

The Rabbit Hole Is Five Claims, Not One: Why Audits, Browsing Panels, Escalation Studies, Exposure Experiments and Discourse Maps of Online Conspiracy Content Keep Disagreeing

Pranay Mahendrakar

SONYTECH · ORCID 0009-0003-7224-029X

pranaymahendrakar@sonytech.in · mahendrakarpranay@gmail.com · +91 6361723454

Abstract. The claim that online platforms draw people down conspiracy "rabbit holes" is among the most repeated in public debate about recommender systems, and the published evidence on it appears to contradict itself. Sock-puppet audits find recommendations that narrow and drift toward problematic channels; browsing panels and counterfactual-bot experiments find that consumption tracks prior preference and that the recommender, if anything, moderates it; a 2025 study of Singapore Telegram groups reports that conspiratorial talk is woven into ordinary conversation rather than sealed in enclaves. In this analysis the author reviews and synthesises published results to show that the metaphor bundles five separable claims: that a recommender followed blindly drifts toward such content (drift), that real consumption of it is caused by recommendation (uptake), that individuals escalate over time (escalation), that exposure changes belief (effect), and that the content lives in isolated spaces (location). Each design in the literature identifies at most two of the five, and most of the apparent conflict comes from reading a result about one link as evidence about another, or across YouTube's 2019 recommender change. The paper contributes a link-by-design map, a consolidated table of more than thirty published results indexed by link and period, a link-attribution procedure for reading any new study, and a re-reading of the Telegram study's own tables, which show conspiratorial share ranging from 5 to 70 percent across its clusters in a corpus 95 percent drawn from one group. The evidence is strong that post-2019 YouTube drift is weak and that exposure is concentrated in a small, motivated tail; it is weak on long-horizon belief effects, on platforms other than YouTube, and on location claims without an unselected comparison set.

Keywords: recommender systems; algorithmic amplification; conspiracy theories; online radicalization; echo chambers; sock-puppet audits; research design; causal identification

SONYTECH ORCID: 0009-0003-7224-029X

1. Introduction

Few claims about recommender systems have travelled as far as the rabbit hole. Journalists, sociologists and legislators have described YouTube's recommendations as a radicalising engine, a framing that Chen et al. (2023) trace through Tufekci's widely quoted line that the system "may be one of the most powerful radicalizing instruments of the 21st century". The concern drove platform policy: in January 2019 YouTube announced that it would recommend less borderline and conspiratorial content, a change whose effects Faddoul, Chaslot and Farid (2020) tracked with fifteen months of emulated watch-next recommendations. It drove a research program as well. A systematic review of 23 studies on YouTube recommendations and problematic content counted 14 that implicated the recommender, seven with mixed results and two that did not (Yesilada & Lewandowsky, 2022), and a broader review of 496 articles on digital media and democracy found associations running in both beneficial and harmful directions depending on the outcome and the political system (Lorenz-Spreen et al., 2022).

The opposite position is now at least as well supported. A panel of more than 300,000 Americans browsing from 2016 to 2019 showed that news viewing on YouTube

was dominated by mainstream sources, that the far-right audience was small and stable, and that its consumption was not systematically traceable to recommendations (Hosseinmardi et al., 2021). Paired browsing and survey data from 2020 found that alternative and extremist videos were watched mostly by subscribers with high prior hostile sexism and racial resentment, who often arrived through external links (Chen et al., 2023). Bots that copied real users and then switched to following only the recommender ended up with less partisan diets than the users they copied (Hosseinmardi et al., 2024). A series of experiments that manipulated real YouTube recommendations for nearly 9,000 participants could not detect consistent attitude effects (Liu et al., 2025), and a review in *Nature* argued that exposure to false and inflammatory content online is low on average and concentrated among a narrow, motivated fringe (Budak et al., 2024).

A third line of work moves the question away from algorithms altogether. Shimgekar et al. (2025) classified 553,648 messages from six Singapore-based Telegram groups, clustered them into seven narrative archetypes, and concluded that conspiratorial talk sits inside everyday discussion of law, finance and daily life, a finding they present as challenging common assumptions about online radicalization. Tschofenig et al. (2026) argue that rabbit holes can arise without any recommender, through interactions among conspiracy narratives spreading as social contagions among X users.

Read as answers to one question, these results are irreconcilable. This paper argues that they are not answers to one question. "Platforms pull people down conspiracy rabbit holes" bundles at least five claims, each with its own estimand, its own counterfactual and its own failure mode. The first is about what the recommender would serve to someone who simply follows it. The second is about whether real people's consumption is caused by what they are served. The third is about whether individuals move toward more extreme or more numerous conspiracy beliefs over time, whatever the cause. The fourth is about whether exposure changes what people believe. The fifth is about where conspiratorial content sits: in sealed enclaves or within ordinary conversation. A sock-puppet audit speaks to the first claim and is silent on the other four. A browsing panel speaks to the second. A comment-migration study speaks to the third. A field experiment speaks to the fourth over the horizon it runs. A discourse map speaks to the fifth, and only for the population it sampled.

In this analysis the author reviews and synthesises published results to show three things. First, once results are indexed by link and by period, most of the apparent contradiction dissolves: drift was real and measurable on YouTube before 2019 and is weak afterwards; uptake is dominated by prior demand in every behavioral dataset that can see it; escalation is documented, but in samples selected on the outcome; effects on attitudes are small and short-horizon wherever they have been tested; and location claims depend entirely on the sampling frame. Second, the evidence complicates both of the stories the field tells. It complicates the rabbit-hole story, because the strongest causal designs do not find the algorithmic push. It equally complicates the "ordinary interaction" counter-story, because the Telegram study's own cluster table shows conspiratorial content heavily concentrated in one archetype, and its corpus is dominated by a single group whose name and theme the sampling did not control. Third, what remains open is specific and can be stated: the heavy-consumer tail, long-horizon belief change, platforms other than YouTube, and the pre-2019 counterfactual that no one can now recreate.

The paper makes six contributions. It stipulates the five links and the evidence each would require (Section 2). It reviews the evidence for each link in turn (Sections 3 to 7), and isolates the 2019 period break as a confound that cuts across all five (Section 8). It consolidates the measured values in Table 1 and the design-to-link map in Table 2, with Figure 1 re-presenting two published distributions that carry the argument and Figure 2 placing each study's data window against the 2019 change (Section 9). It states a link-attribution procedure as Algorithm 1 (Section 10). It identifies where the evidence complicates both prevailing stories (Section 11), and states what would settle the open part (Section 12). Sections 13 to 16 give the strongest case against this paper, what is not known, and its limitations.

The argument is not that recommenders are harmless, and it is not that the Telegram study is wrong about its groups. Several audits document real and persistent drift toward problematic channels for some simulated users (Haroon et al., 2023; Gunasekara et al., 2026), and the Telegram study's message-level finding stands for the groups it sampled. The argument concerns inference: which link a given design can establish, and which it cannot.

2. Five links and what each would require

This section stipulates the five claims and, for each, the counterfactual that a study would need to identify it. The definitions are the author's. They are offered as instruments for reading the literature rather than as a summary of any one source, although each tracks a distinction that at least one cited paper draws.

Link 1, drift. A recommender exhibits drift toward a content class if a user who follows its recommendations, without any preference of their own beyond the starting point, is served a rising share of that class as the session or history lengthens. The estimand is a property of the system conditional on a behavior rule. Identifying it requires control over the user's choices, which is what sock puppets provide. Brown et al. (2026) separate three versions of this claim for YouTube: ideological echo chambers, "content rabbit holes" in which recommendations stay close to what is being watched, and radicalization pathways in which they move toward more extreme material. The distinction matters, because a recommender can produce narrowing without producing escalation.

Link 2, uptake. Consumption of a content class is algorithmically caused, in the sense relevant to the rabbit-hole claim, if real users would have consumed less of it had the recommender behaved differently. The estimand is a counterfactual about real people with real preferences. Hosseinmardi et al. (2024) put the difficulty precisely: panel studies see only what users clicked, not what they were shown, and audit studies see what a rule-following agent is shown but not what the user would otherwise have chosen. Neither isolates the causal share without a counterfactual built from both.

Link 3, escalation. Individuals escalate if, over time, they adopt more conspiracy narratives, more extreme ones, or move from milder to more extreme communities. This is a claim about trajectories and is agnostic about cause. A trajectory can be driven by a recommender, by social contagion, by an external shock or by a pre-existing disposition unfolding. Identifying escalation requires longitudinal data on individuals; identifying its cause requires a comparison group that did not receive the hypothesised driver.

Link 4, effect. Exposure has an effect if it changes beliefs, attitudes or behavior relative to non-exposure. Identifying it requires randomised or as-if-random variation in exposure and a measured outcome. The horizon matters: an experiment that runs for a week can

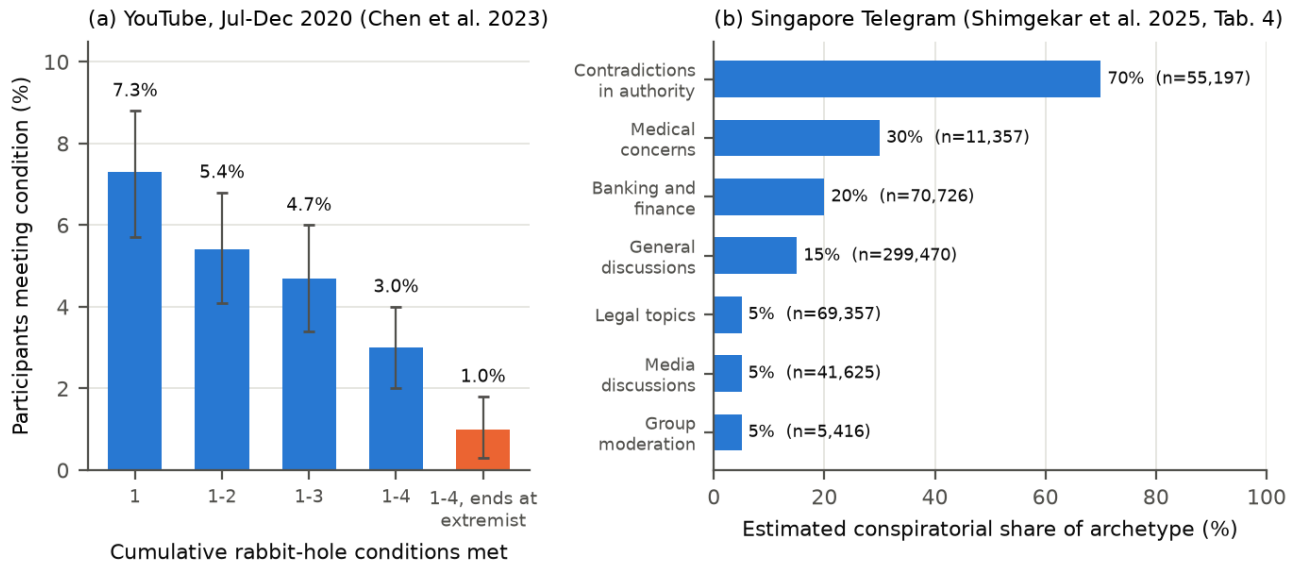


Figure 1. (a) Share of 1,181 YouTube users in July-December 2020 meeting cumulative conditions of a rabbit-hole event: (1) followed a recommendation to an alternative or extremist video, (2) the move was toward a more extreme channel type, (3) not subscribed to that channel, (4) subscribed to no channel of that type. Values and 95% confidence intervals from Chen et al. (2023); the last bar is the subset ending at an extremist video. (b) Estimated share of messages labelled conspiratorial in each of the seven narrative archetypes of Singapore Telegram messages (the archetype counts in the source's Table 4 sum to 553,148, 500 fewer than the 553,648-message corpus), computed by the author from the counts in Shimgekar et al. (2025), Table 4; bar labels give message counts. The source describes the counts as estimates.

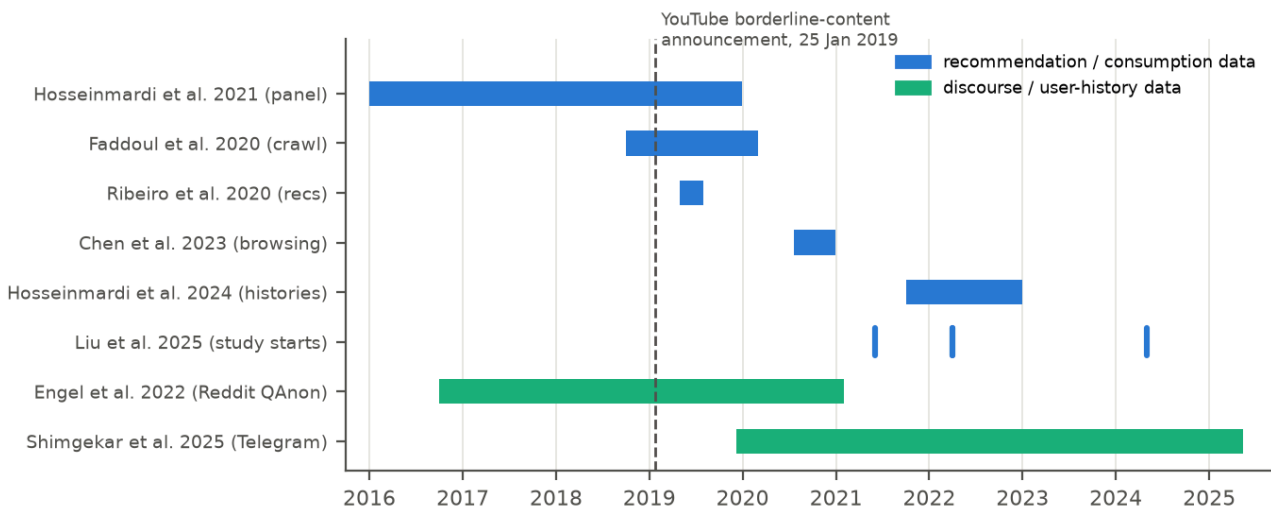


Figure 2. Data-collection windows of eight studies discussed in Sections 3 to 8, as reported by each source, against YouTube's 25 January 2019 announcement on borderline content (dashed line). Liu et al. (2025) report start months only (June 2021, April 2022, May 2024), drawn as short marks. Studies reporting strong recommender drift collected their data before or during the change; studies reporting preference-driven consumption collected after it, with the exception of Hosseinmardi et al. (2021).

rule out large one-week effects and is silent about the cumulative effect of years.

Link 5, location. Conspiratorial content is enclaved if it is concentrated in spaces where participants rarely encounter other content or other people, and embedded if it appears within spaces that also carry ordinary conversation. This is a claim about the joint distribution of content and space. Identifying it requires a sampling frame that includes spaces not selected for their conspiratorial character, because the share of ordinary

talk inside a conspiracy-adjacent space says nothing about how much conspiratorial talk occurs in spaces that are not.

These links are logically independent. A recommender can drift (Link 1) while real users ignore the drift (Link 2 fails), as the amplification-paradox model of Ribeiro, Veselovsky and West (2023) shows is expected when niche content has low utility for most users. Individuals can escalate (Link 3) with no recommender involved, which is the hypothesis of Tschopenig et al. (2026).

Exposure can be heavily algorithmic and still have no detectable attitude effect over the measured horizon (Link 4 fails), which is the pattern the 2020 Facebook and Instagram experiments report (Guess et al., 2023; Nyhan et al., 2023). And content can be both embedded and enclaved, if it appears inside groups that are themselves sealed off from the rest of the platform. The common public claim, that recommendation algorithms radicalise people into conspiracy belief, requires Links 1, 2, 3 and 4 to hold jointly, with Link 2 doing the causal work. No single study in this review tests all four.

Two further terms are used throughout. "Pre-2019" and "post-2019" refer to YouTube data collected before or after the January 2019 announcement on borderline content, which Section 8 treats as a period break. "Problematic" is used as each source uses it, which varies: Haroon et al. (2023) mean channels on curated lists labelled IDW, Alt-right, Conspiracy and QAnon; Chen et al. (2023) mean alternative and extremist channels defined by prior lists; Faddoul et al. (2020) mean videos their classifier scores as conspiratorial. These categories are not interchangeable, and Section 13 returns to what that costs the synthesis.

3. Link 1: what the audits show about drift

Sock-puppet and crawl audits are the evidence most often cited for the rabbit hole, and they are the right design for Link 1. Their results depend sharply on platform, period, personalisation and content class.

Pre-change and transition-period YouTube. Ribeiro et al. (2020) analysed 330,925 videos on 349 channels sorted into media, Alt-lite, Intellectual Dark Web and Alt-right categories, together with more than two million video and channel recommendations collected between May and July 2019. They report that Alt-lite channels were easily reachable from IDW channels, particularly through channel recommendations, and that Alt-right channels could be reached through channel recommendations but were rarely recommended in the video graph. Faddoul et al. (2020) emulated the watch-next algorithm starting from more than a thousand popular informational channels and collected more than eight million recommendations over fifteen months from October 2018 to February 2020. The share of recommendations their classifier scored as conspiratorial peaked near 10 percent in raw frequency and 6 percent when weighted by the popularity of the source video, at the end of 2018. They also report a clear positive association between the conspiracy likelihood of the source video and that of the recommended video, which they read as a filter-bubble effect, strongest in the October 2018 to January 2019 window. Roth, Mazieres and Menezes (2020), crawling non-personalised recommendation graphs from diverse seeds, found them often prone to confinement, with the most confined graphs organised around the most-viewed videos.

After the change. Faddoul et al. (2020) record a consistent decline starting in April 2019, with raw frequency briefly reaching a low of 3 percent in early

June 2019. The popularity-weighted measure then rebounded, and by the end of their window conspiratorial recommendations were only 40 percent less common than when the policy was first announced. The reduction was applied mainly at the channel level: the ten channels that had carried the most conspiratorial recommendations before April 2019, together more than 20 percent of them, fell to under 0.5 percent, while a set of five channels including Fox News and the Science Channel rose from under 0.5 percent to 25 percent of conspiratorial recommendations. Ledwich and Zaitsev (2020), categorising nearly 800 political channels and using anonymous accounts with no watch history, concluded that the algorithm favoured mainstream and cable news over independent channels. Their follow-up with personalised personas reported moderate filter bubbles for most personas but a weak effect for personas engaged with niche content such as Conspiracy and QAnon channels, and a larger personalisation effect on the homepage than on Up Next (Ledwich et al., 2022).

Personalisation and topic. The audits that personalise report more drift than those that do not, and the size of the effect varies by topic. Hussein, Juneja and Mitra (2020) found that age, gender and location had no significant effect on misinformation in search results for new accounts, but that once an account had a watch history, watching misinformative videos led to more of them in Top 5 and Up-Next recommendations for every topic they audited except vaccine controversies. Papadamou et al. (2022), working with 6.6K videos on COVID-19, flat earth, anti-vaccination and anti-mask content, found more pseudoscientific suggestions for long-established topics than for emerging ones, more on the search results page than on the homepage or in the recommendation sidebar, and a substantial effect of watch history. Tomlein et al. (2021) and Srba et al. (2022) showed that misinformation filter bubbles could be entered and then burst by watching debunking content, that bubbles did not appear in some configurations at all, and that the overall quantity of recommended misinformation had not improved relative to an earlier study.

The largest audit. Haroon et al. (2023) trained 100,000 sock puppets across the ideological spectrum. They report that the homepage recommended congenial content: 77 percent of top-1 homepage recommendations for very-left puppets were left or very-left, and 81.7 percent for very-right puppets were right or very-right. Congenial recommendations increased deeper in the Up-Next trail for right-leaning puppets but not for left-leaning ones. They found no meaningful increase in ideological extremity with trail depth, but a growing proportion of recommendations from channels on their problematic lists. That proportion peaked at 2.5 percent, yet over 36.1 percent of puppets encountered at least one such recommendation, rising to 40 percent for very-right puppets. Brown et al. (2026), using a separate audit and a video-level ideology estimate, report the same shape: recommendations stay close to the current video,

which they call content rabbit holes, but do not lead into increasingly extreme material, and with longer recommendation-following all users are pushed toward moderately conservative content within a narrowing ideological range.

Other platforms. TikTok audits show stronger and topic-dependent effects. Gunasekara et al. (2026) report that new accounts with no search history were shown substantially more misinformation than established ones, and that engaging with misinformation on one topic led to misinformation on related and unrelated topics. Pecher et al. (2026), auditing personalisation drift with controlled accounts, report what they call "a strong neutralising effect" for topics themed on misinformation, alongside reinforcement of user stance on US politics. These two audits partly disagree about misinformation, and neither shares a protocol with the YouTube studies.

What the audits cannot say. Bouchaud and Ramaciotti Morales (2024), using a simulated engagement-optimising platform grounded in behavioral data, show that sock-puppet audit outcomes depend on parameters such as the number of followed accounts and session length. More fundamentally, an audit measures what a rule-following agent is served. Ribeiro, Veselovsky and West (2023) give a mechanism for why that can diverge from what people consume: in an agent-based model with collaborative filtering and niche content of low utility to most users, blindly following recommendations reaches the niche, while real users rarely choose it, which can lead the recommender to deamplify it. On this reading, the audit literature's consistent finding of narrowing and occasional drift is a statement about Link 1 that is compatible with Link 2 failing. The audits establish that a path to conspiratorial content exists in the recommendation graph, and on TikTok that the path can be short. They do not establish that people take it.

4. Link 2: whether real consumption is caused by recommendation

Behavioral data on real users is the evidence for Link 2. It comes in three forms: passive browsing panels, browsing paired with captured recommendations, and counterfactual experiments.

Panels. Hosseinmardi et al. (2021) analysed the individual-level browsing of a representative panel of more than 300,000 Americans on and off YouTube from January 2016 through December 2019. News consumption was dominated by mainstream, largely centrist sources; far-right consumers were more engaged than average but a small and stable share; and "anti-woke" content grew steadily and correlated with off-platform far-right consumption. They found no evidence that far-right engagement was systematically caused by recommendations, and no clear evidence that anti-woke channels were a gateway. Their reading is that political consumption on YouTube reflects preferences that extend across the web. The same pattern appears on search. Pairing surveys with captured Google Search results in the 2018 and 2020 US elections, Robertson et

al. (2023) found more identity-congruent and unreliable sources in what participants chose to engage with than in what they were shown, and concluded that exposure to and engagement with such news were driven primarily by users' choices rather than by algorithmic curation.

Browsing with recommendations. Chen et al. (2023) are the closest any study comes to observing Link 2 directly, because their browser extension recorded both what 1,181 participants watched and what YouTube recommended to them, from July to December 2020. Although 91 percent of participants visited YouTube, only 15.4 percent watched any video from an alternative channel and 6.1 percent from an extremist one. Viewing was concentrated: 1.7 percent of participants accounted for 80 percent of time spent on alternative-channel videos and 0.6 percent for extremist-channel videos. It was also subscriber-driven: 60.8 and 54.7 percent of views of alternative and extremist videos came from subscribers to the channel being watched, and 92.9 and 84.7 percent from people subscribed to at least one channel of that type. Off-platform referrers accounted for 52.4 and 46.6 percent of referrals to alternative and extremist videos, against 41.7 percent for mainstream media. Most tellingly, the authors defined a rabbit-hole event by four cumulative conditions and counted how often each was met. A participant following a recommendation to an alternative or extremist video occurred in 0.17 percent of video visits and among 7.3 percent of participants. Adding that the recommendation moved them to a more extreme channel type left 5.4 percent of participants; adding that they did not subscribe to the recommended channel left 4.7 percent; adding that they subscribed to no channel of that type left 3.0 percent, and 0.01 percent of visits. Sequences ending at an extremist video occurred in 0.002 percent of visits, among 1.0 percent of participants. Figure 1(a) re-presents the participant-level funnel. Chen et al. note that their counts should be read as a lower bound and that the susceptible population may already have been radicalised before 2019.

Counterfactual bots. Hosseinmardi et al. (2024) built the counterfactual that panels and audits each lack. Logged-in bots replayed the exact histories of real users from a desktop panel collected between October 2021 and December 2022; after a 60-video learning phase, one bot continued the real user's trajectory while others switched to following the top sidebar recommendation, a random sidebar item or a random homepage item. The difference between the two gives the user's own preference net of the recommender. Across 125 focal users, oversampled from the far-right tail (which made up only 1.12 percent of histories), the estimated preference term was positive for all three rules (0.029 for up next, 0.016 for random sidebar, 0.047 for random homepage, on their partisanship score), meaning that recommendation-following produced less partisan consumption than the users chose themselves. The gap grew with the partisanship of the focal user. When a far-

right history switched to moderate viewing, sidebar recommendations became indistinguishable from a centrist user's within about 30 videos, while homepage recommendations adjusted more slowly and retained a small far-right fraction. The authors restrict their conclusions to the post-2019 period, desktop users and an accelerated replay.

Amplification at the platform level. The two largest randomised platform experiments give a mixed picture that is easy to over-read. Twitter's long-running holdout of nearly two million daily active accounts on a reverse-chronological feed showed that the algorithm amplified the mainstream political right more than the left in six of seven countries, but found no evidence that it amplified far-left or far-right groups more than moderate ones (Huszar et al., 2022). On Facebook in 2020, ideological segregation rose as the analysis moved from content users could have seen, to content shown after ranking, to content they engaged with, and most fact-checked misinformation sat in a homogeneously conservative corner of the news ecosystem (Gonzalez-Bailon et al., 2023). The step from potential to actual exposure is algorithmic; the further step from exposure to engagement is the user's.

Concentration. The behavioral studies agree that exposure to problematic content is concentrated in a small tail. In the 2020 US election, 2,107 registered voters out of a panel of 664,391 accounted for 80 percent of the fake news shared on Twitter, and reached 5.2 percent of registered voters on the platform (Baribi-Bartov et al., 2024). Exposure to untrustworthy websites fell from 44.3 percent of American adults in 2016 to 26.2 percent in 2020, with older adults and conservatives still the most exposed and Facebook playing a smaller referral role than in 2016 (Moore et al., 2023). Budak et al. (2024) generalise the pattern into the claim that exposure is low on average and concentrated among people with strong motivations to seek it out. The deplatforming evidence points the same way: when Parler went offline in January 2021, active Parler users' probability of daily activity on other fringe platforms rose by 10.9 percentage points on desktop and 15.9 on mobile, and overall fringe activity did not fall (Horta Ribeiro et al., 2023). Demand relocated rather than disappeared.

What this establishes. For YouTube after 2019, every design that observes real users finds that consumption of partisan, alternative and extremist content (conspiratorial content has not been measured separately) is dominated by people who already sought it, reached by subscription, search and external link, with the recommender's causal contribution small and, in the one counterfactual design, negative on average. This is strong evidence against Link 2 in that setting. It is weak evidence about the tail, because the tail is a few dozen people in each sample, and it is no evidence about pre-2019 YouTube, about TikTok, or about mobile viewing, which Hosseinmardi et al. (2024) did not observe.

5. Link 3: escalation, and the problem of selecting on the outcome

Escalation is the part of the rabbit-hole claim most directly supported by data, and the part least often connected to its cause.

Migration between communities. Ribeiro et al. (2020) processed more than 72 million comments to track users across their four channel categories. The categories increasingly shared a user base, and users moved from milder to more extreme commenting over time. Of the 1,251,674 users who in 2017 commented only on Alt-lite or IDW videos, roughly 12 percent of those the authors could track went on to comment on Alt-right content, more than 60,000 people. Users who commented only on media channels did so far less often: under 1 percent reached the authors' mild or severe exposure levels, against 3 to 4 percent of Alt-lite or IDW commenters. Mamie, Horta Ribeiro and West (2021), with more than 300 million comments across 115 subreddits and 526 YouTube channels, found large user-base overlap between the Manosphere and the Alt-right and a higher chance of later Alt-right engagement than among matched counterparts, but with migration varying substantially by community and platform. Ota, Chujyo and Toriumi (2025) report for Japanese YouTube that channels framing conspiracy theories as entertainment or spirituality acted as gateways toward more conspiracy-laden content, and that monetised channels promoted conspiracy content most vigorously.

Joining and staying. Klein, Clutton and Dunn (2019) used a retrospective case-control design on Reddit and found that users who would later post in r/conspiracy differed from controls in language and in the communities they frequented before they ever posted there, which they read as self-selection into communities whose shared interests can feed a conspiratorial worldview. Phadke, Samory and Mitra (2020) matched 30K future members of 56 conspiracy subreddits with 30K non-members and found that dyadic interaction with existing members and marginalisation in other communities were the strongest social precursors of joining, outperforming individual-level baselines. After joining, the same group identified four engagement pathways among 36K users, steady high, increasing, decreasing and steady low, with only the first two progressing through successive radicalisation stages (Phadke et al., 2022). Samory and Mitra (2018), following r/conspiracy across ten years and four catastrophic events, distinguished joiners who arrived on Reddit after an event, converts who were already Reddit users, and veterans, and found that joiners were short-lived but intensely active.

Escalation without a recommender. Tschofenig et al. (2026) model escalation as interacting social contagions. Using 7.6 million tweets from 7,416 X users in the first wave of COVID-19, and language-model classification of public endorsement of 15 conspiracy narratives, they find that adopting one narrative raises the hazard of adopting others, that the elevation grows with the

number already adopted, and that transitions concentrate among semantically close narratives. Entry into a semantic region the user had not visited before occurs at roughly half the hazard of adoption within a region already entered, a pattern they call the settler effect. The model involves no recommender. Chandio and Nithyanand (2026), combining a year of YouTube watch history with two survey waves from 1,100 US participants, identify users whose ideology shifted toward extremes and find that their consumption differed from stable users and that the channels they favoured produced more anger and grievance content, while leaving open whether producers drove consumption or followed demand.

What this establishes, and what it does not. These studies document that some people's engagement with conspiratorial and extreme content deepens over time, that it deepens along semantic and community paths, and that social ties and marginalisation predict joining. They establish Link 3 in the populations studied. They do not establish that the recommender caused it, and most of them could not. Comment data records who commented, not who watched or what they were shown. Klein et al. (2019) show that future members differ before they join, which is the signature of selection rather than of exposure effects. Tschofenig et al. (2026) sample users who entered through two COVID-19 conspiracy hashtags, so their population is selected on prior engagement with conspiracy content; their hazards describe escalation among people already near the hole, not entry from outside it. Hosseinmardi et al. (2024) address the closest causal version of this question directly: bursts of far-right viewing predicted subsequent partisan consumption, but bots that followed the recommender after the same burst received less partisan content than the real users chose, so the burst signalled a change in the user's own preferences rather than a change in what the recommender served. Escalation is real. Its attribution to recommendation is, on current evidence, unsupported after 2019 and, before it, tested only observationally (Hosseinmardi et al., 2021), without a counterfactual.

6. Link 4: whether exposure changes belief

The link that gives the rabbit-hole claim its moral weight is the one least often measured.

Feed experiments. In the 2020 US election, moving consenting Facebook and Instagram users to reverse-chronological feeds sharply reduced their time on the platforms, increased the political and untrustworthy content they saw, and decreased uncivil content on Facebook, yet did not significantly change issue or affective polarisation, political knowledge or other key attitudes over three months (Guess et al., 2023). Two errata later corrected how Instagram exposure metrics were described and processed; the errata state that the treatment effect estimates were likely unaffected and that no Facebook results were affected (Science Erratum, 2024; Science Erratum, 2026). Reducing exposure to

like-minded sources by about a third for 23,377 Facebook users had no measurable effect on eight preregistered attitudes, including belief in false claims (Nyhan et al., 2023). Liu et al. (2025) manipulated real YouTube recommendations in a YouTube-like interface across four experiments with nearly 9,000 participants, run from mid-2021 to 2024. Slanted recommendations did shift what people chose to watch: averaged across three studies, liberal respondents chose about 6 percentage points more liberal videos and conservative respondents about 12 points fewer liberal videos, relative to balanced recommendations. But attitude effects were not consistently distinguishable from zero, in studies powered to detect effects of roughly 0.02 to 0.04 on a unit-scaled attitude measure. The authors argue that the burden of proof has shifted to claims of algorithmic polarisation, while noting that the short horizon limits what academic experiments can detect.

The counter-evidence. Not every exposure experiment returns a null. Piccardi et al. (2025) reranked the X feeds of 1,256 participants in real time during the 2024 US campaign, in a ten-day experiment whose reranking ran for seven days, using a language model to raise or lower exposure to posts expressing antidemocratic attitudes and partisan animosity, and found that the manipulation shifted out-party animosity by more than two points on a 100-point thermometer in either direction. Allen, Watts and Rand (2024), combining lab experiments with Facebook exposure data for 13,206 vaccine-related URLs, estimated that the impact of fact-checker-flagged misinformation on vaccination intentions was 46 times smaller than that of unflagged items that still pushed readers toward vaccine scepticism, because flagged content was rarely seen while factually accurate but potentially misleading stories, such as reports of rare deaths after vaccination, were among the most viewed. That result does not concern recommender drift, but it bears on Link 4 directly: exposure to belief-relevant content does move intentions, and the content that moves them most at scale is not the content a conspiracy classifier would flag.

Belief is movable. The most direct evidence that conspiracy belief responds to content comes from the opposite intervention. Costello, Pennycook and Rand (2024) engaged 2,190 conspiracy believers in tailored, evidence-based dialogues with GPT-4 Turbo and reported that belief fell by about 20 percent, persisted at two months, generalised across theories, and spilled over to unrelated conspiracies. The authors deposited a correction and reproducibility package for the article in September 2026 (Costello et al., 2026); this paper cites the original headline figure and does not assess what the correction changes. The relevance here is structural. If belief can be reduced by a few rounds of targeted counter-evidence, it is in principle movable by targeted supporting content, which is why the null results in the feed experiments say more about dose, horizon and content than about whether belief responds to information at all.

Observational correlates. Survey work places the association between platforms and conspiracy belief inside individual predisposition. Enders et al. (2023) find that social-media news use and frequent use are associated with some conspiracy and misinformation beliefs, but that the association strengthens as a respondent's conspiracy thinking intensifies, so the platform association is conditional on the disposition. In a two-wave survey across 17 countries before and after the onset of COVID-19, Theocharis et al. (2021) find that use of most platforms was associated with greater COVID-19 conspiracy belief while Twitter use was associated with less, which they attribute to differences in platform affordances. The psychology literature locates the demand in epistemic, existential and social motives (Douglas, Sutton & Cichocka, 2017), and a broad review attributes conspiracy belief to a combination of psychological, political and social factors (Douglas et al., 2019). At the population level, Uscinski et al. (2022) find no systematic increase in conspiracism across four studies, including 46 US conspiracy theories with observations in some cases spanning half a century, six European countries, and generalised conspiracy thinking in the US from 2012 to 2021. And conspiracy beliefs do not form a single dimension: Enders et al. (2021) find two uncorrelated dimensions, one partisan and one tied to anti-social orientations, with beliefs clustering by content.

What this establishes. Link 4 has been tested over horizons from a single viewing session with attitudes surveyed immediately afterwards (Liu et al., 2025, whose one-week follow-up wave was collected but not analysed), through a ten-day experiment (Piccardi et al., 2025), to three months (Guess et al., 2023; Nyhan et al., 2023), mostly on polarisation rather than on conspiracy belief, and mostly with null or small effects on attitudes despite large changes in exposure. One well-powered reranking study found a small effect on animosity. No experiment has varied recommender drift toward conspiracy content and measured conspiracy belief over a horizon of months or years. The absence of a secular rise in conspiracism over the period in which recommenders spread (Uscinski et al., 2022) is consistent with small effects, but an aggregate trend cannot rule out offsetting movements in subgroups.

7. Link 5: enclave or ordinary conversation

The Telegram study that prompted this analysis makes a claim about Link 5. Its evidence deserves a close reading, because it is the most recent and the most explicitly framed as a challenge to the radicalization account.

What the study did. Shingekar et al. (2025) fine-tuned RoBERTa-large on 2,000 messages labelled by two expert annotators, using a definition of a conspiracy theory as a narrative accusing agents of secret, malevolent action, and reached an F1 of 0.866 on held-out data. They applied the classifier to 553,648 messages from six Singapore-based Telegram groups, built a signed graph over messages in which edge signs encode whether two

messages share a conspiratorial label, learned embeddings with a signed graph network, and clustered them hierarchically into seven archetypes: legal topics, medical concerns, media discussions, banking and finance, contradictions in authority, group moderation, and general discussions. They report that conspiratorial messages appear across all seven archetypes, including mundane ones, and conclude that conspiratorial discourse operates within ordinary social interaction rather than in isolated echo chambers.

What the tables show. The study's Table 4 gives each archetype's message count and an estimated count of conspiratorial messages, totalling 106,931. Dividing one by the other, which the paper does not do, gives the share of each archetype that is conspiratorial. Those shares are 70 percent for contradictions in authority, 30 percent for medical concerns, 20 percent for banking and finance, 15 percent for general discussions, and 5 percent for each of legal topics, media discussions and group moderation, about 19 percent overall. Figure 1(b) re-presents them. The distribution is not uniform. One archetype carrying a tenth of the messages is seven-tenths conspiratorial, and three archetypes are one-twentieth. Read one way, conspiratorial content is everywhere; read the other, it is concentrated in a recognisable space, with a long tail of lower-rate occurrence elsewhere. Both readings are true of the same table. The study chooses the first.

Two features of the table limit how far either reading can go. First, each estimated count equals the archetype's message count multiplied by a round rate (5, 30, 5, 20, 70, 5 and 15 percent) to within one message. The caption calls them estimates, and the text does not say how they were produced; they are evidently not raw classifier counts. The shares are therefore best treated as approximate. Second, the per-group counts in the study's Table 1 show that 524,524 of the 553,648 messages, about 95 percent, come from one group. The finding that conspiratorial content is embedded in everyday talk is therefore, to a first approximation, a finding about one group. The six groups were described as having diverse themes; two of their names refer to COVID-19, and the study's introduction describes Telegram in Singapore during the pandemic as a space where users expressed scepticism toward institutional authority. The sample was not drawn to contain spaces unselected for conspiratorial content, and the paper does not report how often such content appears in groups outside its six.

The inference gap. Link 5 asks whether conspiratorial content is concentrated in sealed spaces. Evidence that a conspiracy-adjacent group also talks about shopping and bank rates is evidence that enclaves have ordinary conversation inside them. It is not evidence that ordinary spaces have conspiracy talk inside them, which would require the denominator the study does not have. Nor does it bear on radicalization in the sense of Links 1 to 4, since the design observes neither recommendations, nor trajectories of individuals, nor belief change. The study's limitations section notes that lurkers are invisible to text

analysis; this paper adds that the unsampled groups are invisible too.

Other evidence on location. The wider literature gives a mixed picture that depends on the metric. Engel et al. (2022) followed 13,000 users from early QAnon subreddits across all of Reddit from October 2016 to January 2021. These users participated in a diverse range of subreddits, often on unrelated topics, which looks like embedding; but most of their submissions were concentrated in subreddits sympathetic to the theory, which looks like enclaving. Breadth of participation and concentration of volume answer different questions. Cinelli et al. (2021), comparing more than 100 million pieces of content across Gab, Facebook, Reddit and Twitter, found that homophilic clusters dominated interactions on Facebook and Twitter, with higher segregation on Facebook than Reddit. Gonzalez-Bailon et al. (2023) found a substantial corner of the Facebook news ecosystem consumed exclusively by conservatives, where most fact-checked misinformation sat. On Telegram itself, a study of 4.4 million messages in 161 QAnon groups and channels found German overtaking English after 2020 and the discourse spanning world politics, COVID-19 and anti-vaccination topics alongside conspiracy theories (Hoseini et al., 2023), and a case study of 4,876 posts in one Irish extreme-right group found US conspiracies adapted to local concerns rather than adopted wholesale (Fattibene et al., 2024). Both describe enclaves with broad topical range, which is the same pattern the Singapore study reports, under a different name.

What this establishes. Conspiratorial content in conspiracy-adjacent spaces is mixed with ordinary talk; that is now documented on Reddit, Telegram and elsewhere. Whether conspiratorial content also appears at meaningful rates in spaces not selected for it, and whether the spaces that carry it are open or closed to the rest of the platform, has not been measured with a sampling frame that could answer it.

8. The period confound

One variable cuts across all five links and is often left out of comparisons between studies: when the data were collected.

YouTube announced on 25 January 2019 that it would reduce recommendations of borderline content and harmful misinformation, and later stated that watch time from such recommendations had fallen by 50 percent, and then that watch time from non-subscribed recommendations had fallen by 70 percent, figures Faddoul et al. (2020) describe as mostly consistent with their own measurements and Chen et al. (2023) report as not independently evaluated. The studies that find strong drift are disproportionately those with pre-2019 or mid-2019 data. Faddoul et al. (2020) observe the peak at the end of 2018 and the decline from April 2019. Ribeiro et al. (2020) collected recommendations in May to July 2019, as the change was being rolled out, and their migration data run through 2018. Hosseinmardi et al.

(2021) span 2016 to 2019 and still find no systematic recommendation-driven pattern, which is the strongest evidence that the pre-2019 effect was not large in real consumption either. The studies that find weak drift and preference-driven consumption are post-2019: Chen et al. (2023) in late 2020, Hosseinmardi et al. (2024) with 2021 to 2022 histories replayed in 2023, Liu et al. (2025) from 2021 to 2024. Each of these authors names the change as a limitation. Chen et al. (2023) add that the susceptible population may have been radicalised already. Hosseinmardi et al. (2024) state that recreating YouTube before the change is, to their knowledge, impossible.

Figure 2 places each study's data window against the January 2019 announcement. The pattern is clear enough that a reader comparing a 2019 audit with a 2023 counterfactual-bot experiment is comparing two different systems. That is not a criticism of either study. It means the field's central causal question about pre-2019 YouTube cannot now be answered with any design that requires the system to be run again, and the evidence about it is limited to what was archived at the time.

The demand side moved as well. Munger and Phillips (2022) report that viewership of far-right videos on YouTube peaked in 2017, before the recommender change. Exposure to untrustworthy websites fell from 2016 to 2020 (Moore et al., 2023). A period comparison therefore conflates changes in the recommender with changes in the supply of channels, in moderation, in user demand and in the wider media environment. Attributing the post-2019 decline to the recommender alone is no better supported than attributing the pre-2019 audience to it.

9. The evidence in two tables

Sections 3 to 7 lean on the numbers gathered in Table 1, one row per published result with its source on that row. Its "Link" column records which of the five links the author judges each design able to address; that column is this paper's judgement, and no source reports it. Table 2 maps designs to the links they can and cannot identify. Figure 1 re-presents two published distributions, and Figure 2 plots data windows as each source reports them.

Three things stand out from the tables. First, the rows that support the rabbit-hole claim are almost all L1 or L3 rows, and the rows that undercut it are almost all L2 or L4 rows; the two camps are, for the most part, measuring different things. Second, no row is a "Yes" for L2 and L4 together on conspiracy content specifically. Liu et al. (2025) come closest, on policy attitudes. Third, apart from Gonzalez-Bailon et al. (2023), whose platform-wide Facebook frame places fact-checked misinformation (not conspiratorial content as such) in a homogeneously conservative corner, and Cinelli et al. (2021), who compare interaction networks on controversial topics across four whole platforms, the L5 rows are discourse maps or user histories within spaces chosen for their content, which is why Table 2 marks every L5 entry as

Table 1. Published results on online conspiracy and extreme content, by link (author's reading). L1 drift, L2 uptake, L3 escalation, L4 effect, L5 location. Period is the data window as reported.

Source	Platform, period	Design	Link	Reported result
Faddoul et al. 2020	YouTube, Oct 2018-Feb 2020	Watch-next crawl, 8M + recs	L1	Conspiratorial recs peak ~10% raw, 6% weighted; low 3% (Jun 2019); weighted only 40% below announcement level by end
Ribeiro et al. 2020	YouTube, recs May-Jul 2019	Crawl + 72M comments	L1, L3	Alt-right reachable only via channel recs; ~12% of 2017 Alt-lite/IDW-only commenters later comment on Alt-right
Ledwich & Zaitsev 2020	YouTube	Anonymous crawl, ~800 channels	L1	Recs favour mainstream and cable news
Ledwich et al. 2022	YouTube	Personas	L1	Moderate bubbles; weak for Conspiracy/QAnon personas
Hussein et al. 2020	YouTube	Audit, 56K videos	L1	Watch history raises misinformation recs; demographics do not (new accounts)
Papadamou et al. 2022	YouTube	Audit, 6.6K videos	L1	More pseudoscience for old topics, on search page
Srba et al. 2022	YouTube	Sock puppets	L1	Bubbles burstable by debunking; no improvement vs prior audit
Haroon et al. 2023	YouTube	100,000 sock puppets	L1	Problematic-channel recs max 2.5%; seen by > 36.1% (40% very-right)
Brown et al. 2026	YouTube	Audit	L1	Content rabbit holes, no radicalization pathways
Gunasekara et al. 2026	TikTok	Audit, 2,893 videos	L1	More misinformation for new accounts; cross-topic gateway
Hosseinmardi et al. 2021	YouTube, 2016-2019	Panel, >300,000	L2	Far-right audience small, stable; no systematic rec effect
Chen et al. 2023	YouTube, Jul-Dec 2020	Browsing + recs, n=1,181	L2	1.7% / 0.6% of people = 80% of alt / extremist time; strict rabbit-hole event among 3.0%
Hosseinmardi et al. 2024	YouTube, 2021-22 histories	Counterfactual bots, 125 users	L2	Rec-following less partisan than users (alpha 0.016-0.047); sidebar forgets in ~30 videos
Robertson et al. 2023	Google Search, 2018, 2020	Exposure vs engagement	L2	Engagement more congruent and unreliable than exposure
Huszar et al. 2022	Twitter	Holdout, ~2M accounts	L1, L2	Mainstream right amplified (6 of 7 countries); far ends not
Gonzalez-Bailon et al. 2023	Facebook, 2020	208M users, aggregate	L2, L5	Segregation rises from potential to exposure to engagement
Baribi-Bartov et al. 2024	Twitter, 2020	Voter panel, 664,391	L2	2,107 voters share 80% of fake news
Moore et al. 2023	Web, 2016 vs 2020	Browsing, n=1,151	L2	Untrustworthy-site exposure 44.3% to 26.2%
Horta Ribeiro et al. 2023	Fringe sites, 2021	Panel, diff-in-diff	L2	Parler ban raised other-fringe activity 10.9 / 15.9 pp
Klein et al. 2019	Reddit	Case-control	L3	Future r/conspiracy posters differ before joining
Phadke et al. 2020	Reddit	30K matched pairs	L3	Dyadic ties, marginalisation predict joining
Phadke et al. 2022	Reddit	36K users	L3	Four pathways; two progress, one disengages
Tschofenig et al. 2026	X, 2020	7,416 users, hazard models	L3	Adoption raises later adoption; new region ~half hazard
Guess et al. 2023	Facebook, Instagram, 2020	Chronological-feed RCT	L4	Exposure changes; no significant attitude change
Nyhan et al. 2023	Facebook, 2020	RCT, 23,377 users	L4	Like-minded exposure cut ~1/3; no effect on 8 attitudes
Liu et al. 2025	YouTube-like, 2021-2024	4 RCTs, ~9,000	L4	Choice shifts 6-12 pp; attitudes null (MDE 0.02-0.04)
Piccardi et al. 2025	X, 2024	Reranking RCT, 1,256	L4	Animosity shifts >2 points (100-pt scale)
Allen et al. 2024	Facebook	Lab + exposure model	L4	Flagged misinformation impact 46x below unflagged sceptical content
Costello et al. 2024	Chat with GPT-4 Turbo	RCT, 2,190 believers	L4	Belief down ~20%, holds 2 months
Uscinski et al. 2022	US, Europe surveys	Trend analysis	L4	No systematic rise in conspiracism
Shimgekar et al. 2025	Telegram (SG), 2019-2025	Classifier + clustering	L5	Conspiratorial share 5-70% by archetype; ~95% of messages one group
Engel et al. 2022	Reddit, 2016-2021	13,000 QAnon users	L5	Broad participation; submissions concentrated in sympathetic subs
Cinelli et al. 2021	Gab, Facebook, Reddit, Twitter	100M+ items	L5	Homophilic clusters dominate on Facebook, Twitter

partial.

10. A link-attribution procedure

The map in Table 2 becomes useful when it is applied to a specific result before that result is cited for a general claim. Algorithm 1 states the procedure the author

proposes for doing so. It takes a published study and a claim someone wants to support with it, and returns the links the study can bear on, the period and population it covers, and the strongest conclusion it licenses. The procedure is deliberately mechanical: each step asks a question that can be answered from a methods section.

Table 2. Which link each design can identify (author's analysis). Yes: identifies the link under its stated assumptions. Part: speaks to it with a named gap. No: silent.

Design	L1 drift	L2 uptake	L3 escalation	L4 effect	L5 location
Sock-puppet or crawl audit	Yes	No	No	No	No
Browsing panel	No	Part: no recs seen	Part	No	No
Browsing with captured recs	Part	Yes, post hoc	Part	No	No
Counterfactual bots	Yes	Yes	Part	No	No
Comment or post trajectories	No	No	Yes, selected	No	Part
Feed or rec RCT	Part	Yes	No	Yes, short	No
Discourse map of chosen spaces	No	No	No	No	Part: no frame

Algorithm 1. Attributing a rabbit-hole result to a link Input: study S; claim K that S is cited for Output: links S can bear on; admissible conclusion for K

```

Step 1: Unit of observation.
  if S records what an agent is SHOWN under a fixed rule:
    candidate links <- {L1}
  if S records what real users CHOSE:
    candidate links <- candidate links + {L2-descriptive}
  if S records the same individuals over time:
    candidate links <- candidate links + {L3}
  if S randomises exposure and measures belief or attitude:
    candidate links <- candidate links + {L4}
  if S maps where content sits across spaces:
    candidate links <- candidate links + {L5}
Step 2: Counterfactual.
  if S compares real choice with rule-following on the same history,
  or randomises the recommender:
    upgrade L2-descriptive to L2-causal
  else: L2 stays descriptive; recommender share unidentified
Step 3: Selection.
  if S sampled users or spaces on prior conspiracy engagement:
    mark L3 and L5 as conditional on being inside already
    require an unselected comparison before any claim on entry
Step 4: Period and platform.
  record platform, device and data window; flag pre/post Jan 2019
  for YouTube; do not transfer across the flag or across platforms
Step 5: Horizon and outcome.
  for L4: record horizon and outcome measured
  if outcome is not conspiracy belief: mark L4 as adjacent only
Step 6: Admissibility.
  if K requires a link not in candidate links: K is unsupported by S
  else: state K restricted to the platform, period, population,
        horizon and outcome recorded in Steps 3-5

```

Applied to the four studies the public debate leans on most, the procedure gives the following. Ribeiro et al. (2020) supports L1 for YouTube in mid-2019 and L3 among commenters, with L3 unattributed. Chen et al. (2023) supports descriptive L2 for late-2020 YouTube users of desktop browsers, with a strict count of rabbit-hole events that the authors themselves describe as a lower bound. Hosseinmardi et al. (2024) supports causal L2 for 2021 to 2022 desktop histories replayed in 2023, oversampled from the partisan tail. Shingekar et al. (2025) supports partial L5 for six chosen Singapore groups, dominated by one, with no unselected comparison; it supports nothing on L1 to L4, so its framing as a challenge to radicalization accounts is admissible only for the version of those accounts that asserts enclave location.

The procedure has two uses beyond reading. For authors, it is a checklist for the scope sentence of a discussion section. For reviewers, Step 6 is the test of whether a framing sentence in an abstract has outrun the design.

11. Where the evidence complicates both stories

The popular account holds that recommendation algorithms pull ordinary users into conspiracy belief. The emerging counter-account holds that recommenders are a sideshow and conspiracy talk is simply part of ordinary life online. The evidence reviewed here complicates both.

Against the popular account. The strongest causal designs do not find an algorithmic push after 2019. Real users who sought far-right content chose more of it than the recommender would have served them (Hosseinmardi et al., 2024). Strict rabbit-hole events occurred among 3.0 percent of participants and 0.01 percent of visits (Chen et al., 2023). Slanted recommendations shifted choices but not attitudes, in experiments powered for small effects (Liu et al., 2025). Escalation is documented mainly among people already inside conspiracy spaces, who differed from others before they arrived (Klein et al., 2019; Phadke et al., 2020). And population-level conspiracism did not rise

over the period in which recommenders spread (Uscinski et al., 2022).

Against the counter-account. Four findings resist the conclusion that recommenders do not matter. First, drift is not zero even after 2019: problematic-channel recommendations reached over a third of sock puppets (Haroon et al., 2023), and the popularity-weighted share of conspiratorial recommendations rebounded after its 2019 low (Faddoul et al., 2020). Second, recommenders are not neutral pipes even when they moderate. They amplified the mainstream right on Twitter (Huszar et al., 2022), pushed long-running sessions toward moderately conservative content on YouTube (Brown et al., 2026), and served more misinformation to new TikTok accounts (Gunasekara et al., 2026). Third, the concentration that is usually cited to minimise the problem is also the reason it matters. Exposure to alternative and extremist content is concentrated among a small group with high prior resentment (Chen et al., 2023), the sharing of fake news among a few thousand voters (Baribi-Bartov et al., 2024), and recommendations to such content are disproportionately served to subscribers, who are the people most likely to follow them. A system that reinforces a committed tail and leaves everyone else alone is not radicalising the median user, but it may still be sustaining the tail. Fourth, Piccardi et al. (2025) show that feed ranking can move affective polarisation within ten days, which is a direct demonstration that ranking can have attitude effects when the manipulated content is chosen for its likely effect rather than for its ideology.

Against the "ordinary interaction" reading. The Telegram study's own table shows a 14-fold range in conspiratorial share across its archetypes, with the highest-rate archetype at 70 percent (Shingekar et al., 2025). Embedding and concentration are both present, and which one a reader sees depends on whether they count presence or rate. The QAnon Reddit data show the same split between breadth of participation and concentration of volume (Engel et al., 2022). The more defensible conclusion is that conspiracy-adjacent spaces are socially rich, which is consistent with the social-precursor findings of Phadke et al. (2020) and with the social-contagion model of Tschofenig et al. (2026). Socially rich enclaves may be more durable than sealed ones, not less, which is an argument for taking them more seriously, not for treating their content as ordinary.

The common ground. Both accounts can agree on the following. Demand drives most consumption after 2019. A small, motivated tail carries most exposure. Recommenders shape what that tail is served, and on some platforms what newcomers are served, in ways that audits can measure and that platform policy demonstrably changed. Social ties predict entry and escalation. Belief responds to content in both directions under some conditions, as the dialogue study shows. What neither account can yet claim is a measured, long-horizon causal effect of recommender drift on conspiracy belief.

12. What would settle the open part

Four questions remain open after the reconciliation. Each can be stated as a design.

The tail. The population that matters most for harm is the one that every representative sample captures in tens, not thousands. Hosseinmardi et al. (2024) show that oversampling the tail from a large panel makes counterfactual estimates feasible; their highest far-right stratum contained 17 histories. A design that recruits from the tail directly, through subscribers to channels on curated lists, and runs counterfactual bots on their histories with conspiracy-specific content labels, would estimate the recommender's causal share for the people the rabbit-hole claim is about. It would need conspiracy labels at the video level rather than the channel level, which Hosseinmardi et al. (2024) name as a limitation of channel-level scoring.

The horizon. No experiment has run long enough to test cumulative effects. Liu et al. (2025) note that their short-term perturbations set practical limits on detectable effect sizes. A field design that randomises a mild, persistent recommender adjustment over months, using the reranking infrastructure of Piccardi et al. (2025) and measuring conspiracy belief rather than polarisation, would test Link 4 at the horizon the claim implies. The dialogue study's two-month follow-up (Costello et al., 2024) shows that durable belief change can be measured.

Platforms other than YouTube. The causal designs are almost all on YouTube. TikTok audits disagree with each other about misinformation drift (Gunasekara et al., 2026; Pecher et al., 2026), and no TikTok study pairs recommendations shown with real choices. Telegram has no recommender in the YouTube sense, so its rabbit holes, if they exist, run through forwarding and group invitations, a social channel closer to the contagion model of Tschofenig et al. (2026) than to algorithmic drift. Each platform needs its own Link 2 study.

Location with a frame. A location claim needs a sampling frame that includes spaces not chosen for conspiratorial content. On Reddit, where the full set of communities is enumerable, a study could estimate the rate of conspiratorial content in a random sample of communities and compare it with the rate in conspiracy-adjacent ones, using the same classifier. On Telegram, where public groups can be discovered by forwarding chains, a snowball frame starting from non-political groups would provide the missing denominator. Either would turn the Singapore finding from a description of six groups into a test of the enclave hypothesis.

The pre-2019 counterfactual. This one cannot be settled by new data collection. The archived crawls of Faddoul et al. (2020) and Ribeiro et al. (2020) are the evidence that exists about what the system served, and the 2016 to 2019 panel of Hosseinmardi et al. (2021) is the evidence about what people chose. A joint re-analysis of the two, aligning recommendation archives with panel viewing over the same months, is the closest

available approximation to a counterfactual for that period.

13. The strongest case against this paper

The strongest objection is that the five-link decomposition is an artefact of the analyst's taxonomy and that the field already knows it. Brown et al. (2026) separate echo chambers, content rabbit holes and radicalization pathways; Hosseinmardi et al. (2024) explicitly contrast what audits and panels can identify; Ribeiro, Veselovsky and West (2023) explain why audits and user traces diverge; Budak et al. (2024) separate exposure from effects. On this view the paper re-describes known distinctions. The reply is that the existing distinctions are each partial. Brown et al. subdivide Link 1; Hosseinmardi et al. separate Links 1 and 2; Budak et al. separate Links 2 and 4. None places escalation and location in the same frame as drift, uptake and effect, and none is applied to discourse-map studies like the Telegram paper, whose framing as a challenge to radicalization accounts is exactly the kind of cross-link inference the decomposition is meant to catch. Whether that is a sufficient contribution is a fair question for a reader to judge.

A second objection is that the synthesis compares incommensurable categories. "Problematic", "alternative", "extremist", "far-right", "conspiratorial" and "untrustworthy" are defined differently in every source, and Table 1 sets them side by side. That is correct, and the table does not average across rows or treat any two values as measuring the same quantity. The conclusions drawn from it are about which link each design can identify, which does not depend on the content category.

A third objection is that the reading of the Telegram study is unfair to a paper whose main contribution is methodological. Its signed graph network and clustering method are its central contribution, and nothing here bears on them. The critique is confined to the substantive framing in its abstract and conclusion, and the numbers used come from the paper's own tables.

14. What is not known

The causal effect of recommender drift on conspiracy belief has not been measured over any horizon. The recommender's causal share of consumption among the heaviest consumers of conspiratorial content specifically, as opposed to partisan content, has not been measured on any platform. Whether pre-2019 YouTube drove the audience that later studies observe as already committed is unknown and, by the account of those studies' authors, cannot now be tested directly. The rate of conspiratorial content in spaces not selected for it is unknown for Telegram and has not been estimated with a common classifier for any platform. No study pairs recommendations shown with real choices on TikTok.

15. Limitations

The first limitation is the one a hostile reviewer would raise first: the evidence base is dominated by YouTube and by US samples, and the conclusions about Link 2 are much stronger for post-2019 YouTube desktop use than for anything else. The paper says so where each conclusion is drawn, but the reader should discount the general claims accordingly. Second, the link assignments in Tables 1 and 2 are the author's reading of each design and are not values the sources report; a different reader could assign some studies differently, particularly the browsing-with-recommendations design, which this paper treats as identifying L2 post hoc rather than causally. Third, several sources are recent and some are preprints (Tschofenig et al., 2026; Shimgekar et al., 2025; Phadke et al., 2020 in its arXiv form); their numbers may change. Fourth, the per-archetype shares in Figure 1(b) are computed from counts the source labels as estimates, and the paper treats them as approximate. Fifth, the paper uses abstracts rather than full texts for several sources whose role is a single qualitative claim, and its review is not systematic: it does not apply a PRISMA-style protocol, and the 23 studies in Yesilada and Lewandowsky (2022) and the 496 in Lorenz-Spreen et al. (2022) are drawn on only through their summary findings. Sixth, the Costello et al. (2024) figure is the originally published one; a correction package exists and was not assessed.

16. Conclusion

The rabbit-hole metaphor joins five claims that the literature tests separately: that recommenders drift toward conspiratorial content, that real consumption is caused by that drift, that individuals escalate, that exposure changes belief, and that the content lives in sealed spaces. Each research design identifies at most two of them. Once results are sorted by link and by the January 2019 period break, the literature's apparent contradictions largely resolve. Drift was measurable before 2019 and is weak and channel-specific after it on YouTube; uptake is dominated by prior demand in every dataset that can see it; escalation is real but documented among people already inside; attitude effects of large exposure changes are small over short horizons; and claims about location turn on a sampling frame no study of conspiratorial content yet has. The Telegram study that prompted this analysis is evidence about the inside of conspiracy-adjacent spaces, not about whether they are enclaves or how people arrive in them. What would settle the remaining question is not another audit or another discourse map, but a long-horizon causal design aimed at the committed tail, run on more than one platform.

References

- Allen, J., Watts, D. J., & Rand, D. G. (2024). Quantifying the impact of misinformation and vaccine-skeptical content on Facebook. *Science*, 384(6699), eadk3451. doi:10.1126/science.adk3451.
- Baribi-Bartov, S., Swire-Thompson, B., & Grinberg, N. (2024). Supersharers of fake news on Twitter. *Science*, 384(6699),

- 979-982. doi:10.1126/science.adl4435.
- Bouchaud, P., & Ramaciotti Morales, P. (2024). Auditing the audits: evaluating methodologies for social media recommender system audits. *Applied Network Science*, 9(1). doi:10.1007/s41109-024-00668-6.
- Brown, M. A., Bisbee, J., Lai, A., Bonneau, R., Nagler, J., & Tucker, J. A. (2026). Evaluating Echo Chambers, Rabbit Holes, and Radicalization Pathways on YouTube. *Political Communication*. doi:10.1080/10584609.2026.2671765.
- Budak, C., Nyhan, B., Rothschild, D. M., Thorson, E., & Watts, D. J. (2024). Misunderstanding the harms of online misinformation. *Nature*, 630(8015), 45-53. doi:10.1038/s41586-024-07417-w.
- Chandio, S., & Nithyanand, R. (2026). Examining the Role of YouTube Production and Consumption Dynamics on the Formation of Extreme Ideologies. *Proceedings of the International AAAI Conference on Web and Social Media*, 20(1), 407-420. doi:10.1609/icwsm.v20i1.42645.
- Chen, A. Y., Nyhan, B., Reifler, J., Robertson, R. E., & Wilson, C. (2023). Subscriptions and external links help drive resentful users to alternative and extremist YouTube channels. *Science Advances*, 9(35), eadd8080. doi:10.1126/sciadv.add8080.
- Cinelli, M., De Francisci Morales, G., Galeazzi, A., Quattrociocchi, W., & Starnini, M. (2021). The echo chamber effect on social media. *Proceedings of the National Academy of Sciences*, 118(9). doi:10.1073/pnas.2023301118.
- Costello, T. H., Pennycook, G., & Rand, D. G. (2024). Durably reducing conspiracy beliefs through dialogues with AI. *Science*, 385(6714), eadq1814. doi:10.1126/science.adq1814.
- Costello, T. H., Pennycook, G., & Rand, D. G. (2026). Durably reducing conspiracy beliefs through dialogues with AI: Correction and reproducibility materials. *Zenodo*. doi:10.5281/zenodo.22754983.
- Douglas, K. M., Sutton, R. M., & Cichocka, A. (2017). The Psychology of Conspiracy Theories. *Current Directions in Psychological Science*, 26(6), 538-542. doi:10.1177/0963721417718261.
- Douglas, K. M., Uscinski, J. E., Sutton, R. M., Cichocka, A., Nefes, T., Ang, C. S., & Deravi, F. (2019). Understanding Conspiracy Theories. *Political Psychology*, 40(S1), 3-35. doi:10.1111/pops.12568.
- Enders, A. M., Uscinski, J. E., Klostad, C., Seelig, M. I., Wuchty, S., Murthi, M. N., Premaratne, K., & Funchion, J. (2021). Do conspiracy beliefs form a belief system? Examining the structure and organization of conspiracy beliefs. *Journal of Social and Political Psychology*, 9(1), 255-271. doi:10.5964/jsp.5649.
- Enders, A. M., Uscinski, J. E., Seelig, M. I., Klostad, C. A., Wuchty, S., Funchion, J. R., Murthi, M. N., Premaratne, K., & Stoler, J. (2023). The Relationship Between Social Media Use and Beliefs in Conspiracy Theories and Misinformation. *Political Behavior*, 45(2), 781-804. doi:10.1007/s11109-021-09734-6.
- Engel, K., Hua, Y., Zeng, T., & Naaman, M. (2022). Characterizing Reddit Participation of Users Who Engage in the QAnon Conspiracy Theories. *Proceedings of the ACM on Human-Computer Interaction*, 6(CSCW1), 1-22. doi:10.1145/3512900.
- Faddoul, M., Chaslot, G., & Farid, H. (2020). A Longitudinal Analysis of YouTube's Promotion of Conspiracy Videos. *arXiv:2003.03318*. doi:10.48550/arXiv.2003.03318.
- Fattibene, G., Windle, J., Lynch, O., Helm, G., Purvis, J., & Seppa, L. (2024). The online exchange of conspiracy theories within an Irish extreme right wing Telegram group during the COVID-19 pandemic. *Behavioral Sciences of Terrorism and Political Aggression*. doi:10.1080/19434472.2024.2409185.
- Gonzalez-Bailon, S., Lazer, D., Barbera, P., Zhang, M., Allcott, H., Brown, T., et al. (2023). Asymmetric ideological segregation in exposure to political news on Facebook. *Science*, 381(6656), 392-398. doi:10.1126/science.ade7138.
- Guess, A. M., Malhotra, N., Pan, J., Barbera, P., Allcott, H., Brown, T., et al. (2023). How do social media feed algorithms affect attitudes and behavior in an election campaign? *Science*, 381(6656), 398-404. doi:10.1126/science.abp9364.
- Gunasekara, S., Pareek, S., Kelly, R., Kostakos, V., & Goncalves, J. (2026). Auditing TikTok's recommendations: Understanding the prevalence of misinformation on the platform. *International Journal of Human-Computer Studies*, 218, 103929. doi:10.1016/j.ijhcs.2026.103929.
- Harooun, M., Wojcieszak, M., Chhabra, A., Liu, X., Mohapatra, P., & Shafiq, Z. (2023). Auditing YouTube's recommendation system for ideologically congenial, extreme, and problematic recommendations. *Proceedings of the National Academy of Sciences*, 120(50), e2213020120. doi:10.1073/pnas.2213020120.
- Hoseini, M., Melo, P., Benevenuto, F., Feldmann, A., & Zannettou, S. (2023). On the Globalization of the QAnon Conspiracy Theory Through Telegram. *Proceedings of the ACM Web Science Conference 2023*, 75-85. doi:10.1145/3578503.3583603.
- Horta Ribeiro, M., Hosseinmardi, H., West, R., & Watts, D. J. (2023). Deplatforming did not decrease Parler users' activity on fringe social media. *PNAS Nexus*, 2(3), pgad035. doi:10.1093/pnasnexus/pgad035.
- Hosseinmardi, H., Ghasemian, A., Clauset, A., Mobius, M., Rothschild, D. M., & Watts, D. J. (2021). Examining the consumption of radical content on YouTube. *Proceedings of the National Academy of Sciences*, 118(32). doi:10.1073/pnas.2101967118.
- Hosseinmardi, H., Ghasemian, A., Rivera-Lanas, M., Horta Ribeiro, M., West, R., & Watts, D. J. (2024). Causally estimating the effect of YouTube's recommender system using counterfactual bots. *Proceedings of the National Academy of Sciences*, 121(8), e2313377121. doi:10.1073/pnas.2313377121.
- Hussein, E., Juneja, P., & Mitra, T. (2020). Measuring Misinformation in Video Search Platforms: An Audit Study on YouTube. *Proceedings of the ACM on Human-Computer Interaction*, 4(CSCW1), 1-27. doi:10.1145/3392854.
- Huszar, F., Ktena, S. I., O'Brien, C., Belli, L., Schlaikjer, A., & Hardt, M. (2022). Algorithmic amplification of politics on Twitter. *Proceedings of the National Academy of Sciences*, 119(1). doi:10.1073/pnas.2025334119.
- Klein, C., Clutton, P., & Dunn, A. G. (2019). Pathways to conspiracy: The social and linguistic precursors of involvement in Reddit's conspiracy theory forum. *PLoS ONE*, 14(11), e0225098. doi:10.1371/journal.pone.0225098.
- Ledwich, M., & Zaitsev, A. (2020). Algorithmic extremism: Examining YouTube's rabbit hole of radicalization. *First Monday*, 25(3). doi:10.5210/fm.v25i3.10419.
- Ledwich, M., Zaitsev, A., & Laukemper, A. (2022). Radical bubbles on YouTube? Revisiting algorithmic extremism with personalised recommendations. *First Monday*, 27(12). doi:10.5210/fm.v27i12.12552.
- Liu, N., Hu, X. E., Savas, Y., Baum, M. A., Berinsky, A. J., Chaney, A. J. B., et al. (2025). Short-term exposure to filter-bubble recommendation systems has limited polarization effects: Naturalistic experiments on YouTube. *Proceedings of the National Academy of Sciences*, 122(8), e2318127122. doi:10.1073/pnas.2318127122.
- Lorenz-Spreen, P., Oswald, L., Lewandowsky, S., & Hertwig, R. (2022). A systematic review of worldwide causal and correlational evidence on digital media and democracy. *Nature Human Behaviour*, 7(1), 74-101. doi:10.1038/s41562-022-01460-1.
- Mamie, R., Horta Ribeiro, M., & West, R. (2021). Are Anti-Feminist Communities Gateways to the Far Right? Evidence from Reddit and YouTube. *Proceedings of the 13th ACM Web Science Conference 2021*, 139-147. doi:10.1145/3447535.3462504.
- Moore, R. C., Dahlke, R., & Hancock, J. T. (2023). Exposure to untrustworthy websites in the 2020 US election. *Nature Human Behaviour*, 7(7), 1096-1105. doi:10.1038/s41562-023-01564-2.
- Munger, K., & Phillips, J. (2022). Right-Wing YouTube: A Supply and Demand Perspective. *The International Journal of Press/Politics*, 27(1), 186-219. doi:10.1177/1940161220964767.

- Nyhan, B., Settle, J., Thorson, E., Wojcieszak, M., Barbera, P., Chen, A. Y., et al. (2023). Like-minded sources on Facebook are prevalent but not polarizing. *Nature*, 620(7972), 137-144. doi:10.1038/s41586-023-06297-w.
- Ota, K., Chujyo, M., & Toriumi, F. (2025). Pathways to Conspiracy Theorizing on YouTube in Japan. *Proceedings of the International AAAI Conference on Web and Social Media*, 19, 2635-2641. doi:10.1609/icwsm.v19i1.35965.
- Papadamou, K., Zannettou, S., Blackburn, J., De Cristofaro, E., Stringhini, G., & Sirivianos, M. (2022). It Is Just a Flu: Assessing the Effect of Watch History on YouTube's Pseudoscientific Video Recommendations. *Proceedings of the International AAAI Conference on Web and Social Media*, 16, 723-734. doi:10.1609/icwsm.v16i1.19329.
- Pecher, B., Bindas, A., Jakubcik, J., Tuna, M., Tibensky, M., Liska, S., et al. (2026). Algorithmic Audit of Personalisation Drift in Polarising Topics on TikTok. *Proceedings of the ACM Conference on User Modeling, Adaptation and Personalization (UMAP 2026)*. doi:10.1145/3774935.3806161.
- Phadke, S., Samory, M., & Mitra, T. (2020). What Makes People Join Conspiracy Communities?: Role of Social Factors in Conspiracy Engagement. *arXiv:2009.04527*. doi:10.48550/arXiv.2009.04527.
- Phadke, S., Samory, M., & Mitra, T. (2022). Pathways through Conspiracy: The Evolution of Conspiracy Radicalization through Engagement in Online Conspiracy Discussions. *Proceedings of the International AAAI Conference on Web and Social Media*, 16, 770-781. doi:10.1609/icwsm.v16i1.19333.
- Piccardi, T., Saveski, M., Jia, C., Hancock, J. T., Tsai, J. L., & Bernstein, M. S. (2025). Reranking partisan animosity in algorithmic social media feeds alters affective polarization. *Science*, 390(6776), eadu5584. doi:10.1126/science.adu5584.
- Ribeiro, M. H., Ottoni, R., West, R., Almeida, V. A. F., & Meira, W. (2020). Auditing radicalization pathways on YouTube. *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, 131-141. doi:10.1145/3351095.3372879.
- Ribeiro, M. H., Veselovsky, V., & West, R. (2023). The Amplification Paradox in Recommender Systems. *Proceedings of the International AAAI Conference on Web and Social Media*, 17, 1138-1142. doi:10.1609/icwsm.v17i1.22223.
- Robertson, R. E., Green, J., Ruck, D. J., Ognyanova, K., Wilson, C., & Lazer, D. (2023). Users choose to engage with more partisan news than they are exposed to on Google Search. *Nature*, 618(7964), 342-348. doi:10.1038/s41586-023-06078-5.
- Roth, C., Mazieres, A., & Menezes, T. (2020). Tubes and bubbles: topological confinement of YouTube recommendations. *PLoS ONE*, 15(4), e0231703. doi:10.1371/journal.pone.0231703.
- Samory, M., & Mitra, T. (2018). Conspiracies Online: User Discussions in a Conspiracy Community Following Dramatic Events. *Proceedings of the International AAAI Conference on Web and Social Media*, 12(1). doi:10.1609/icwsm.v12i1.15039.
- Science Erratum. (2024). Erratum for the Research Article "How do social media feed algorithms affect attitudes and behavior in an election campaign?" by A. M. Guess et al. *Science*, 386(6726), eadu8261. doi:10.1126/science.adu8261.
- Science Erratum. (2026). Erratum for the Research Article "How do social media feed algorithms affect attitudes and behavior in an election campaign?". *Science*, 391(6791), eah2575. doi:10.1126/science.eah2575.
- Shingekar, S. R., Goyal, A., Lee, R. K.-W., Saha, K., Zonooz, P., & Kumar, N. (2025). Belief Is All You Need: Modeling Narrative Archetypes in Conspiratorial Discourse. *arXiv:2512.10105*. doi:10.48550/arXiv.2512.10105.
- Srba, I., Moro, R., Tomlein, M., Pecher, B., Simko, J., Stefancova, E., et al. (2022). Auditing YouTube's Recommendation Algorithm for Misinformation Filter Bubbles. *ACM Transactions on Recommender Systems*, 1(1), 1-33. doi:10.1145/3568392.
- Theocharis, Y., Cardenal, A., Jin, S., Aalberg, T., Hopmann, D. N., Stromback, J., et al. (2021). Does the platform matter? Social media and COVID-19 conspiracy theory beliefs in 17 countries. *New Media & Society*, 25(12), 3412-3437. doi:10.1177/14614448211045666.
- Tomlein, M., Pecher, B., Simko, J., Srba, I., Moro, R., Stefancova, E., et al. (2021). An Audit of Misinformation Filter Bubbles on YouTube: Bubble Bursting and Recent Behavior Changes. *Proceedings of the Fifteenth ACM Conference on Recommender Systems*, 1-11. doi:10.1145/3460231.3474241.
- Tschofenig, F., Vicinanza, P. J., Greve, H. R., Rao, H., & Guilbeault, D. (2026). Conspiracy Theory Rabbit Holes Emerge via Interacting Contagions. *arXiv:2607.19131*. doi:10.48550/arXiv.2607.19131.
- Uscinski, J. E., Enders, A. M., Klostad, C., Seelig, M. I., Drochon, H., Premaratne, K., & Murthi, M. N. (2022). Have beliefs in conspiracy theories increased over time? *PLoS ONE*, 17(7), e0270429. doi:10.1371/journal.pone.0270429.
- Yesilada, M., & Lewandowsky, S. (2022). Systematic review: YouTube recommendations and problematic content. *Internet Policy Review*, 11(1). doi:10.14763/2022.1.1652.