Pernicious Personalization
A Political Experiment with the Twitter Recommender System

Benjamin Guinaudeau, Simon Roth and Fabio Votta
University of Konstanz and University of Amsterdam

May, 3rd 2021

Abstract

Although social media only recently emerged, the accumulation of evidence undermining the 'echo chamber' hypothesis is striking. While self-selective exposure to congruent content - the echo chamber - is not as salient as expected, the ideological bias induced primarily by algorithmic selection - the filter bubble - has been less scrutinized in the literature. In this study, we propose a new experimental research design to investigate recommender systems. To avoid any behavioral confounder, we rely on automated agents, which 'treat' the algorithm with ideological and behavioral cues. For each agent, we compare the ideological slant of the recommended timeline with the ideological slant of the chronological timeline and, hence, isolate the ideological bias of the recommender system. This allows us to investigate two main questions : (1) how much bias is induced by the recommender system? (2) what role is played by implicit and explicit cues, when triggering ideological recommendations?

The experiment has been pre-registrated[1] and features 170 automated agents, which were active for three weeks before and three weeks after the 2020 American presidential election. We find that, after three weeks of delivering ideological cues (following accounts and liking content), the average algorithmic bias is about 5%. In other words, the timeline as structured by the algorithm entails 5% less cross-cutting content than it does when it is structured chronologically. While the algorithm relies on both implicit and explicit cues to formulate recommendations, the effect of implicit cues is significantly stronger. This study is, up to our knowledge, the first experimental assessment of the ideological bias induced by the recommender system of a major social media platform. Recommendations are biased and rely above all on behavioral cues unwarily and passively shared by the user. As affective polarization becomes a greater contemporary challenge, our results raise important normative questions about the possibility of opting-out from the ideological bias of recommender systems. In addition, it points out that more transparency is urgently needed around the recommendation questions: How are algorithms trained? What cues or features do they use? Against which biases have they been tested? In parallel, the results demonstrate the failure of 'in-house bias correction' and calls for an external auditing framework, that would facilitate this kind of research and crowd-sources the scrutiny of recommender systems.

[1] https://osf.io/5kwpr

1 Introduction

Social media and the role it plays in our lives has evolved drastically in the past decade. With the advent of YouTube, Twitter and Instagram, the social media world has further diversified since it's inception of its modern variant with the launch of Facebook in 2004. All grown up now in 2020, the new social media hit TikTok is one of the most downloaded smartphone applications in the world. The numbers are impressive: an average American spends more than 2 hours a day on social media (Statista 2019). Even though each platform has its own affordances, they all need to solve the problem of content oversupply. Social media differs from classical media outlets, such as newspaper or television, because it democratizes content production. Paired with increasing access to the internet and availability of smartphones, social media has dramatically sunk the cost of producing online content in the last twenty years. Almost anybody can now start a YouTube channel, comment on politics on Twitter or publish pictures on Instagram. In 2015, Youtube estimated that 500 hours of videos were uploaded every minute. One year before Twitter declared that about 500 million tweets were produced every day. These mind-boggling numbers go far beyond what any human could consume, even if they would spend 100% of its time on social media.

To facilitate navigation of the enormous volume of content, each social media platform had to adopt a specific set of tools. On most platforms, users are asked to select specific content producers and opt-in to see their content (for example, by becoming friends on Facebook, following Twitter or Instagram accounts, subscribing to a YouTube channel). However, the produced volume of individually selected producers still exceeds any reasonable consumption capacity. Therefore, on top of the active selection performed by the users, social media platforms use algorithmic recommendation to reduce the content supply to a manageable size.

Recommender systems (RS) deployed by social media, follow one main purpose: optimizing user experience on the platform. The opacity around these tools makes it very hard to describe them, as most questions remain unanswered: what method do they rely on? What kind of data is the RS trained on? Which metrics are used for optimization? To what extent are they tested for inherent bias and what is implemented to mitigate such biases? For instance, in the context of this paper we tried to gather all the information we could on the RS deployed by Twitter. After skipping all the technical academic papers proposing new network architecture to perform 'better' recommendation, information becomes scarce. In a Twitter blog post from 2020, two machine learning engineers at Twitter acknowledge that RS "aim to maximize user satisfaction as well as other key business objectives" (Twitter 2020b). Another blog post from 2019 indicates that user engagement is one of the core metrics on which RS are trained (Twitter 2019b). In 2019, Ashish Bansal, Senior Engineer Manager in charge of recommendations at Twitter, gave a talk on challenges of scaling up recommender systems (Ashish Bansal 2019). He states that collaborative filtering is particularly well-suited for Twitter data (in opposition to content-based recommendation). Finally, a blog post from 2017 presents in greater detail how Twitter recommends tweets (Twitter 2017). We learn here that models are informed by ”a list of considered features and their varied interactions”, which includes the tweet itself, its author and the consumer. Once again, the post speaks of a ”set of metrics we use here usually relate more directly to usage and enjoyment of Twitter”. To summarise, Twitter only discloses very imprecise information on its recommender system, in the absence of more details, we can only rely on guessing when investigating their RS.

Because RS prioritize content, they are prone to introduce systematic bias in their recommendations. For example in September 2020, the picture cropping system used by Twitter was the subject of a controversy. The underlying algorithm is meant to automatically segment uploaded pictures and produce a standardized preview that would fit well with Twitter’s user interface. When testing the cropping system with pictures representing the portraits of several persons, the algorithm seemed to systematically prioritize white people over people of color. In a response to the controversy, Twitter’s Chief Design Officer Dantley Davis, explained that they tested the algorithm for such bias. Right after he claims: ”while our analyses to date haven’t shown racial or gender bias, we recognize that the way we automatically crop photos means there is a potential for harm” (Twitter 2020a). This example demonstrates that recommendation bias can arise without the explicit intent of discrimination. Recommender systems are trained on past human behavior or digital traces and therefore reproduce human bias. This example raises many normative questions about the potential bias of recommender systems: Beyond racial bias, what kind of other bias could algorithmic recommendations harbor? How can we properly test for bias? And if biases are found, how can social media platforms be hold accountable for their tools? To what extent do RS not only reproduce but also amplify bias?

In a nutshell, as our time on social media increases, recommender systems become increasingly relevant in our lives. There is a tendency to believe that algorithms ”decide rationally” because it is based on data as opp osed to human decisions that are seen as more skeptical (Voort et al. 2019). But as we’ve seen RS also tends to reproduce human bias and should be investigated accordingly. A closer look at the information disclosed by Twitter on their recommender system does not provide a single element about the strategies adopted by Twitter to mitigate recommendation bias. Reviewing and investigating recommendation biases will become as important for social scientist as quantifying human biases. This paper focuses on the ideological bias of Twitter's recommender system and proposes an innovative experimental research design. In doing so, it answers the following two research questions: to what extent do recommender systems respond to ideological preferences by recommending ideological homogeneous content? When responding to ideological preferences, does the Twitter recommender system rely on explicit as well as implicit cues? The paper makes three different contributions to the literature. First, the design of this study allows to isolate recommendation bias from confounding human behaviours. It is particularly well suited for Twitter but can be adapted to other platforms. Second, we show that 'filter bubbles' are real and that after three weeks of ideologically-sided interaction, the proportion of recommended crosscutting content is reduced by 6%. Finally and as discussed in our conclusion, we want to highlight the nontransparent policy of social media platforms and call for the development of a dedicated framework allowing the independent auditing of recommender systems. The rest of this paper proceeds as follows. After discussing the relevance of investigating ideological bias, we propose a brief theoretical framework which first distinguishes algorithmic personalization from self-selective exposure and second explicit from implicit ideological cues. After these theoretical considerations, we present the research designs and the results of the experiment. We conclude by discussing the implication of these study, including the formulation of broad, but nevertheless important policy recommendations.

2 Theory

The theoretical risks of over-personalized or biased content on social media has already been broadly discussed. Most of what we know on the link between homogeneous content and political behavior has been gathered in studies investigating so-called ”echo chambers”. Even though, ”filter bubble” and ”echo chambers” are expected to affect the consumed content in a similar way, we believe the two concepts should be kept apart, because they result from two different mechanisms.

2.1 Echo Chambers

Scholars use the concept of ”echo chambers” to characterize situations in which individuals only receive ”echoes of their own voices” (Sunstein 2017). This can happen if a person only frequents groups of friends with similar interests or always consumes the same sources for news and information.

Researchers have, in the last years heavily investigated, how echo chambers influence the attitudes of individuals and their political behavior. Studies conducted by (Tewksbury 2003, R Kelly Garrett 2009, Beam 2014) confirm the assumption that users tend to access content that underpins their political attitudes. In a study that measures the effects of exposure to congruent partisan content in the US and Israel, evidence suggests that exposure to supportive information increases affective polarization (R. Kelly Garrett et al. 2014) and polarization via homogeneous personal networks (Druckman, Levendusky, and McLain 2018). Others have found that the selective use of congruent media does not necessarily lead to the active avoidance of incongruous content (R Kelly Garrett, Carnahan, and Lynch 2013).

A review by Zuiderveen et al. (2016) investigates the current empirical evidence of echo chamber effects on political opinions (Zuiderveen Borgesius et al. 2016). While it appears that there are measurable and statistically significant effects of echo chambers on political attitudes of individuals, they remain rather low or moderate. The authors conclude that there is no empirical evidence that would warrant great concern about echo chambers. One of the more recent systematic assessments of the filter bubble hypothesis comes from Möller et al. (2018), who test content diversity among news stories selected by human editors and multiple set-ups of automatized recommendation systems (Möller et al. 2018). The authors conclude that both human and algorithmic methods of prioritization provided a good diversity of opinion. There is also debate in the literature on whether echo chambers are a problem at all. For example, Guess et al. 2018 describe the somewhat limited evidence of echo chambers online and even talk about an ”echo chamber about echo chambers” where people lament the supposed problematic nature of echo chambers without considering systematic evidence in the academic literature (Guess, Nyhan, and Reifler 2018).

2.2 Filter bubbles

Often conflated with the concept of ”echo chamber”, filter bubbles actually describe a different mechanism. Echo chambers explains the emergence of homogeneous political environments by individual’s tendency to expose themselves to content, reinforcing, or confirming their views and avoiding sources that challenge them (Beam 2014, Frimer, Skitka, and Motyl 2017). Filter bubbles, on the contrary, are not caused by self-selective exposure but are the result of algorithmic personalization. If recommending algorithms rank content, so that it pleases the user, this also reduces the likelihood of crosscutting content appearing in the timeline.

In the end, echo chambers and filter bubbles produce very similar outcomes: they control what kind of content a specific user is exposed to. The main difference refers to the authority in charge of selection. While in an echo chamber, users are mainly in charge of selecting the content they will be exposed to, in the case of filter bubbles, an individual does not directly control which content is selected and how the selection happens. In fact, as algorithms are the default on most prominent social media pages, users could be totally unaware of algorithmic selection.

One mechanism does not preclude the other. A user can pre-select an ideologically diverse pool of content producer, but still end up in a homogeneous space because of algorithmic personalization. The assumption that these two mechanisms act independently from each other is the foremost motivation of this study.

We know painfully little on the systems that drive user experiences on social media. We can only assume that they rely on cues provided by the user behavior to deduce the content that will maximise the pleasure of the user, his engagement on the platform or the advertisement revenue associated with his behavior. When navigating an online platform, individuals adopt behavior that are typical of specific ideological position. They follow and interact in a gentle manner with other individuals that share their views. Conversely, they are likely to adopt a harsh tone with individual holding counter-attitudinal views. Any type of behavior, be it clicking, waiting, opening a link, rewatching a video for a second time can inform on the ideology of the individual. Taken all together, the complete set of behavior we adopt online informs about our own ideology. Consequently, each action undertaken on a social media platform feeds the algorithm with information about its own preferences, which itself enables the algorithm to better personalize the content proposed to the user. This leads us to our main expectation: the more ideological cues are provided to the recommender system, the bigger will be the recommendation bias.

2.3 Implicit vs. Explicit Cues

Now, because of the technical limitations (structural complexity, computational costs of training and deployment) surrounding the development of recommender systems, each type of behavioral cues is unlikely to be treated equally by the algorithm. For instance, it is very easier to directly asks users whether they prefer sport or politics content, than deducing it from their behavior. Why explicitly asking users for their preferences has the advantage of providing highly valid information, it is obvious that a platform cannot constantly ask its users for their preferences. The scarcity of explicitly stated preferences reduces the quality of personalization that can be performed. Social media platforms are therefore incentivized to base their recommender systems not only on explicitly stated preferences but also on implicit preferences, unknowingly conveyed through user behavior on the platform. Implicit cues include among others click, mouse movement, pause, historic data transmitted by third-party software. Although these kinds of cues are much noisier than explicit preferences - a click can be a mistake, a pause in scrolling down a feed can be random, etc. -, although the assumption by social media algorithms is that they come in sufficient numbers to clean out the noise and extract precise signals on the preferences of a user. In other words, knowing that one user visited the website of Fox News in the past does not allow to accurately infer her ideological position; yet, knowing that a user visited Fox News 150 times and MSNBC only 4 times in a month, allows to interpolate a rather conservative ideology. This explains why social media platform have good reasons to personalise not only on the basis of explicitly stated preferences, but also with implicit cues.

Beyond increasing the personalization of content, using implicit in addition to explicit cues could skew users' perception by reducing the awareness of encountering content personalised to their preference. We understand here that the types of cues used by algorithms to recommend content do not only have technical implications, they might affect users' behavior and thus have normative implications. Studies on awareness of personalization are hard to come by. However, digital journalism researcher Elia Powers collected a small sample size of 147 college students, which revealed that they were mostly unaware of the existence and workings of personalization techniques on platforms they use (Powers 2017). Other (small scale) studies find similar results when individuals are asked how their (social media) newsfeeds are organized (Eslami et al. 2015, Rader and Gray 2015). Should this unawareness be more widespread in society, the normative implications of algorithmic selection are even more significant since the user might think that they get a representative view of what is produced on a platform and, by extension, the public forum. At the same time, under the hood, the algorithm shows them a biased and inaccurate representation of the world.Although we do not expect implicit cues to be more relevant than explicit cues, these normative considerations brings us to investigate independently the recommendation bias implied by implicit and explicit cues.

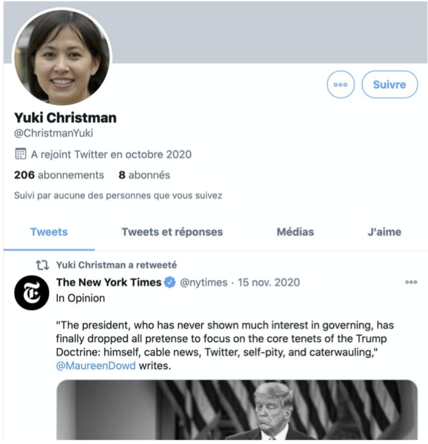
3 Methods

In order to explore the amount of bias as well as the respective roles played by implicit and explicit cues, we design an experiment, where we use automated Twitter agents to treat Twitter's algorithm with implicit and explicit cues. We chose to focus on Twitter, which is one of the leading social-platform for micro-blogging (TwitterGov 2016, Twitter 2019a). Additionally, Twitter's data API made it reasonably easy to access the full tweet history of the content producer encountered in our sample, which eased the synthetic reconstruction of chronological timelines as described later. Lastly, the existence of short keys to navigate Twitter facilitated the automation of agents.

3.1 Infrastructure

To investigate Twitter's algorithm, we generated 170 Twitter accounts [2] that were randomly allocated to one of the eight treatment groups listed below. Each agent consisted of a random name, a random avatar image [3], and a real cell phone number. To prevent a geographical spill-over between agents, we also allocate a VPN server to each agent to ensure that two separate agents do not have the same IP when connecting to Twitter. Figure 1 presents a screenshot of an agents, accompanied by the metadata that we randomly assigned. [4] During all the experiment and up to six times a day, each agent will (1) log in, (2) scroll down for about 50 tweets and, if necessary, (3) follow the twitter account assigned to it.

Figure 1: Example Account and Meta Data



Yuki Christman
@ChristmanYuki
A rejoint Twitter en octobre 2020
206 abonnements 8 abonnés
Suivi par aucune des personnes que vous suivez

Yuki Christman a retweeté
The New York Times @nytimes · 15 nov. 2020
In Opinion

"The president, who has never shown much interest in governing, has finally dropped all pretense to focus on the core tenets of the Trump Doctrine: himself, cable news, Twitter, self-pity, and caterwauling," @MaureenDowd writes.

```
[
  {
    "user": "christman34",
    "firstname": "Yuki",
    "middlename": "",
    "lastname": "Christman",
    "password": "...",
    "sim_id": 198,
    "phone": "+447459842494",
    "mail": "yuki.christman.2503",
    "liberal": true,
    "explicit": true,
    "implicit": true,
    "rpi_index": 2,
    "tz": "America/Los_Angeles",
    "country": "United States",
    "city": "San Jose",
    "region": "California",
    "vpn": "us5304.nordvpn.com.tcp443.ovpn",
    "host": "192.168.1.15",
    "profile": "ChristmanYuki"
  }
]
```

The experiment is separated into two stages, which took three weeks each. The

[2] Although pre-registration mentioned 300 agents, we only managed to create just over 200 agents. Several technical reasons caused the attrition of 30 agents: banned by Twitter, loss password, malfunctioning sim card, etc...

[3] Picture were randomly drawn from the website 'thispersondoesnotexist.com'

[4] To prevent Twitter from linking the agents together, we also randomized connection time and the user agents of the browsers used to connect to Twitter.

first lasted from 13 October to 3 November 2020 and was only used to administer treatments (treatment phase). The second phase began on 3 November and finished on 24 November 2020 (measurement phase). During this period, the treatments stopped (no additional following and no interaction) and we simply collected the recommended timeline without any further manipulation. The results presented in the next section corresponds to the data collected in this second phase. We also uniquely focused on the American political context, which was thriving the political content during the experiment because of the 2020 presidential election, which took place at the beginning of the second phase. The bipartisanship of American politics made the measure of ideology as well as the assignment of ideological treatments more manageable. Since American politics is organized around one axis, we assign one ideological side to each agent and deem all content from the other side to be counter-attitudinal. Thus, immediately after being created, each agent was first given an ideological position: 'Liberal' or 'Conservative.'

To administer implicit and explicit cues, we took advantage of the structure of Twitter. Twitter main feed consist of the content produced, shared or liked by a pool of account exclusively selected by the user (followed accounts). As a consequence, upon landing on Twitter, a new user has to pick accounts, she wants to 'follow'. Eventually, the content generated by the chosen accounts will end up on the timeline of these new users. The action of following an account is seen as an explicit cue, because it actively and explicitly selects a content producer. Agents assigned to the explicit treatment would provide explicit cues about their ideology by following 80% of congruent accounts and 20% of cross-cutting account. So for example, a conservative agent would follow 80% conservative and 20% liberal accounts, and vice versa for liberal agents.

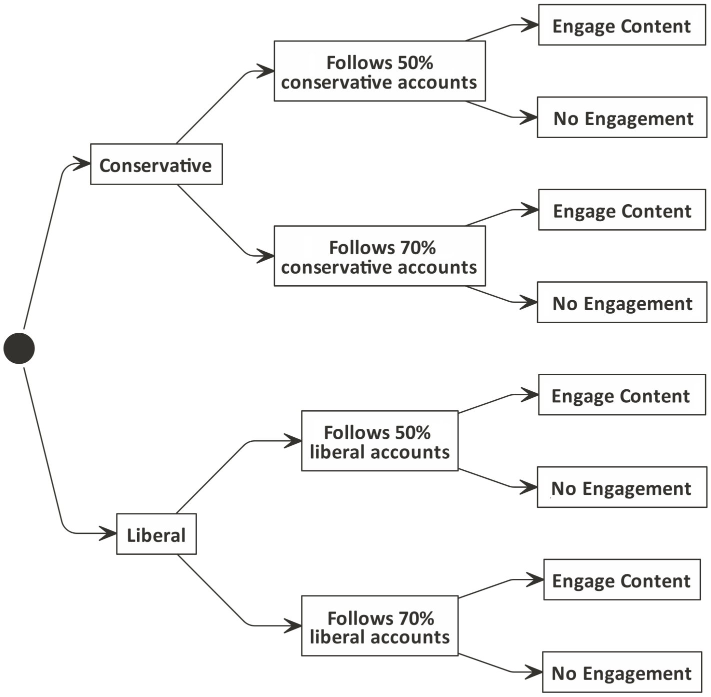
Table 1: Descriptive Statistics

condition	ideology	accounts	sessions	tweets
Control	Liberal	19	871	92990
Control	Conservative	22	906	96700
Only Explicit	Liberal	19	1065	113244
Only Explicit	Conservative	25	1011	107854
Only Implicit	Liberal	18	800	85668
Only Implicit	Conservative	20	707	75458
Explicit and Implicit	Liberal	18	857	91364
Explicit and Implicit	Conservative	29	1161	124064

These followed accounts were randomly drawn from a two pre-defined pools consisting of about 600 liberal and 600 conservative accounts (See Appendix for more info). Agents, who were not assigned to the explicit treatments could not provide any explicit cues about their ideology. Now, because Twitter requires you to follow accounts to have anything appear on your timeline these accounts need to follow some accounts to get started. Consequently, since an agent that follows no accounts cannot receive content, the control group for the explicit treatment consists of neutral explicit cues (50% liberals and 50% conservative).

Implicit cues consist of behaviors informing about a users' preferences, even if they are not meant to specifically select content. On Twitter, this can include scrolling rhythm (pausing longer on congruent content or scrolling faster on counter-attitudinal content), clicking on links, like, retweeting, etc. We chose to focus on two implicit behaviors: liking and retweeting. These two behaviors are essential to the Twitter community and were also convenient to implement technically. Consequently, agents assigned to the implicit treatment had a 30% probability to retweet or like congruent content. Agents who were not assigned to the implicit treatment simply scrolled down without interacting at all with content, thus providing no implicit cues.

Figure 2: Experimental Design

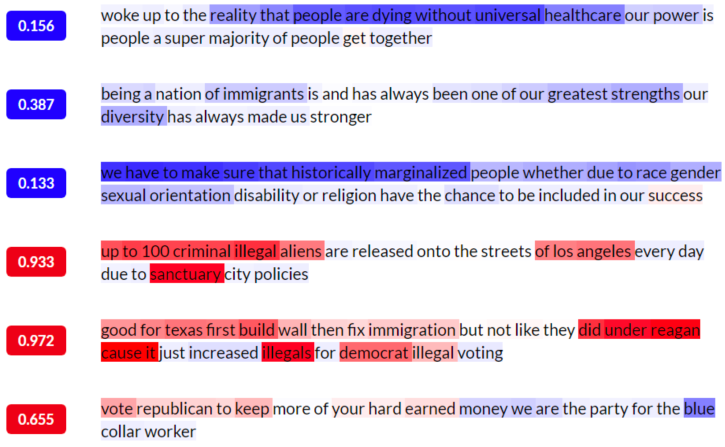


In summary, we adopt a factorial experimental design with three branches, as presented in Fig. 2. The first treatment is the ideological position of the agent (Liberal vs. Conservative) ; the second treatment is in regards to the administration of explicit cues (following 80% congruent accounts vs. following 50% congruent accounts) and the last treatment regards implicit cues (liking and retweeting congruent content vs. no interaction at all).

3.2 Measuring Ideology

Both the implementation of this design, as well as the analysis of the results require an automated and robust strategy to measure the ideology of the content. Particularly challenging is our need to measure ideology at the tweet level to administer the implicit treatment. Indeed, while an agent is scrolling, we need to instantly know whether a viewed tweet is congruent or counter-attitudinal. This excludes the involving of hand-coding as well as account-level procedures as proposed by Barberá 2015 and others. Following Gottlieb 2018, we used supervised machine learning and train a classifier to predict the ideology of a tweet given its text. The classifier was trained on all available tweets produced by acting members of congress. Each tweet was labeled 'Liberal' or 'Conservative' following the party affiliation of its author. The model combines convolutional layers with a LSTM layers, which have been proven to deliver excellent predicting performance in the context of text-classification (Goldberg 2016). The output layer relies on a sigmoid activation function, that fits the binary structure of the target.

Figure 3: Qualitatively sampled Tweets with predicted Ideology Score



0.156 woke up to the reality that people are dying without universal healthcare our power is people a super majority of people get together

0.387 being a nation of immigrants is and has always been one of our greatest strengths our diversity has always made us stronger

0.133 we have to make sure that historically marginalized people whether due to race gender sexual orientation disability or religion have the chance to be included in our success

0.933 up to 100 criminal illegal aliens are released onto the streets of los angeles every day due to sanctuary city policies

0.972 good for texas first build wall then fix immigration but not like they did under reagan cause it just increased illegals for democrat illegal voting

0.655 vote republican to keep more of your hard earned money we are the party for the blue collar worker

To evaluate the model, we initially kept the tweets of 20% of the congress members aside as hold-out dataset. We use these 60.000 tweets, which are not seen by the model while training, to evaluate the performance of the models (as can be seen in Fig. 5). After tuning a basic set of hyperparameters, we reached an out-of-sample accuracy of 86% on content level. To ensure the calibration of the model on the tweets seen by the agents during the experiment, we draw a sample of 1000 random tweets and ad hock coded them. Our hand-coding agreed 74.2% of the time with the classifier.

3.3 Recommendation Bias

The fact that the twitter timeline, even when recommendation is activated, only entails content produced by accounts selected by the user proves really helpful to distinguish echo chambers from filter bubbles. Because Twitter recommendation system only reorganizes content pre-selected by the user, we can compare the timeline as structured by the algorithm and its raw and chronological counterpart, to analyse exactly the influence of the recommender system. Comparing the algorithmic timeline with the chronological ones allows to control for any self-selective exposure and volume effect. If we observe that the timeline of a liberal user is overwhelmingly liberal, it is tempting to conclude that the algorithm priorizes liberal content according to this users' preferences. But this result could be independent from the algorithmic recommendation if the user follows more liberal account or simply if liberal accounts produce more content. When comparing the content as recommended with the content as it would appear chronologically, we can control for both self-selective exposure and volume effect, hence isolating the mere influence of the recommendation on the timeline. In other words, we approach the chronological timeline as a counter-factual timeline free from any algorithmic recommendation. The difference between this counter-factual chronological timeline and the algorithmic and observed timeline is interpreted as recommendation bias and constitutes the main outcome of this experiment.

Capturing recommendation bias consequently requires to observe both the algorithmic timeline and the chronological timeline. Algorithmic timeline are constituted by the first 50 tweets appearing on an agents' timeline after log-in. Chronological timelines were artificially reconstructed. During the experiment, we gather all the tweets produced and shared by the accounts in the followed pool as it is happening. Using the information on (1) which account were followed by an agent and (2) when the agent logged-in to gather the algorithmic timeline, we reconstructed what this agent would have seen, if its timeline was structured chronologically. To make it comparable to the algorithmic timelines, we restricted the chronological timelines to 50 tweets. Using this technique, we were able to match each algorithmic timeline with its chronological counterpart.

Once we had matched each collected algorithmic timeline with its counterpart, we compare the proportion of cross-cutting content entailed in the two timelines. Using the above-mentioned classifier, we label each observed tweet as liberal or conservative. In doing so, we compute for each agent the proportion of counter-attitudinal (cross-cutting) tweets observed in the algorithmic and chronological timelines. Recommending bias is consequently estimated through the difference between these two proportions. For instance, if the chronological timeline of a liberal agent entailed 40% cross-cutting - conservative in this case - content and its algorithmic timeline contained 30% cross-cutting tweets, we estimate an recommending bias of 10%. This means that for this agent, algorithmic timelines entailed 10% less cross-cutting tweets than in its chronological timeline.

Following our factorial design with three levels, we measure the causal effect of implicit and explicit cues on the recommending bias, captured as the difference in proportion of crosscutting content in the chronological and the algorithmic timelines. Treatment assignment was fully randomized, which allows us to simply estimate Average Treatment Effects (ATEs). Each treatment branch is compared to the control group within a linear regression (OLS).

4 Results

We are mainly interested in the average treatment effect, respectively the average recommender bias per condition. First we averaged the predictions per timeline, session and bot, as the latter is the unit of analysis. A simple linear model helps in describing average effects of which we conducted four different ones with increasing complexity as can be seen by the regression table 2 in the appendix.

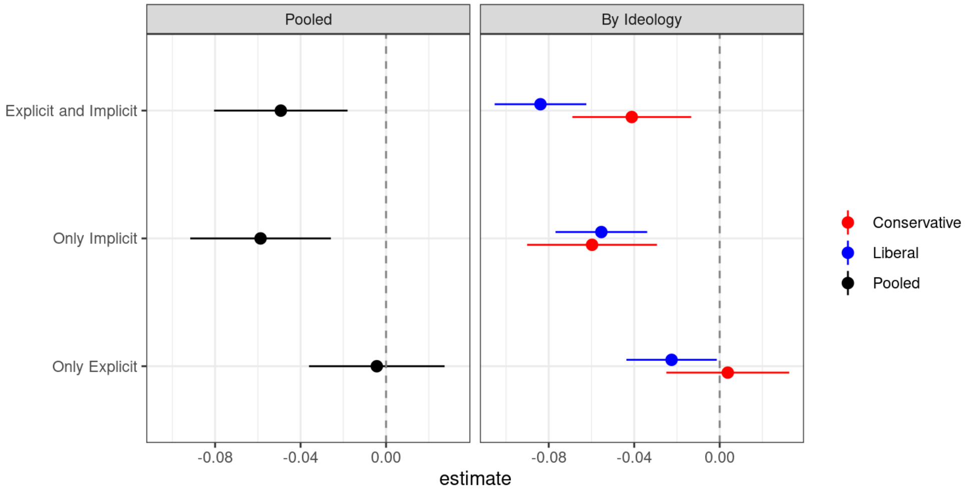
In order to get a more clear view on the causal effects we created a linear coefficient plot in Fig. 9. The left hand side shows the pooled estimates over conservatives and liberals for our treatment conditions. Compared to the chronological timeline and the control group we can not observe any significant effect regarding the explicit cues only branch. That means without additional preferences, the twitter algorithm will serve on average the proportion of ideological content that a user self selected into. Neither congruent nor divergent ideological content is prioritized more.

In contrast, we can observe a strong negative effect for implicit cues. Only providing behavioral preferences by liking or retweeting the algorithm is reducing cross cutting content (political opposite views) by about 6% on average. That means interacting with ideologically loaded content amplifies the algorithmic personalization bias compared to our control group (No Interaction).

If both effects are combined, the effect is comparable to the only implicit cues condition. This allows the inference that most of the personalization bias on Twitter is driven by implicit human behavior.

The best in sample model fit can be reached by simply combining our engagement treatment conditions with the ideological treatment. Additionally, adding an interaction term (multiplication) between our two treatment levels does not further increase model fit. The observed treatment effects remain relatively stable over all model specifications. Through additional bootstrapping we can show that the result of the experiment is significantly different from 0 and entails out of sample predictive (explanatory) power even under worst case scenarios.

Figure 4: ATE for Crosscutting Content by Condition



The right hand side of figure 9 shows the conditional treatment effects (CATE) for Liberals and Conservatives. The overall pattern remains stable but with minor deviations for Conservatives. Whether the difference between Liberals and Conservatives is due to altered political context over time, Twitter volume effects or measurement bias cannot be stated for certain.

5 Conclusion

In this study, we set up an experimental study to estimate the ideological bias of Twitter's recommender system. We find out, that following users from one ideological side does not trigger any ideological bias of the recommender system. A non-interacting user will have recommendation representative of her chronological timeline. On the other hand, we find strong evidence that interacting with tweets (liking, retweeting) triggers large recommendation bias. After three weeks of interaction with congruent content, recommended timelines entailed between 4 and 9% less crosscutting content than their chronological counterpart. We observe rather similar patterns for liberal and conservative agents. This findings have three different implications.

First, algorithmic recommendation amplification is a mechanism independent from self-selective exposure, that can explain ideologically homogeneous diets on social media. Because algorithmic recommendation affects the appearing content and not the produced content, it is crucial to investigate online ideological diversity at the level of the consumed content and not at the level of the producers, as some studies have done so far.

Second, we need to better understand how individuals react to personalized recommendation and especially whether the awareness of the recommendation plays a role in the way content is processed. As shown in this study, recommendation on Twitter - based on collaboartive filtering - mostly react to implicit cues, that are not explicitly and awarely meant to feed a recommender system. In this setting, recommendation and personalization can happen without anybody noticing and eventually induce misleading interpretation of reality. For instance, if a conservative person only sees conservative news without knowing that it has been personalized, she might be tempted by interpreting that this conservative opinion is representative of the overall pool of accounts she is following, including the liberal ones. This dissonance between the reality as perceived online and the reality as described by other actors - survey companies, media, politician, etc. - might increase mistrust or polarization. It is consequently very important not only to study people's awareness of recommendation but also the ways how online dissonance affect political behaviors.

Finally, the findings of this study demonstrate a systematic bias of Twitter's recommending system and, thus, highlight Twitter's failure to either detect the bias or mitigate it appropriately. In this context, we need to review, how recommending system are deployed on social media. Third parties actors such as academics or NGO-employees can play a crucial role in identifying the biases. We should follow a transparency model, where social media companies publish detailed reports about the information used to recommend, about the types of models used and about the search/mitigation of systematic biases. As transparency standard in the matter are increased, it will become easier to run studies as this one. Nonetheless, it would be very beneficial to both social media users and social media companies, to co-develop a framework allowing external testing, able to detect biases as soon as they appear. It was a herculean task to set up the study conducted in this paper and it is of public interest to make the auditing of social media more accessible. Put briefly, an external auditing framework will ensure that the negative public consequences of algorithmic bias are not underestimated in favor of short-term profits. Right now, nothing ensures that this is the case beyond the word of social media companies.

References

Ashish Bansal (2019). "Challenges in Recommender Systems at Twitter Scale". In: URL: http://tiny.cc/ncmbtz.

Barberá, Pablo (2015). "Birds of the same feather tweet together: Bayesian ideal point estimation using Twitter data". In: *Political analysis* 23.1, pp. 76–91.

Beam, Michael A (2014). "Automating the news: How personalized news recommender system design choices impact news reception". In: *Communication Research* 41.8, pp. 1019–1041.

Druckman, James N, Matthew S Levendusky, and Audrey McLain (2018). "No need to watch: How the effects of partisan media can spread via interpersonal discussions". In: *American Journal of Political Science* 62.1, pp. 99–112.

Eslami, Motahhare et al. (2015). "" I always assumed that I wasn't really that close to [her]" Reasoning about Invisible Algorithms in News Feeds". In: *Proceedings of the 33rd annual ACM conference on human factors in computing systems*, pp. 153–162.

Frimer, Jeremy A, Linda J Skitka, and Matt Motyl (2017). "Liberals and conservatives are similarly motivated to avoid exposure to one another's opinions". In: *Journal of Experimental Social Psychology* 72, pp. 1–12.

Garrett, R Kelly (2009). "Echo chambers online?: Politically motivated selective exposure among Internet news users". In: *Journal of Computer-Mediated Communication* 14.2, pp. 265–285.

Garrett, R Kelly, Dustin Carnahan, and Emily K Lynch (2013). "A turn toward avoidance? Selective exposure to online political information, 2004–2008". In: *Political Behavior* 35.1, pp. 113–134.

Garrett, R. Kelly et al. (2014). "Implications of Pro- and Counterattitudinal Information Exposure for Affective Polarization". In: *Human Communication Research* 40.3, pp. 309–332. DOI: 10.1111/hcre.12028. URL: http://tiny.cc/ocmbtz.

Goldberg, Yoav (2016). "A primer on neural network models for natural language processing". In: *Journal of Artificial Intelligence Research* 57, pp. 345–420.

Gottlieb, Alex (2018). "Private Partisan, Public Moderate: Preference Falsification on Twitter". In: *Princeton University DataSpace*. URL: https://github.com/alex-gottlieb/deepIdeology.

Guess, Andrew, Brendan Nyhan, and Jason Reifler (2018). "Selective exposure to misinformation: Evidence from the consumption of fake news during the 2016 US presidential campaign". In: *European Research Council* 9.

Möller, Judith et al. (2018). "Do not blame it on the algorithm: An empirical assessment of multiple recommender systems and their impact on content diversity". In: *Information, Communication & Society* 21.7, pp. 959–977.

Powers, Elia (2017). "My News Feed is Filtered?" In: *Digital Journalism* 5.10, pp. 1315–1335. DOI: 10.1080/21670811.2017.1286943.

Rader, Emilee and Rebecca Gray (2015). "Understanding user beliefs about algorithmic curation in the Facebook news feed". In: *Proceedings of the 33rd annual ACM conference on human factors in computing systems*, pp. 173–182.

Statista (2019). "Daily social media usage worldwide 2012-2019". In: URL: http://tiny.cc/ecmbtz.

Sunstein, Cass R (2017). *# Republic: Divided democracy in the age of social media*. Princeton University Press.

Tewksbury, David (2003). "What do Americans really want to know? Tracking the behavior of news readers on the Internet". In: *Journal of communication* 53.4, pp. 694–710.

Twitter (2017). "Using Deep Learning at Scale in Twitter's Timelines". In: URL: http://tiny.cc/mcmbtz.

— (2019a). "About your Twitter timeline". In: *Twitter Help Center*. URL: http://tiny.cc/fcmbtz.

— (2019b). "Improving engagement on digital ads with delayed feedback". In: URL: http://tiny.cc/lcmbtz.

— (2020a). "Transparency around image cropping and changes to come". In: URL: http://tiny.cc/kcmbtz.

— (2020b). "What Twitter learned from the Recsys 2020 Challenge". In: URL: http://tiny.cc/jcmbtz.

TwitterGov (2016). "Tuesday was the most-Tweeted election day ever: 75+ million global election-related Tweets sent through 3am ET as Trump claimed victory". In: URL: http://tiny.cc/hcmbtz.

Voort, H.G. van der et al. (Jan. 2019). "Rationality and politics of algorithms. Will the promise of big data survive the dynamics of public decision making?" en. In: *Government Information Quarterly* 36.1, pp. 27–38. ISSN: 0740624X. DOI: 10.1016/j.giq.2018.10.011. (Visited on 01/31/2021).

Zuiderveen Borgesius, F et al. (2016). "Should we worry about filter bubbles?" In: *Internet Policy Review* 5.1.

Appendix

Figure 5: Ideology Classifier Performance

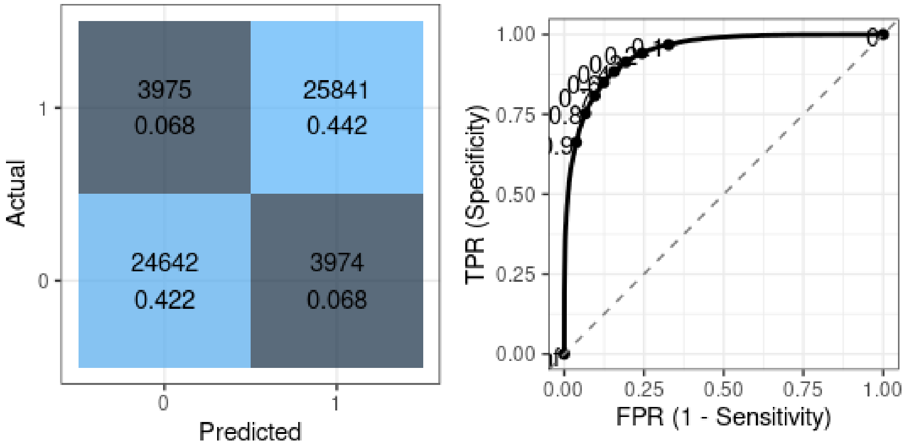


Table 2: ATE Linear Model (Pooled)

	Dependent variable:		
	Percent Crosscutting Content		
	(1)	(2)	(3)
Only Explicit	-0.004		-0.008
	(0.016)		(0.010)
Only Implicit	-0.059^{***}		-0.057^{***}
	(0.017)		(0.010)
Explicit and Implicit	-0.049^{***}		-0.059^{***}
	(0.016)		(0.009)
Conservative		0.118^{***}	0.120^{***}
		(0.008)	(0.007)
Constant	-0.001	-0.095^{***}	-0.065^{***}
	(0.012)	(0.006)	(0.008)
Observations	170	170	170
R^2	0.110	0.566	0.687
Adjusted R^2	0.094	0.563	0.679
Note:	*p<0.1; **p<0.05; ***p<0.01		

Table 3: CATE Linear Model (Liberal vs Conservative)

	Dependent variable:	
	Percent Crosscutting Content	
	(Liberal)	(Conservative)
Only Explicit	-0.023^{**}	0.004
	(0.011)	(0.015)
Only Implicit	-0.055^{***}	-0.060^{***}
	(0.011)	(0.016)
Explicit and Implicit	-0.084^{***}	-0.041^{***}
	(0.011)	(0.014)
Constant	-0.056^{***}	0.047^{***}
	(0.008)	(0.011)
Observations	74	96
R^2	0.492	0.222
Adjusted R^2	0.470	0.196
Note:	$^{*}p<0.1$; $^{**}p<0.05$; $^{***}p<0.01$	

Figure 6: Randomized Inference: Under the Null Hypotheses any system that is close to a distribution of random x and y is ”useless”. Generating random data from the existing once allows to have the similar data properties but without robust relationship in the data. In the end we can compare our model in red with the distributions of 10000 random systems that realize around zero and show that the data behind our models is far away from a random observation.

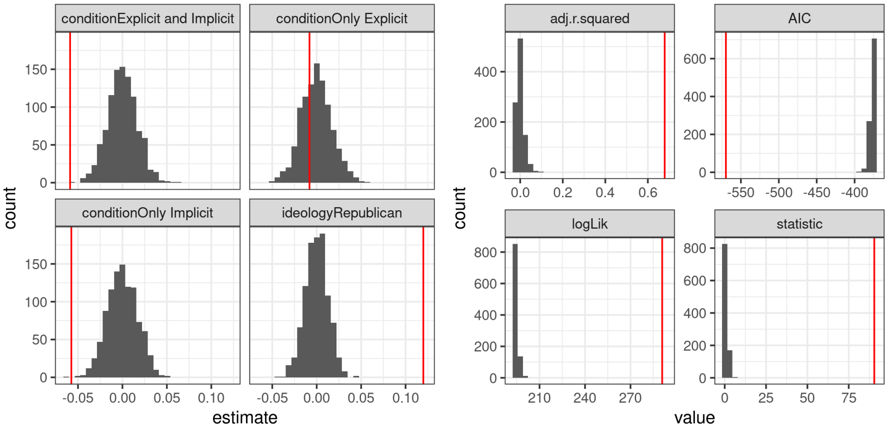


Figure 7: Leave-one-out Validation: measures the impact of single observation on the stability of estimated parameters. Although, we can observe little variation in the parameter estimates, a single bot is unlikely to bias the result substantially given a simulation of 10000 trials.

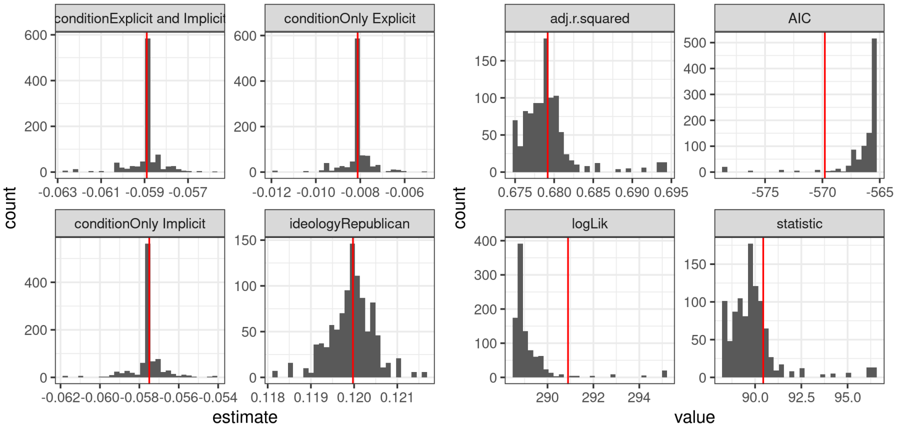


Figure 8: Bootstrap with Replacement: is intended for estimating confidence intervals for complex statistical models or arbitrary functions whose variance properties are difficult to analytically derive. The provided boundaries from 10000 trials match the standard errors from a linear model. In addition we can check whether a underlying data sample is only one lucky realization of the unknown population, which seems not to be the case.

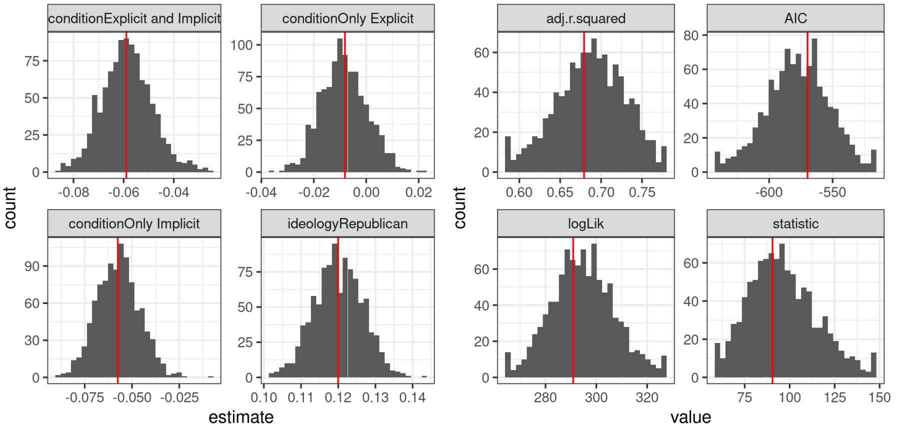
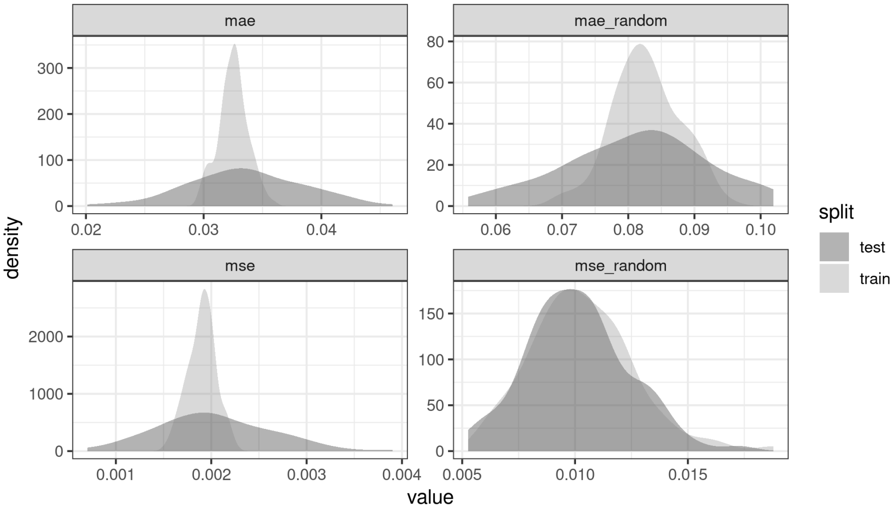


Figure 9: Predictive Validation: Splitting the bot sample into two separate folds before fitting allows to evaluate a model on completely different data. This helps to reduce overfly optimistic metrics and yield a better estimate of a models out of sample performance. We can see that the mean squared or absolute error are much lower than the random data models and by the factor 3-5, which is another excellent performance aspect beside the relative high in-sample R^2.


mae
mae_random
mse
mse_random
density
value
split
test
train