PNAS RESEARCH ARTICLE | COMPUTER SCIENCES SOCIAL SCIENCES

Causally estimating the effect of YouTube's recommender system using counterfactual bots

Homa Hosseinmardi[a,b,1], Amir Ghasemian[c], Miguel Rivera-Lanas[d], Manoel Horta Ribeiro[e], Robert West[e], and Duncan J. Watts[a,b,f,1]

Edited by Christopher A. Bail, Duke University, Durham, NC; received August 8, 2023; accepted December 13, 2023, by Editorial Board Member Mark Granovetter

In recent years, critics of online platforms have raised concerns about the ability of recommendation algorithms to amplify problematic content, with potentially radicalizing consequences. However, attempts to evaluate the effect of recommenders have suffered from a lack of appropriate counterfactuals—what a user would have viewed in the absence of algorithmic recommendations—and hence cannot disentangle the effects of the algorithm from a user's intentions. Here we propose a method that we call "counterfactual bots" to causally estimate the role of algorithmic recommendations on the consumption of highly partisan content on YouTube. By comparing bots that replicate real users' consumption patterns with "counterfactual" bots that follow rule-based trajectories, we show that, on average, relying exclusively on the YouTube recommender results in less partisan consumption, where the effect is most pronounced for heavy partisan consumers. Following a similar method, we also show that if partisan consumers switch to moderate content, YouTube's sidebar recommender "forgets" their partisan preference within roughly 30 videos regardless of their prior history, while homepage recommendations shift more gradually toward moderate content. Overall, our findings indicate that, at least since the algorithm changes that YouTube implemented in 2019, individual consumption patterns mostly reflect individual preferences, where algorithmic recommendations play, if anything, a moderating role.

algorithmic audits | experiment design | recommender systems | online extremism

Significance

When problematic behaviors are observed on online platforms such as YouTube, it is generally unclear to what extent they reflect biases in the platform's algorithms versus user preferences. Effective interventions require answering such questions, but disentangling algorithmic influence from user intentions with non-experimental data is extremely difficult. Here we introduce an experimental method for causally estimating the effect of platform recommendations that explicitly compares real user behavior with "counterfactual" bots that first imitate users and then rely exclusively on recommendations. We find that recommendations, on average, push users to more moderate content, suggesting that user preferences play the dominant role in determining consumption. More broadly, our method has implications for studying situations in which user preferences and algorithms interact.

With over 250 million active users in the United States and over 2.6 billion worldwide, YouTube is among the world's largest and most engaging social media platforms. Moreover, while news and other related content account for a relatively small share of both production and consumption, the sheer scale of the platform means that YouTube is also one of the largest online sources of political information for Americans, roughly equivalent to X (formerly Twitter) (1–3). Finally, while on-platform news consumption is dominated by mainstream and moderate sources (4), a relatively small but still substantial population of YouTube users consume concerning amounts of ideologically extreme (5), conspiratorial (6), and inflammatory content (7). The ready availability of problematic content, along with the pervasive presence of algorithmically generated recommendations on the site, has led to prominent speculation that YouTube is actively radicalizing its users via its recommender system (8, 9). As has been pointed out (10–14), however, the content that users consume is some unobserved combination of their own preferences and the platform design, including the recommender, each of which influences the other in a complex feedback loop with potentially emergent properties. Careful empirical work is therefore needed to estimate the effect of platform design on user consumption in a way that accounts for user preferences.

To date, empirical studies using different methodological approaches have reached somewhat different conclusions regarding the relative importance of algorithmic recommendations. While no studies find support for the alarming claims of radicalization that characterized early, anecdotal accounts, audit studies in which bots (15) or humans (5) follow rule-based viewing patterns—and platform recommendations are systematically recorded—have found that blindly following the recommender system results in ideologically biased recommendations, implying that the recommender is at least partly responsible. In contrast, panel studies (4, 16) based on real user traces over many months show that the consumption of "radical" content on YouTube does not increase over time or with session length (on average) and is highly correlated with off-platform consumption, suggesting that user preferences are more to blame than biased recommendations (17).

Critically, neither type of study is sufficient to resolve the key causal question: How much bias do recommenders cause? By design, panel studies only observe what users actually clicked on, not what was recommended to them. As a result, they cannot rule out

Author contributions: H.H., A.G., M.R.-L., M.H.R., R.W., and D.J.W. designed research; H.H. and A.G. performed research; H.H. and A.G. analyzed data; M.H.R. designed bots; and H.H. and D.J.W. wrote the paper.

The authors declare no competing interest.

This article is a PNAS Direct Submission. C.A.B. is a guest editor invited by the Editorial Board.

[1]To whom correspondence may be addressed. Email: homahm@seas.upenn.edu or djwatts@seas.upenn.edu.

This article contains supporting information online at https://www.pnas.org/lookup/suppl/doi:10.1073/pnas.2313377121/-/DCSupplemental.

Published February 13, 2024.

that the platform is recommending more extreme content than is visible in the consumption patterns, nor can they reveal what a user would have watched in the absence of recommendations. Audit studies, meanwhile, also cannot estimate the causal effect of the recommender on biased consumption. Say, for example, that a hypothetical user who ignored all recommendations ended up consuming content that is at least as biased as an otherwise identical user who only clicked on recommended content. In that case, one would not conclude that the algorithm itself is biased even if the "algorithmic" user also consumed biased content: Only if the latter were more biased than the former would the recommender be responsible for the residual bias. Just as with panel studies, audit studies do not create counterfactual comparisons of this sort and hence cannot identify the cause of the observed bias. A second, related shortcoming of audit studies is that the causal (i.e., counterfactual) effect of the recommender likely depends on the type of user; specifically, how much moderate vs. extreme content they would have consumed even in the absence of recommendations. Here, audit studies struggle to find the right balance between capturing rare and highly unrepresentative users who are unlikely to show up in surveys (5) while also not assuming far higher concentrations of extreme content than is consumed by any real user (15).

In this paper, we propose an experimental approach, which we call "counterfactual bots," designed to causally estimate the effect of algorithmic recommendations independent of user intentions. The bots in question are logged-in, programmatic users, each trained on the exact historical trajectory of a real user, drawn from empirical panel data encompassing 15 mo (October 2021 to December 2022) of desktop browsing behavior by 87, 988 users (see *Materials and Methods* and *SI Appendix* for more details and a discussion of the benefits of using logged-in vs. logged-out users, as was done in ref. 15).

Each experiment proceeds in two phases. First, during an initial "learning" phase, all bots follow the same sequence of videos, ensuring that they present indistinguishable "preferences" to YouTube's recommender system. However, in a second "observation" phase, each bot is assigned to one of two types of treatment: the "user" treatment, in which the bot continues to follow the historical trajectory of the focal user; or a counterfactual treatment in which they follow some predefined rule such as clicking on the top-ranked sidebar (i.e., up next) video or imitating a different type of user, (see Fig. 1). Upon completion of each experiment, we use the YouTube API to retrieve metadata associated with each video ID in our collection, which we use to estimate the partisanship of the content (see *Methods and Materials* for details). By measuring the difference in the partisanship of watched and recommended videos between user and counterfactual treatments in the observation phase, our approach eliminates the preference or choice component of observed consumption, allowing us to estimate the causal effect of algorithmic recommendations. An additional advantage of our design is that by training our bots on historical user data, our results have high ecological validity, meaning that they describe the effects of recommendations on real users rather than hypothetical ones. Finally, leveraging a large, representative historical panel allows us to estimate the effect of the recommender for different types of users—in particular, users who consume the largest amounts of problematic content. As noted above, these users are rare and hence are unlikely to volunteer for online experiments or surveys; however, by oversampling the "tail" of the distribution, we can obtain accurate estimates even for rare cases (16, 18).

Our analysis yields four main findings. First, we find that algorithmic bots, on average, receive less partisan recommendations and consume less partisan content than the corresponding "real" users—a result that is stronger for heavier consumers of partisan content. Second, we find that real users who consume "bursts" of highly partisan videos subsequently consume more partisan content than identical bots who subsequently follow algorithmic viewing rules. Third, we find that when a user switches their diet from one dominated by far-right news content to one dominated by moderate news content, recommendations of far-right content essentially disappear from the sidebar within 30 videos, but linger for longer in homepage recommendations. Fourth, we show that longer histories of prior far-right consumption result in longer "forgetting" times of homepage recommendations but have no impact on the forgetting time of sidebar recommendations. Together, our results show that platform recommendations serve,

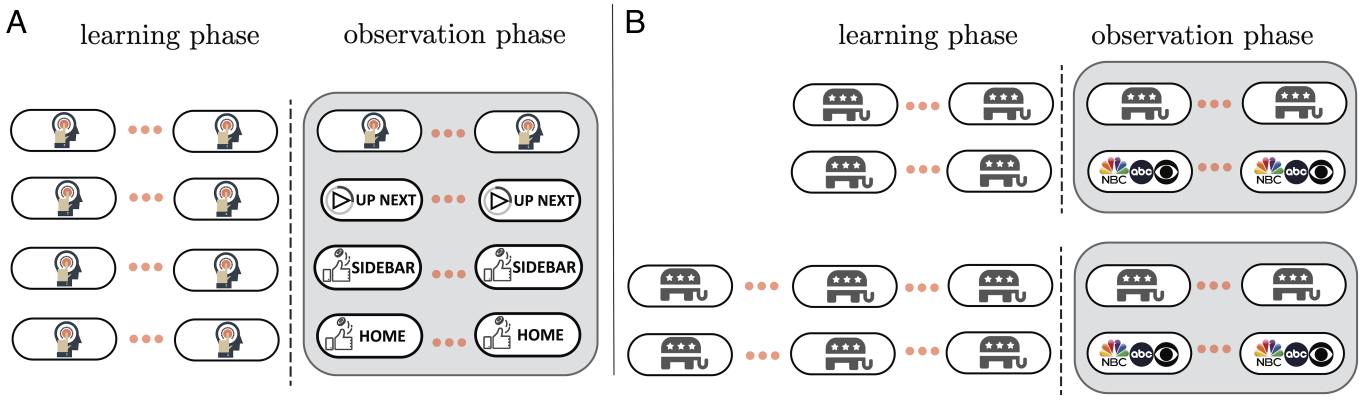

Fig. 1. Overview of the counterfactual bot method to disentangle YouTube's recommender system from user preferences utilizing counterfactual bots. Each panel shows the trajectories (one per row) that bots traverse within the corresponding experiment. By measuring the difference in the partisanship of watched videos by control bot ($y^{\text{cont.}}$) and watched videos by algorithmic counterfactual bots ($y^{\text{alg.}}$), our design eliminates the "preference" or "choice" component ($\widehat{y}^{\text{pref.}} = y^{\text{cont.}} - y^{\text{alg.}}$) of observed consumption, allowing us to estimate the causal effect of algorithmic recommendations. (*A*) Estimating bias of the recommender: Four bots watch the same history in the learning phase, whereas in the observation phase, the control bot continues to follow the real user's historical trajectory and the "counterfactual" bots follow simple algorithmic rules: "up next" (choosing the top-ranked video from the sidebar), "random sidebar" (choosing a random video from the sidebar), and "random home" (choosing a random video from the homepage). (*B*) Estimating "forgetting time" of the recommender: Two bots start at the same time, watching the same trajectory in the learning period. The control bot will continue watching from the same trajectory in the observation phase, while the counterfactual bot will switch to watching videos of moderate content. To estimate the effects of different-length histories, half the bots have "short" (30 video) histories prior to switching (top two rows), while the other half have "long" (120 video) histories (bottom two rows).

on average, to moderate a user's experience relative to following their own exogenous preferences, where the moderating effect is predominately experienced by extreme users. Noting that in 2019 YouTube made substantial changes to their recommendation algorithm that may have impacted the likelihood of recommending highly partisan content (16), our results suggest that at least in the post-2019 era, a user's preferences are the primary determinant of their experience.

Results

Our four findings derive from two main experiments, each of which leverages counterfactual bots in somewhat different ways. In the first main experiment, shown schematically in Fig. 1*A*, the bots simulate a user who switches from replicating the behavior of a real user (during the learning phase) to one who follows a simple heuristic (during the observation phase) such as clicking on the top-ranked (aka "up next") video on the right side of the screen. Leveraging this design, we extract our first two main results: one estimating the causal effect of the recommender for different types of users and one estimating the causal effect of users consuming bursts of far-right videos. In the second main experiment, shown schematically in Fig. 1*B*, the bots simulate a user "switching" from one set of preferences (dominated by far-right consumption) to another (moderate consumption) and measure the "forgetting time" of the recommender, defined as the number of post-switch videos before the recommendations become indistinguishable from those for a moderate user with no far-right history. As with the first main experiment, we leverage the design to extract two findings: one estimating the forgetting time for a user with a short (30 video) history of far-right consumption and one comparing the forgetting times of short and long (120 video) history consumers.

Both experiments leverage the same sample of 4,583 users who watched at least 140 YouTube videos during October 2021 to December 2022, drawn from a much larger ($N = 87{,}988$) US representative desktop panel (*Methods and Materials*). From this sample, we then further sampled trajectories with a length of exactly 120 videos from each of these users by choosing a random start point between 1 and $M_i - 120$, where M_i is the total number of video views for the ith user, and taking the next 120 videos. The number of sampled trajectories from each user is proportional to the user's lifetime in the panel, resulting in 24,871 unique user histories (*Methods and Materials*). We use channel labels provided by ref. 19 and assign all videos produced by a given channel the same partisan score. Next, we clustered these histories into eight news consumption "archetypes" ψ^X ranging from ψ^{fL}, characterized by mostly far-left with some centrist content, to ψ^{fR}, characterized by mostly far-right content (see *Methods and Materials*, and *SI Appendix*, Fig. S2 and Table S1 for details). Recognizing that within the ψ^{fR} archetype there remains considerable heterogeneity regarding the relative consumption of *fR* vs. other content as well as the total volume of *fR* videos, we further decompose ψ^{fR} into ψ^{fR}_{low}, $\psi^{fR}_{\text{medium}}$, and ψ^{fR}_{high} (see *SI Appendix*, Fig. S3 and Table S2 for details).

Estimating Bias of the Recommender. In this experiment, we sampled randomly 32 histories from ψ^C (characterized almost exclusively by centrist consumption) and ran a stratified sampling from ψ^{fR}, choosing 35 random histories from the ψ^{fR}_{low} group and taking all 41 and 17 histories from each of the $\psi^{fR}_{\text{medium}}$ and ψ^{fR}_{high} ones respectively, yielding a final sample of 125 "focal" users. We note that ψ^C accounted for roughly 66% of all histories in our sample, whereas ψ^{fR} accounted for only 1.12%; thus, our final sample over-represents heavy consumers of far-right content, who otherwise would not appear in sufficient numbers to power our analysis.

As noted above, the experiment comprised two phases. In the first half, the learning phase, four logged-in bots simultaneously and independently followed the trajectory of the focal user for the first $N_{\text{learning}} = 60$ videos of the focal user history. In this way, the recommender system had ample time to learn the preferences of each of the bots, but because all bots had the exact same history, they all presented the same preferences. In the second half, the "observational" phase, one of the bots (control bot) continued to watch videos from the trajectory of the same user for an additional $N_{\text{observation}} = 60$ videos, while the other three bots (counterfactual bots) switched to one of the following rule-based trajectories: up next, in which the bot deterministically selected the first video from the sidebar recommendations; random sidebar, in which the bot randomly selected one of the top 30 videos listed in the sidebar recommendations; and random homepage, in which the bot randomly selected a video from the top 15 videos listed in the homepage recommendations. For each of the selected focal users, we conducted three replications of this experiment, where each replication began with identical initial conditions but varied depending on the stochastic responses of YouTube's recommender system (i.e., if two hypothetical users created the exact same profile and watched the exact same sequence of videos, their recommendations would still not be identical). In total, our experiment comprised four bots per replication with an average of 2.61 replications per focal user for 125 focal users (*SI Appendix*, Table S3), yielding 1,304 independent trajectories of 120 videos each and an estimated cumulative watch time of over 640,975 min.

Fig. 2 shows four instances of the experiment for one focal user from each of the ψ^C, ψ^{fR}_{low}, $\psi^{fR}_{\text{medium}}$, and ψ^{fR}_{high} archetypes. As expected, the average partisanship of the videos consumed during the observation period increases with the partisanship of the focal user: whereas the bots replicating the ψ^C users generally consume videos that fluctuate around a partisan score

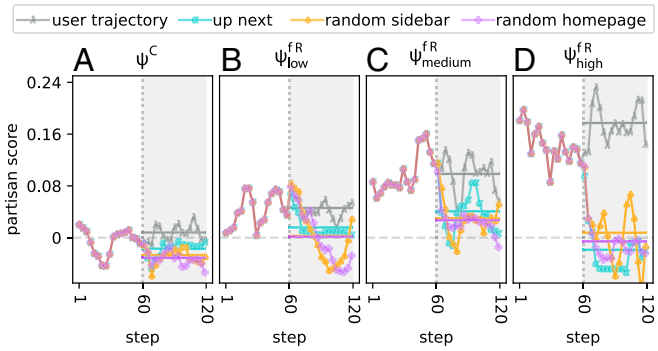

Fig. 2. Examples of traversed trajectories for four focal users with different mixtures of (*A*) center (ψ^C) and (*B–D*) far-right (ψ^{fR}) consumption in the counterfactual experiment, Fig. 1*A*. The first half is the learning phase (all four bots watch the same videos at each step) and the second half (shaded gray area) is the observation phase (each of the four bots follows a separate rule). The *y*-axis shows the partisanship of watched videos at each step. The dashed line shows zero partisanship. Solid lines show the average partisan score of all 60 watched videos in the observation phase for each path.

of 0 (Fig. 2*A*), the bots replicating the ψ^{fR}_{low}, $\psi^{fR}_{\text{medium}}$, and ψ^{fR}_{high} users consume progressively more partisan content (Fig. 2 *B*–*D*, respectively). Also as expected, the trajectories of all four bots are indistinguishable during the learning period, reflecting that they are all viewing the same sequence of videos. In the observation period, however, the bot trajectories diverge: whereas the control bot (gray line) continues on a similar path to the learning phase, the three counterfactual bots—up next (blue line), random sidebar (yellow line), and random homepage (purple line)—take somewhat different paths, both from the control and from each other. Notably, all three counterfactual bots trend toward less partisan content than the control, where the gap is small in the case of the ψ^{C} user (Fig. 2*A*) but becomes increasingly pronounced as the partisanship of the focal user increases. In the case of the ψ^{fR}_{high} user (Fig. 2*D*) the difference is highly pronounced and suggests that for extremely partisan users, the recommender actively promotes more moderate content than what the user would otherwise consume.

Fig. 3 shows these differences more systematically: each box-plot shows the median, interquartile range, and full range of the average partisanship for the watched videos by each of the four bots during the observation phase. Fig. 3*A* reveals that for ψ^{C} users, both counterfactual bots and control bots received relatively non-partisan recommendations, on average, and the differences between control and counterfactual bot experiences were small (see *SI Appendix*, Table S5 for more details on *P*-value and effect size for Fig. 3). Fig. 3 *B*–*D* shows that as the partisanship of the focal users increases, the gap between partisanship of control and counterfactual bots increases, suggesting that the net effect of the recommender was, if anything, to moderate the partisanship of the user experience. To quantify this qualitative observation we first compute user preference $\hat{y}^{\text{pref.}} = y^{\text{cont.}} - y^{\text{alg.}}$ as the gap between the partisanship of the control bot trajectory $y^{\text{cont.}}$ and the partisanship of a counterfactual bot $y^{\text{alg.}}$ (algorithmic or rule-based path); thus, a positive value of $\hat{y}^{\text{pref.}}$ corresponds to an intrinsic preference for partisan content relative to what the recommender system is recommending.

Next we regress $\hat{y}^{\text{pref.}}_{t} = \alpha + \beta_1 t + \beta_2 n^{\text{learning}}_{C} + \beta_2 n^{\text{learning}}_{R} + \beta_2 n^{\text{learning}}_{fR}$ on historical features of the learning phase, including: the step t at which the video was watched; and the number of center videos n^{learning}_{C}, the number of R videos n^{learning}_{R}, and the number of fR videos n^{learning}_{fR} in the learning phase. Table 1 shows that α is positive for all three types of recommendations (up next, random sidebar, and random homepage), confirming that recommendations have moderating effects relative to the focal users' intrinsic preferences. Furthermore, a higher number of C videos in the learning phase (n^{learning}_{C}) results in a smaller difference between control and counterfactual bots, while a higher number of fR videos (n^{learning}_{fR}) has the opposite effect. Table 1 also shows that β_1, the coefficient for the number of steps into the observed trajectory, is not significantly different from 0, consistent with the null hypothesis that trajectories do not become more or less extreme over time. Finally, α is larger for random homepage than up next, which in turn is larger than for random sidebar, suggesting that homepage recommendations are more moderate than sidebar but that the top-ranked sidebar recommendation is more moderate than the rest of the sidebar.

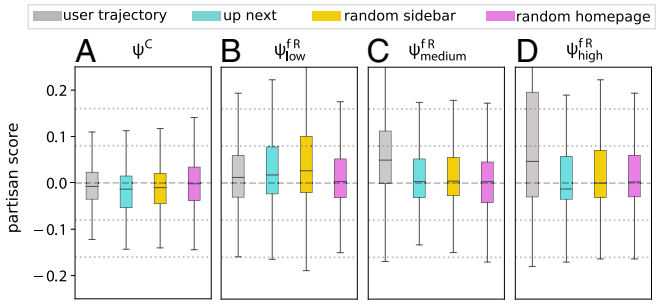

Fig. 3. Partisan score of the 60 watched videos for the control and counterfactual bots during the observation phase for focal users with different mixtures of (*A*) center (ψ^{C}) and (*B*–*D*) far-right (ψ^{fR}) video consumption. Each box-plot shows the median, interquartile range, and full range of the average partisanship (the *y*-axis range is limited to [−0.24,0.24] for better visualization). The dashed line shows zero partisanship, and the dotted lines represent one and two standard deviations away from the mean (zero) of partisan scores (see *SI Appendix* for details of the sample of videos and partisan scoring method).

Table 1. We define user preference as the difference in partisan score between the trajectory that the user has traversed (control bot) and the rule-based trajectory (counterfactual bot), which follows the recommendation only

	Up next	Random sidebar	Random homepage
Preference (α)	0.029*** [0.027, 0.032]	0.016*** [0.013, 0.019]	0.047*** [0.045, 0.050]
Depth (β_1)	0.000 [0.000, 0.000]	0.000 [0.000, 0.000]	0.000 [0.000, 0.000]
n^{learning}_{C}	−0.001*** [−0.002, −0.001]	0.001*** [0.001, 0.001]	−0.001*** [−0.001, −0.001]
n^{learning}_{R}	0.000 [0.000, 0.001]	0.001*** [0.000, 0.001]	0.000 [0.000, 0.000]
n^{learning}_{fR}	0.003*** [0.002, 0.003]	0.003*** [0.003, 0.004]	0.002*** [0.002, 0.003]
R^2	0.055	0.034	0.066

User Preference is positive for all three types of recommendations (up next, random sidebar, and random homepage). A higher number of C videos in the learning phase results in a smaller difference between control and counterfactual bots, while a higher number of fR videos has the opposite effect.
$^{+}P < 0.1$; $^{*}P < 0.05$; $^{**}P < 0.01$; $^{***}P < 0.001$

To examine the robustness and generalizability of these findings, we also conducted three supplemental analyses that we report in the (*SI Appendix*, section 4). First, we re-analyzed the data from our experiment replacing the partisan score with (a) an "establishment" score that captured the extent to which channel owners position themselves as non-partisan "anti-establishment" figures; and (b) a popularity score based on views, likes, and comments. For the establishment score, we found similar results to partisanship (*SI Appendix*, Fig. S4), whereas for popularity, we found no consistent effect of the recommender in either direction (*SI Appendix*, Fig. S5). Second, to check that our findings generalize to other parts of the ideological spectrum, we conducted an additional experiment for consumers of predominantly "far-left" partisan content, finding similar results to Fig. 3 (*SI Appendix*, Fig. S7). Third, to check the effect of channel subscriptions, we conducted another experiment in which the 17 fR-high focal users also subscribed to their three most visited channels, again finding very similar results to Fig. 3 (*SI Appendix*, Fig. S8). We thank two anonymous reviewers for suggesting these supplemental analyses.

Bursty viewership effect. Even if the recommender moderates a user's experience on average, it may be the case that it overreacts to bursts of partisan consumption, defined as viewership of highly partisan videos in near succession. Previous work (4) has found that bursts of this sort (for lengths 2, 3, and 4) predict subsequent higher consumption of partisan content but could not determine if the cause was endogenous user preferences or the exogenous response of the recommender. Here, we revisit this question by exploiting the presence of real users in our data who consumed bursts of up to six videos from one of {*C, R, fR*} categories during the last six videos of the learning phase. We then regress $\hat{y}^{\text{pref.}} = \alpha + \beta_1 n^{\text{learning}}_{C:6} + \beta_2 n^{\text{learning}}_{R:6} + \beta_3 n^{\text{learning}}_{fR:6}$, where j:6 represents the number of videos from category $j \in \{C, R, fR\}$ in the burst. Fig. 4 shows that the marginal prediction of preference increases for $n^{\text{learning}}_{fR:6} \in \{2, 3, 4\}$ and all three types of recommendation and is positive except for $n^{\text{learning}}_{fR:6} = 2$ for up next recommendations, which is not distinguishable from 0. Similar to our main analysis, therefore, recommendations following bursts of highly partisan consumption offer greater moderating effects than for non-bursty consumption. Put another way, bursts of partisan consumption predict future consumption because they signal a change in user preferences toward more extreme content, not because the recommender is suddenly recommending more such content.

Estimating Forgetting Time of the Recommender. Recommendation algorithms have been criticized for continuing to recommend problematic content to previously interested users long after they have lost interest in it themselves (9). To understand the extent to which this is the case, we again train the bots on the trajectory of a user from the far-right end of the political spectrum, where half the bots ("short history") imitate the user for 30 videos and the other half ("long history") do so for 120 videos. In the second phase, both sets of bots switch to the trajectory of a different user, whose consumption is dominated by moderate and mainstream sources, and follow this user for another 120 videos. Throughout both phases, we tracked the recommended items in the sidebar and homepage at each step and measured the progress of the average partisanship of recommended videos. In this way, we measured the rate at which the recommender "forgets" the prior preferences of the focal user for users with different length histories. We conducted the experiment for 44 focal users—17 drawn from ψ^{fR}_{high} and 27 from $\psi^{fR}_{\text{medium}}$—where in each case, the counterfactual bot was supplied by a randomly selected history from ψ^{C} (*SI Appendix*, Table S4). Replicating the experiment for each focal user three times yielded a total of 233 trajectories comprising 45,435 watched videos and an estimated watch time of 170,381 min. We leveraged this setup to simulate two related experiments (*SI Appendix*, Fig. S9), which used data from the same underlying design in different ways.

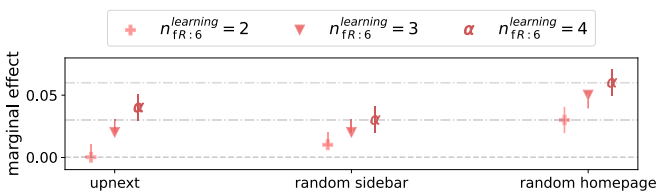

Fig. 4. Marginal effect of bursty viewership of partisan videos (calculated using ggeffects R-package) on the user preference role in future consumption. Preference increases with higher bursts of partisan consumption.

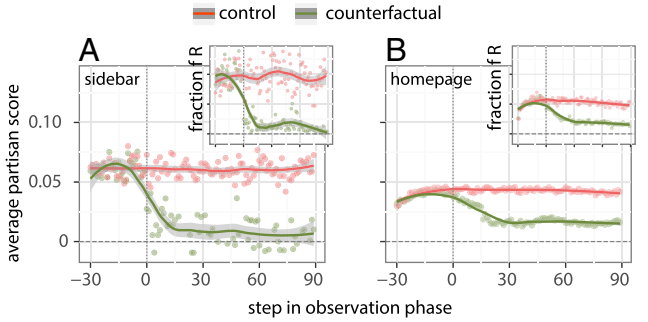

Fig. 5. Forgetting time: A comparison of the average partisan score and the fraction of recommended *fR* videos (insets) for control (red line) and counterfactual (green line) bots for sidebar (*A*) and homepage (*B*) recommendations respectively. The control bot watches 120 videos from a *fR* focal user, while the counterfactual bot after watching the same 30-video history as in the control bot, transitions to videos from a *center* focal user spanning 90 videos. (*A*) Sidebar response to this change in consumption is immediate and partisan score converge to zero. (*B*) For homepage, the average partisan score converges to moderate range; however, even after 90 post-switch videos the average fraction of *fR* videos remains nonzero (albeit much lower than for the control). For better visualization, the *y*-axis range across all panels is the same.

First, we assigned a short-history bot to be the counterfactual bot and a long-history bot as the control. For both bots, therefore, the learning phase comprised $N_{\text{learning}} = 30$ and the observational phase comprised $N_{\text{observation}} = 90$, where the control watched 120 videos from ψ^{fR} group while the counterfactual bot watched 30 ψ^{fR} videos from the same group (ψ^{fR}) and then switched to watching 90 videos from ψ^{C} i.e., moderate content, (Fig. 1*B*). Fig. 5 shows the average partisanship of sidebar and homepage recommendations for control (red line) and counterfactual (green line) bots. For sidebar recommendations (Fig. 5*A*), the counterfactual bot experienced a large and rapid decrease in partisanship relative to the control bot: Within roughly 30 videos, sidebar recommendations had become indistinguishable from those recommended to a ψ^{C} user, whereas those for the control bot remained almost as partisan as during the learning phase. Homepage recommendations (Fig. 5*B*), meanwhile, also decreased in partisanship for the counterfactual bot relative to the control, but tended to be less sensitive to user behavior than the sidebar: They were less partisan to begin with but also adjusted less rapidly to any changes, taking roughly 30 videos to become neutral on average. Fig. 5 *A* and *B Insets* show a similar pattern holds for the fraction of *fR* videos displayed: On average, *fR* videos disappeared from the sidebar recommendations between 30 and 40 into the observation phase; however, a small but non-zero fraction of *fR* videos continued to appear on the homepage until the end of the 90-step observation phase.

Effect of history-length in forgetting time. To examine whether the forgetting time of the recommender depends on the length of the learning phase, we now assign the short-history bot to the control condition with $N_{\text{learning}} = 30$ and the long-history bot to the counterfactual condition with $N_{\text{learning}} = 120$, where both bots then have $N_{\text{observation}} = 120$ (Fig. 1*C*). Thus, the control bot in this experiment watches a total of 150 videos (30 from ψ^{fR} followed by 120 from ψ^{C}) while the counterfactual bot watches a total of 240 (120 from ψ^{fR} followed by 120 from ψ^{C}). If a longer history of viewing *fR* videos causes the recommender to "remember" the user's preference for longer, we ought to see a slower decrease in partisanship during the observation phase for the counterfactual than for the control bot. In contrast, Fig. 6*A* shows no such effect in the case of sidebar

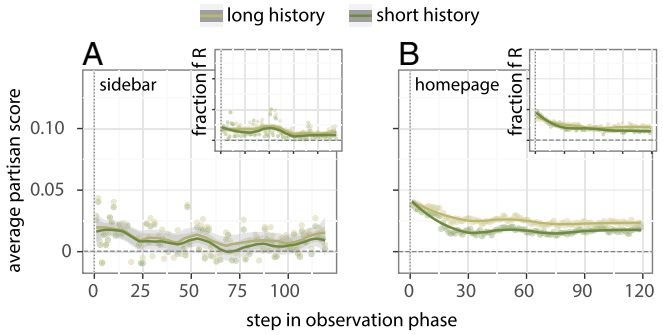

Fig. 6. Effect of history length: A comparison of the average partisan score and the fraction of recommended *far-right* videos (insets) for counterfactual bots only with short (green) and long (purple) histories of *fR* viewership for sidebar (*A*) and homepage (*B*) respectively. In the control arm, a bot watches a 30-video *fR* history followed by a 120-video *center* history, while in the treatment arm, the bot is exposed to an additional *fR* history lasting 120 videos. (*A*) On the sidebar, both longer and shorter history exhibit the same drop rate in terms of average partisan score and average fraction of *fR* content. (*B*) On the homepage longer history reduces the drop rate of partisanship on both metrics. Even after 90 steps, the average partisan score of bots with longer history remains higher than that of the shorter path. For better visualization, the y-axis range across all panels is the same.

recommendations: although the path with a longer viewing history exhibits slightly higher average partisanship scores, both paths exhibit a similar drop rate, and both paths converge toward zero recommendations of *fR* videos (Fig. 6 *A*, *Inset*). On the other hand, Fig. 6*B* shows that homepage recommendations do display a slower drop rate: whereas the average partisanship of videos in both paths stabilizes around step 30, the path with a longer history exhibits a gradual decrease that persists until the end of the observation window. The fraction of *fR* videos drops along the trajectory, where from step 70 they diverge slightly (Fig. 6 *B*, *Inset*).

Discussion

Online platforms such as YouTube are regularly accused of amplifying politically extreme content via their recommender systems and thereby driving their users toward conspiratorial beliefs. Superficially, these accusations appear plausible: Many users rely on recommendations to find new content; some of that content is indeed objectionable; and some users do indeed consume it. It is important to note, however, that even strong correlational evidence of this sort does not constitute evidence that the recommender itself is causing the problematic behavior. Users of online platforms also exhibit considerable agency and might have consumed the same content, or worse, even in the absence of recommendations.

Previous empirical work has struggled to tease out the specific causal role of platform recommendations in large part because of the absence of a proper counterfactual. In some cases (4), we can observe the real users' consumption but not the consumption of a counterfactual user who relied exclusively on recommendations. In other cases (5, 15), the opposite applies: We observe what a synthetic rule-following user (either a bot or a human) would be exposed to, but not what a counterfactual user who only followed their own preferences would see. Ideally, we would like to see both the real user and their rule-following counterfactual: If the latter receives more extreme recommendations than the former, that would be evidence that the recommender is amplifying extreme content; and if it receives less extreme recommendations, that would be evidence that the recommender exerts a moderating effect.

In this paper, we have implemented precisely this design using a combination of real user data and automated bots: Logged-in, programmatic users capable of following arbitrary viewing patterns. In our experiments, the bots are assigned to one of two conditions: the control bot imitates the behavior of a real focal user, whereas the counterfactual bot initially imitates the behavior of the same user but then switches to a different behavior such as clicking on the top-ranked sidebar (aka up next) recommendation (Fig. 1). By comparing the experience of the counterfactual bot with that of the control, we can estimate the causal effect of the recommender. Moreover, by selecting different types of focal users—defined by the amount of far-right (*fR*) vs. centrist (*C*) content they consumed—we can measure how the causal effect varies with user behavior.

Our results suggest that, on average, relying solely on the recommender results in a more moderate experience on YouTube relative to the real user, where the effect is mostly driven by extreme users (Fig. 3 and Table 1) and for users who consume bursts of *fR* videos (Fig. 4). Further, we find that when consumers of partisan content change to moderate content, the sidebar reacts quickly and *fR* content, on average, decreases to zero after 30 steps, while homepage recommendations react more slowly (Fig. 5). We also find that the "forgetting rate" for the homepage is longer for users with longer histories, whereas sidebar recommendations are unaffected (Fig. 6).

Overall, our study reinforces previous work (4, 11, 17, 20, 21) that places individual human preferences at the center of platform dynamics. While recommendations and other platform affordances no doubt shape user experiences to some degree (22), our results suggest that popular narratives (8, 9, 23, 24) about the widespread and profound manipulative impact of algorithms are overstated. This is not to say that highly problematic content does not exist on social media platforms, that it does not have harmful effects on those who consume it, or that platforms should not be held responsible for mitigating these effects. Rather, by shifting the emphasis of the concern from presumed biases in algorithms to the factors governing the supply and demand of problematic content, social media companies and their critics can more accurately target the source of the problem, which may transcend any one platform however large. For example, recent work (25) shows that shutting down the right-wing social media site Parler had little impact on the overall consumption of conspiratorial content, as users simply replaced their diets of such content via other sources on the web.

Although we believe our contribution constitutes a meaningful advance for studying the causal effects of platform design, it nonetheless has limitations. First, as noted earlier, in 2019 YouTube implemented significant changes to its algorithm that it claimed reduced watch time of "borderline content and harmful misinformation" by 50 to 70% (16). It is therefore possible that to some extent the difference between our findings and pre-2019 claims of the radicalizing effects of YouTube's algorithm can be attributed to changes to the algorithm. Unfortunately, testing this hypothesis would require recreating YouTube as it existed prior to the change, which is to our knowledge impossible; thus, our findings should be interpreted as applying only to the post-2019 period. Second, our experiments were conducted in early 2023, whereas our empirical data were recorded between October 2021 through December 2022. Although we are not aware of any major changes to YouTube's moderation policy or recommendation system in the intervening months, and we conducted multiple

iterations of each experiment in order to account for randomness and other time-varying factors, our experiment was not a true field experiment. Third, our empirical panel data are restricted to desktop users and hence do not include YouTube consumption on mobile devices, which could potentially be different. Fourth, for feasibility, we sped up the bot viewing to simulate several months of real user activity in two to three days. Although we do not believe that speeding up the watch time meaningfully altered the recommender's reactions, we cannot rule out that the same experiment conducted over many months would yield different results. Fifth, the scoring is done at the channel level, which is not entirely accurate as there may be differences in partisanship levels across videos within a channel. Future work would benefit from a video-level scoring approach to identify partisanship of content more precisely. In spite of these limitations, we hope our work will stimulate researchers of socio-technical systems to adopt counterfactual bot designs. We believe these designs strike a useful balance between taking real user behavior seriously and exploiting the flexibility, speed, and data-recording capabilities of programmatic users. In this sense, our study can also be viewed as a proof of concept for an approach to studying the interactions between humans and algorithms across many online platforms and services, not just YouTube.

Materials and Methods

Dataset. Our data are derived from Nielsen's nationally representative desktop web panel, which tracks individuals' visits to URLs from October 2021 to December 2022, including a total of 87,988 panelists. Each YouTube video has a unique identifier embedded in its URL. By parsing the recorded URLs, we find the subset of 48,026 users who have at least one recorded YouTube video viewership. To post a video on YouTube, a user must create a channel with a unique name and channel ID. For all unique video IDs collected from Nielsen and recorded in the experiments, we used the YouTube API to retrieve the corresponding channel ID, as well as metadata such as the video's category, title, and duration. We then use the channel IDs to assign a partisanship score to each video based on the political leaning of its channel. Table 2 provides more details on data statistics.

"User History" Selection. Our unit of analysis in this paper is user history, where we focused on heavy consumers of *far-right* content. To ensure a comprehensive and representative selection of user histories, we employed a systematic approach. Initially, we searched across all 4,583 users who had watched a minimum of 140 YouTube videos and sampled trajectories with a length of 120 videos by choosing a random start point between 1 and $M_i - 120$, where M_i is the total number of YouTube video views for the ith user. From each user, we randomly selected multiple histories according to their lifetime on the panel, resulting in 24,871 histories, with 12,969 having at least 1 min of news consumption (from 3,089 unique news users). We continued by grouping histories based on their news consumption archetype using the first $N_{\text{learning}} = 60$ videos. We did so to avoid looking into future consumption, which will be used for evaluation purposes (in this way, when a user history is assigned to an archetype, the observation period is not known, and there is no leakage of future information in the assignment of users to experiments, i.e., we do not keep users who already have a high consumption of *fR* in the observation period). Following the same approach as ref. 4, we characterized every user history in terms of their normalized news viewership vector. We adopted a source-based approach where we assigned all videos produced by a channel the same partisan score. To derive the political partisanship scores of channels, we leveraged the embeddings of approximately 7.5 million channels provided by ref. 19, which incorporated the Reddit embeddings developed by Waller and Anderson (26). The scores were validated using existing lists of left- and right-wing YouTube channels (e.g., ref. 4), resulting in a rank correlation of 0.65. Further, in a crowdsourcing task, the authors found agreement between embedding and crowd workers to be above 80%, indicating the robustness of their approach (details can be found in *SI Appendix*, section 2). Overall, 20% of the collected video IDs do not have a partisan score attached to them, and for the presented results in the main text, we have dropped such videos from our analysis. To validate the robustness of these findings, we have replicated our analysis where missing values are imputed (*SI Appendix*, Tables S6 and S7 and section *Estimating Bias of the Recommender* for more details). With the average partisan score zero and the SD $\sigma = 0.08$, any video with a partisan score in range $(-\sigma, \sigma)$ is labeled as *C*; $(-\sigma, -2\sigma)$ and $(\sigma, 2\sigma)$ are labeled as *L* and *R*, respectively; anything to the left of left $(-\sigma, -2\sigma)$ is labeled as far left, *fL*, and anything to the right of right is labeled as far-right, *fR*. The normalized viewership vector of ith history is v_i, whose jth entry v_{ij} corresponds to the fraction of viewership of ith user-history from jth category ($j \in \{fL, L, C, R, fR\}$). We then used hierarchical clustering to assign each user history to one of $K = 8$ communities of similar YouTube news diets (*SI Appendix*, Fig. S2). We ended up with 144 histories with heavy *fR* consumption (from 90 unique YouTube users), which we used to select histories for this study. To better understand the underlying patterns within this category of behavior, again, we employed a hierarchical clustering algorithm, grouping *fR* histories into three distinct archetypes, each representing a unique pattern of consumption of *fR* videos, as depicted in *SI Appendix*, Fig. S3. To ensure a balanced and representative analysis of the results, we either select all histories or randomly select a subset, depending on the size of the cluster. In all subsequent analyses, we weighted these samples to accurately reflect the true distribution of histories within the overall population.

Table 2. Data descriptive statistics

Number of panelists	87,988
Number of YouTube consumers	48,026
Total number of watched trajectories	1,537
Number of watched videos by bots	201,915
Estimated total watched time (min)	811,356

Designing Bots. Overall, we created more than 150 Google accounts for this study, and each account has been used across multiple experiments. Our test experiments show that logged-out browsing behaves differently from logged-in accounts regarding the similarity of recommended items with the watched history. Further, as personalization in the logged-out approach is via browser cookies, it is specific to a particular browser session. It imposes technical limitations for very long sessions, which may take days, as any interruption and browser reset may lead to loss of historical information. Therefore, we run all experiments with logged-in bots (*SI Appendix*, Fig. S1). The web crawler includes functionality designed to reset YouTube accounts to a clean state. This feature enables the crawler to log into the user's account, access the "Your data in YouTube" section, and clear the watch history. By doing so, all user activity data on YouTube are effectively removed, and the recommendations are reset to a state similar to that of an incognito window, based on our knowledge. The empirical validity of this approach has been confirmed by the experiments, where initial measurements do not indicate any presence of previously watched topics on this account. See *SI Appendix*, section 1 for more details on the data collection pipeline.

In all experiments, each trajectory is divided into two parts: learning and observation. In the learning phase, bots are first "primed" with real user histories by watching their videos from the corresponding focal user. This is equivalent to creating multiple copies of the same account with personalized recommendations. Recognizing the importance of variations in the "Watch Time" for the recommendation system to learn users' interests in different topics (27, 28), we allocate a watch time to each video that is proportionate (half of the actual watch time) to the real user's video viewing duration from Nielsen data. Moreover, we introduce pauses (half of the actual pause duration) between videos that correspond to the behavior of real users, thereby enhancing the accuracy of mimicking their viewing patterns. For both watch time and idle times, we set a maximum limit of 10 and 20 min, respectively, to ensure the feasibility of the experiments. Upon completing each experiment, we retrieve metadata associated with each video ID in our collection using the YouTube API. Only a small percentage, less than 3.1%, of video IDs do not produce metadata from the API.

Data, Materials, and Software Availability. Our data are derived from Nielsen's nationally representative desktop web panel, which tracks individuals' visits to URLs from October 2021 to December 2022, including a total of 87,988 panelists (48,026 users who have at least one recorded YouTube video viewership) encompassing a total of 351,096,850 rows of viewership from the web on their desktop device. Each row includes the visited URL, the activity start time, and the duration (credited to an in-focus page). The data are made available to us by the Nielsen Corporation under an agreement with the University of Pennsylvania that prohibits sharing with any third parties without Nielsen's prior consent. Interested parties should contact the corresponding author for further information.

ACKNOWLEDGMENTS. We are grateful to the Nielsen Company for access to their desktop panel data and to B. Sissenich, S. Sherman, H. Baberwal, and E. Grimaldi for ongoing support. Additionally, H.H., M.R.-L., and D.J.W. are grateful for the financial support provided by Richard Jay Mack and the Carnegie Corporation of New York (Grant G-F-20-57741). A.G. is supported by the NSF under Grant No. 2030859 to the Computing Research Association for the CIFellows Project. M.H.R. and R.W. acknowledge support from the Swiss NSF (Grant 200021_185043) and the European Union (TAILOR, Grant 952215).

Author affiliations: [a]Department of Computer and Information Science, University of Pennsylvania, Philadelphia, PA 19104; [b]Annenberg School of Communication, University of Pennsylvania, Philadelphia, PA 19104; [c]Yale Institute for Network Science, Yale University, New Haven, CT 06511; [d]Heinz College of Information Systems and Public Policy, Carnegie Mellon University, Pittsburgh, PA 15213; [e]School of Computer and Communication Sciences, École Polytechnique Fédérale de Lausanne, 1015 Ecublens, Switzerland; and [f]Operations, Information, and Decisions Department, University of Pennsylvania, Philadelphia, PA 19104

1. A. Schomer, "US YouTube advertising 2020" *eMarketer* (2020). https://www.emarketer.com/content/us-youtube-advertising-2020. Accessed 27 December 2021.
2. T. Konitzer *et al.*, Measuring news consumption with behavioral versus survey data. Pew Research. https://www.pewresearch.org/journalism/2020/12/08/measuring-news-consumption-in-a-digital-era/. Accessed 27 December 2023.
3. M. Iqbal, Twitter revenue and usage statistics. Business of Apps. https://www.businessofapps.com/data/twitter-statistics/. Accessed 27 December 2023.
4. H. Hosseinmardi *et al*., Examining the consumption of radical content on YouTube. *Proc. Natl. Acad. Sci. U.S.A.* **118**, e2101967118 (2021).
5. M. A. Brown *et al.*, Echo chambers, rabbit holes, and algorithmic bias: How YouTube recommends content to real users. SSRN. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4114905. Accessed 27 December 2023.
6. E. Hussein, P. Juneja, T. Mitra, Measuring misinformation in video search platforms: An audit study on YouTube. *Proc. ACM Human-Comput. Int.* **4**, 1–27 (2020).
7. M. H. Ribeiro, R. Ottoni, R. West, V. A. Almeida, W., Meira Jr, "Auditing radicalization pathways on YouTube" in *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency* (2020), pp. 131–141.
8. Z. Tufekci, Youtube, the great radicalizer. *NY Times*, 10 March 2018. https://www.nytimes.com/2018/03/10/opinion/sunday/youtube-politics-radical.html. Accessed 27 December 2023.
9. K. Roose, The making of a YouTube radical. *NY Times*, 8 June 2019. https://www.nytimes.com/interactive/2019/06/08/technology/youtube-radical.html. Accessed 27 December 2023.
10. A. Narayanan, Understanding social media recommendation algorithms, *Knight First Amendment Institute*, 9 March 2023. https://knightcolumbia.org/content/understanding-social-media-recommendation-algorithms. Accessed 27 December 2023.
11. M. Horta Ribeiro, V. Veselovsky, R. West, The Amplification Paradox in Recommender Systems. *Proc. Int. AAAI Conf. Weblogs Soc. Media* **17**, 1138–1142 (2023).
12. A. D'Amour *et al.*, "Fairness is not static: Deeper understanding of long term fairness via simulation studies" in *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency* (2020), pp. 525–534.
13. A. Sinha, D. F. Gleich, K. Ramani, "Deconvolving feedback loops in recommender systems" in *Advances in Neural Information Processing Systems* (2016), p. 29.
14. K. Garimella, T. Smith, R. Weiss, R. West, "Political polarization in online news consumption" in *Proceedings of the International AAAI Conference on Web and Social Media* (2021), vol. 15, pp. 152–162.
15. M. Haroon *et al*., Auditing YouTube's recommendation system for ideologically congenial, extreme, and problematic recommendations. *Proc. Natl. Acad. Sci. U.S.A.* **120**, e2213020120 (2023).
16. A. Y. Chen, B. Nyhan, J. Reifler, R. E. Robertson, C. Wilson, Subscriptions and external links help drive resentful users to alternative and extremist YouTube channels. *Sci. Adv.* **9**, eadd8080 (2023).
17. K. Munger, J. Phillips, Right-wing YouTube: A supply and demand perspective. *Int. J. Press/Polit.* **27**, 186–219 (2022).
18. T. Yang, S. González-Bailón, Online media boosts exposure to news but only for a small minority of hyper-consumers. SSRN. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3954565. Accessed 27 December 2023.
19. L. Boesinger, M. H. Ribeiro, V. Veselovsky, R. West, Tube2Vec: Social and semantic embeddings of YouTube channels. arXiv [Preprint] (2023). https://doi.org/10.48550/arXiv.2306.17298 (Accessed 27 December 2023).
20. A. M. Guess *et al*., How do social media feed algorithms affect attitudes and behavior in an election campaign? *Science* **381**, 398–404 (2023).
21. R. E. Robertson *et al*., Users choose to engage with more partisan news than they are exposed to on Google search. *Nature* **618**, 342–348 (2023).
22. S. González-Bailćn *et al*., Asymmetric ideological segregation in exposure to political news on Facebook. *Science* **381**, 392–398 (2023).
23. E Pariser, *The Filter Bubble: How the New Personalized Web is Changing What We Read and How We Think* (Penguin Books, 2012).
24. J. Haidt, J. Twenge, Social media and mental health: A collaborative review [Unpublished manuscript, New York University] (2023) (Accessed 27 December 2023).
25. M. Horta Ribeiro, H. Hosseinmardi, R. West, D. J. Watts, Deplatforming did not decrease Parler users' activity on fringe social media. *PNAS Nexus* **2**, pgad035 (2023).
26. I. Waller, A. Anderson, Quantifying social organization and political polarization in online platforms. *Nature* **600**, 264–268 (2021).
27. P. Covington, J. Adams, E. Sargin, "Deep neural networks for YouTube recommendations" in *Proceedings of the 10th ACM Conference on Recommender Systems* (2016), pp. 191–198.
28. C. Goodrow, "On YouTube's recommendation system" *YouTube Official Blog* (2015). https://blog.youtube/inside-youtube/on-youtubes-recommendation-system/. Accessed 27 December 2023.