratings and latent taste spaces. In *Proceedings of the 33rd Annual ACM Symposium on Applied Computing*, 1383–1392.

Tang, J.; Hu, X.; and Liu, H. 2013. Social recommendation: a review. *Social Network Analysis and Mining*, 3(4): 1113–1133.

Tewksbury, D.; and Riles, J. M. 2015. Polarization as a Function of Citizen Predispositions and Exposure to News on the Internet. *Journal of Broadcasting & Electronic Media*, 59(3): 381–398.

Tsai, C.-H.; and Brusilovsky, P. 2019. Exploring Social Recommendations with Visual Diversity-Promoting Interfaces. *ACM Trans. Interact. Intell. Syst.*, 10(1).

Törnberg, P. 2022. How digital media drive affective polarization through partisan sorting. *Proceedings of the National Academy of Sciences*, 119(42): e2207159119.

Venkatesh, V.; and Bala, H. 2008. Technology acceptance model 3 and a research agenda on interventions. *Decision sciences*, 39(2): 273–315.

Venkatesh, V.; Morris, M. G.; Davis, G. B.; and Davis, F. D. 2003. User Acceptance of Information Technology: Toward a Unified View. *MIS Quarterly*, 27(3): 425–478.

Wang, W.; Feng, F.; Nie, L.; and Chua, T.-S. 2022. User-controllable Recommendation Against Filter Bubbles. *arXiv preprint arXiv:2204.13844*.

Ying, R.; He, R.; Chen, K.; Eksombatchai, P.; Hamilton, W. L.; and Leskovec, J. 2018. Graph convolutional neural networks for web-scale recommender systems. In *Proceedings of the 24th ACM SIGKDD international conference on knowledge discovery & data mining*, 974–983.

Zhang, H.; Zhu, Z.; and Caverlee, J. 2023. Evolution of Filter Bubbles and Polarization in News Recommendation. In Kamps, J.; Goeuriot, L.; Crestani, F.; Maistro, M.; Joho, H.; Davis, B.; Gurrin, C.; Kruschwitz, U.; and Caputo, A., eds., *Advances in Information Retrieval*, 685–693. Cham: Springer Nature Switzerland.

Ziegler, C.-N.; McNee, S. M.; Konstan, J. A.; and Lausen, G. 2005. Improving recommendation lists through topic diversification. In *Proceedings of the 14th international conference on World Wide Web*, 22–32.

7 Ethics Checklist

1. For most authors...
   (a) Would answering this research question advance science without violating social contracts, such as violating privacy norms, perpetuating unfair profiling, exacerbating the socio-economic divide, or implying disrespect to societies or cultures? Yes, the study conducted was granted IRB approval by the authors' institutions.
   (b) Do your main claims in the abstract and introduction accurately reflect the paper's contributions and scope? Yes.
   (c) Do you clarify how the proposed methodological approach is appropriate for the claims made? Yes, see "Evaluation Methodology and Findings" section.
   (d) Do you clarify what are possible artifacts in the data used, given population-specific distributions? Yes, see "Our Approach" section.
   (e) Did you describe the limitations of your work? Yes, see "Discussion, Social Impacts, Limitations" section.
   (f) Did you discuss any potential negative societal impacts of your work? Yes, see "Discussion, Social Impacts, Limitations" section.
   (g) Did you discuss any potential misuse of your work? Yes, see "Discussion, Social Impacts, Limitations" section.
   (h) Did you describe steps taken to prevent or mitigate potential negative outcomes of the research, such as data and model documentation, data anonymization, responsible release, access control, and the reproducibility of findings? Yes, see "Our Approach" section.
   (i) Have you read the ethics review guidelines and ensured that your paper conforms to them? Yes.
2. Additionally, if your study involves hypotheses testing...
   (a) Did you clearly state the assumptions underlying all theoretical results? Yes, see "Our Approach" and "Evaluation Methodology and Findings" sections.
   (b) Have you provided justifications for all theoretical results? Yes, see "Related Work" and "Our Approach" sections.
   (c) Did you discuss competing hypotheses or theories that might challenge or complement your theoretical results? Yes, see "Related Work" section.
   (d) Have you considered alternative mechanisms or explanations that might account for the same outcomes observed in your study? Yes, see "Discussion, Social Impacts, Limitations" section.
   (e) Did you address potential biases or limitations in your theoretical framework? Yes, see "Our Approach" section.
   (f) Have you related your theoretical results to the existing literature in social science? Yes, see "Introduction" and "Related Work" sections.
   (g) Did you discuss the implications of your theoretical results for policy, practice, or further research in the social science domain? Yes, see "Discussion, Social Impacts, Limitations" section.
3. Additionally, if you are including theoretical proofs...
   (a) Did you state the full set of assumptions of all theoretical results? N/A
   (b) Did you include complete proofs of all theoretical results? N/A
4. Additionally, if you ran machine learning experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? Yes, the code and the anonymous user survey and interaction data are available at https://github.com/IIT-ML/icwsm-2024-filter-bubbles.
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? Yes, see "Our Approach" section.
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? Yes, please see Figures 2 and 3.
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? N/A
   (e) Do you justify how the proposed evaluation is sufficient and appropriate to the claims made? Yes, please see "Measures" subsection in the Approach section (Section 3)
   (f) Do you discuss what is "the cost" of misclassification and fault (in)tolerance? N/A
5. Additionally, if you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? Yes, see the "Datasets" subsection of Section 3.
   (b) Did you mention the license of the assets? NA
   (c) Did you include any new assets in the supplemental material or as a URL? Yes, the code and the anonymous user survey and interaction data are available at https://github.com/IIT-ML/icwsm-2024-filter-bubbles.
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? Yes, see "Technical Appendix" section. We also restate here that the study was granted IRB approval by the authors' institutions.
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? Yes, please see the Appendix.
   (f) If you are curating or releasing new datasets, did you discuss how you intend to make your datasets FAIR)? N/A
   (g) If you are curating or releasing new datasets, did you create a Datasheet for the Dataset)? N/A
6. Additionally, if you used crowdsourcing or conducted research with human subjects...
   (a) Did you include the full text of instructions given to participants and screenshots? Yes, see "Technical Appendix" section.

(b) Did you describe any potential participant risks, with mentions of Institutional Review Board (IRB) approvals? Yes.

(c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? Yes, please see the "Compensation" subsection in the appendix.

(d) Did you discuss how data is stored, shared, and deidentified? Yes, please see Appendix.

Technical Appendix

User recruitment

The study conducted was granted IRB approval by the authors' institutions. The study was conducted through Amazon Mechanical Turk and no personally identifiable information was collected at any time. The initial screening survey contained six demographic questions: gender, age, race, self-identified political stance, education level, and annual income. Additionally, we ask these individuals the following three questions:

1. *Where do you get most of your information about current news events?* Response options include *printed, online, TV,* among others.
2. *How often do you read or watch news about U.S. politics, policies, or the economy?* The response is a five-point Likert scale ranging from *Never* to *Always*.
3. *How often do you use fact-checking websites (e.g., PolitiFact, Snopes, FactCheck, etc.)?* The response option is a five-point Likert scale ranging from *Never* to *Always*.

The survey ends with a political literacy qualification section containing three questions meant to assess a basic knowledge of U.S. politics. Those who answered at least two of the three following questions correctly are invited to participate in the full-scale recommendation study:

1. *Which of the following is the most conservative news source?* Response options are *MSNBC, New York Times, Fox News, The Guardian*.
2. *Among the following, who is the most liberal politician?* Response options are *Ted Cruz, Bernie Sanders, Donald Trump, Lindsey Graham.*
3. *Which state among the following recently enacted a restrictive abortion law?* Response options are *Texas, Massachusetts, New York, California*. (The study was conducted immediately after Texas drafted its Texas Heartbeat Act in September 2021.)

User interface

An example recommendation page is shown in Figure 4. Users have access to the title, date, and content of the article. At the bottom of the page are the up-vote, down-vote, and skip buttons. The example in Figure 4 is for a user in the treatment group; those in the control group see a similar page with the sentence and hyperlink at the top (beginning with "You may see how...") removed.

Figure 5 shows a sample transparency figure shown to users in the control group.

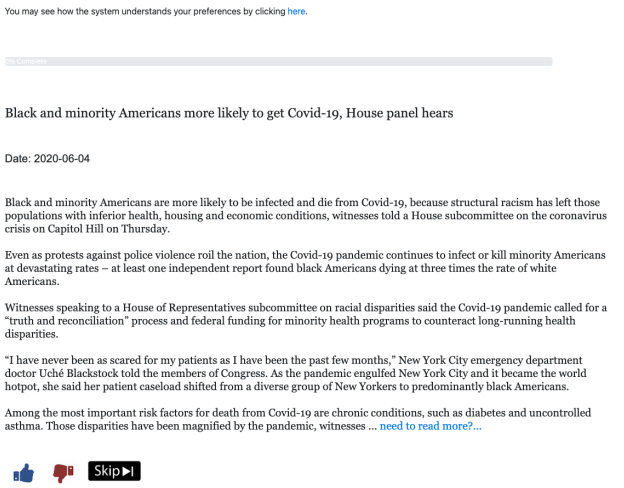


Figure 4: An example recommendation page for a user in the treatment group. The sentence at the top "You may see how the system understands your preferences by clicking here," was available only to users in the treatment group, connecting them to the transparency and interaction tool.

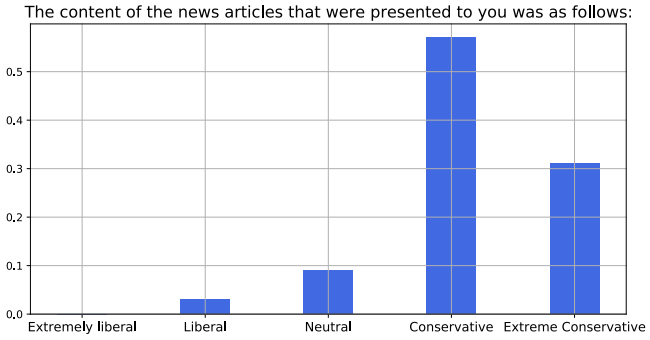


Figure 5: A sample transparency figure provided to users in the control group between answering *Qb* and *Qc*.

Instructions to users

Here are the instructions users were provided at each stage.

Before pre-Questionnaire *Thank you for your participation. Please take 3-5 minutes to answer the following questions about American policy issues. Click the "Begin" button to start.*

Selecting interests *Regardless of your previous responses, how interested would you be in reading news articles about the following topics?*

Before the first recommendation is shown Both control and treatment users: *Thank you for your responses thus far. You will now assess a number of newspaper articles according to your personal preferences. Your options after reading each article are to* give a thumbs-up *(you enjoyed reading the article),* give a thumbs-down *(you did not enjoy reading the article), or* skip *(you had no strong feelings about the article).* The users in the treatment group was shown the

following additional two sentences *You will periodically be exposed to information about how the system understands your preferences about the news, and you will have opportunities to view and modify this information. Please click the button below to begin.*

On the interaction page (treatment group only) *Below is a description of how the system understands your news-related preferences. You have two options.: 1) You may move the "Political Stance" slider to receive more articles in your preferred stance. 2) You may move the "Interest" slider to adjust the number of articles on a topic. When you have completed making changes, if any, click the SUBMIT button to read more news articles After all changes have been made, click SUBMIT to continue reading newspaper articles. You can revert to your original preferences based on your survey responses by clicking REVERT button.*

After all recommendations and before post-questionnaire *Thank you for your participation. Now you are going to answer the last several post-questionaires to get your unique token. If you have any questions of this survey, please contact author-info-removed-for-anonymous-review.*

Post-questionnaire *To what extent do you agree or disagree with the following statements.*

At the end of the study *Thank you for your participation. Here is your hash string that you need to copy and paste in Amazon Mturk website. If you have any questions of this survey, please contact author-info-removed-for-anonymous-review.*

Attention check

We implemented an attention-checking mechanism to ensure that users are not randomly clicking up/down-vote buttons. Ten articles focusing on science-related news, a politically neutral topic, had embedded in the article content instructions for the user to respond in a specific way regardless of their views, i.e. to up-vote, down-vote, or skip the article. Users who failed to click on the correct button on five articles or more were excluded from the study. Of the 146 users who participated, 44 failed this attention check and were removed from the study.

Compensation

Participants were paid $1 for completing the brief pre-screening survey and $15 for completing the full user study. Based on the time spent in the study, we estimate this wage to be $15/hour.