nature computational science

Review article

https://doi.org/10.1038/s43588-026-01038-1

Algorithmic systems, human agency and the future of platform research

Received: 10 March 2026

Accepted: 14 July 2026

Published online: 21 September 2026

Check for updates

Homa Hosseinmardi[1] ✉, Upasana Dutta[2], David Rothschild[3] & Duncan J. Watts[2,4,5] ✉

As the volume of online content grows exponentially, algorithmic curation has become necessary for delivering content to billions of users daily, such as via recommendation and ranking algorithms. However, there are widespread concerns that these algorithms amplify problematic content or prioritize content that reinforces users' existing beliefs. Although legitimate, these concerns have overshadowed the role that users' own preferences play in determining these same outcomes. Here we argue for a more balanced approach to understanding online content consumption and engagement that considers both algorithmic mechanisms and human agency. We review the existing literature on the role of algorithmic recommendation and users' self-selection in the patterns of users' online diets and show that evidence linking algorithmic curation to problematic outcomes is inconclusive. We highlight the gaps in the existing literature and call for the need for more large-scale data-driven empirical research and experimentation, with the goal of refining the debate on the societal impact of digital platforms and providing a path for future research.

Algorithmic content curation is central to the functioning of modern digital platforms, determining what content users see and do not see online, as well as the order in which they see it. Leveraging user data such as search histories, viewing behaviors, subscriptions, social connections and demographics, algorithmic systems are trained to predict and recommend content tailored to individual users. At a high level, these algorithms serve two broad purposes: the first is to rank the content users encounter on homepages, search results or news feeds, and the second is to recommend new content to which users have not yet been exposed[1,2]. Because ranking and recommendation systems are credited with improving users' experiences, they have become ubiquitous, shaping how billions of users interact with online environments daily.

Accordingly, a large scholarly literature has emerged on the individual and societal consequences of these algorithms[3–11]. The complexity and opacity of machine-learning algorithms make it difficult for users, regulators and even developers to understand their functionality and impact[12]. Combined with the vast scale and diversity of content on modern platforms, this lack of transparency has prompted a wide range of concerns to be raised both about the technical features of recommendation systems[13–15] and also the intentions and business models of the companies that own them[16–18]:

- Meta, X (formerly Twitter) and Alphabet faced scrutiny for facilitating the spread of Russian misinformation during the 2016 US presidential election, prompting questions about their influence on electoral outcomes[19,20].
- YouTube has been described as 'one of the most powerful radicalizing instruments of the 21st century', due to its alleged exposure of users to ideologically extreme and conspiracy-laden content[21–23].
- Google has been criticized for perpetuating racial and gender stereotypes through biases embedded in its search results[7].
- Facebook has been linked to declines in subjective well-being[11], while Instagram has been accused of promoting bullying and undermining the mental health of teenage girls[24–26].

[1]Department of Communications, University of California Los Angeles, Los Angeles, CA, USA. [2]Department of Computer and Information Science, University of Pennsylvania, Philadelphia, PA, USA. [3]Microsoft Research New York, New York, NY, USA. [4]Annenberg School of Communication, University of Pennsylvania, Philadelphia, PA, USA. [5]Operations, Information, and Decisions Department, University of Pennsylvania, Philadelphia, PA, USA. ✉e-mail: homahm@ucla.edu; djwatts@engineering.upenn.edu

https://doi.org/10.1038/s43588-026-01038-1

- X has been cited for propagating hate speech and racist sentiments, potentially inciting real-world violence against minority communities[6].
- The engagement-maximizing design of social media platforms—such as autoplay, infinite scroll and algorithmic personalization optimized for retention—has come under regulatory and legal scrutiny, with European regulators classifying such features as systemic risks under the Digital Services Act [27] and US lawsuits alleging that algorithmic recommendation systems have contributed to addictive use and mental health harms among minors[28].
- Social media and search algorithms in general have been accused of creating echo chambers[3–5,29–31], also often referred to as 'filter bubbles'[32], in which individuals are steered toward like-minded people or sources that reinforce existing beliefs, potentially exacerbating political polarization[33–36] and undermining support for democracy[37–41].

Together, these critiques highlight widespread concerns about the potential negative societal impacts of algorithm-driven systems. In this Review we categorize these concerns into two distinct but related sets of problems. First, algorithmic systems may intentionally or unintentionally promote 'problematic content'—defined as misleading, harmful or otherwise undesirable material[42–44]. For example, inflammatory or polarizing videos on YouTube may receive disproportionate algorithmic promotion simply because they retain viewers' attention longer than more balanced alternatives. Similarly, on platforms like X, content generating high engagement—such as sensationalized narratives or controversial opinions—may be amplified, potentially biasing public discourse and elevating divisive voices. Exposure to problematic content, however, should not be conflated with its effects. While exposure can be measured through browsing or recommendation data, downstream effects such as radicalization or attitude change require separate evidence and may vary substantially across individuals. Second, even when the content itself is not inherently problematic, its selective presentation can still lead to adverse societal effects. For example, if users predominantly engage with articles that align with their existing beliefs, algorithms may amplify such content to optimize engagement, leading to the aforementioned echo chambers that limit exposure to diverse viewpoints[8–10,45,46]. Exacerbating this problem, users in such echo chambers may make progressively biased selections, which then further inform subsequent algorithmic recommendations, creating a reinforcing cycle[47,48] in which individuals are driven toward increasingly homogeneous and potentially polarizing equilibria[49,50].

Scenarios such as these are legitimately concerning. In focusing so much on the role that algorithms play in causing these problems, however, the public discourse has frequently glossed over another potential cause—namely, the role that users themselves play in shaping their online experiences. Individuals often initiate search or browsing sessions with clear intentions about the content they wish to discover. Moreover, having entered a session, users do not respond deterministically to algorithmic recommendations but instead select among available offerings in ways that also reflect their preferences, interests, offline experiences, social connections and even some degree of randomness. Finally, platforms are highly incentivized to learn the preferences of their users in order to more effectively engage them; thus, even when users do engage with recommended content, the reason may be that it resembles what they would have sought out anyway. At the same time, the observed user preferences should not be interpreted as fixed or exogenous. Preferences themselves may be shaped by previous platform experiences, social environments and historical algorithmic exposure. For all these reasons, although recommendation systems undoubtedly influence users' behavior, the extent to which observed outcomes derive from algorithms versus humans remains deeply unclear[45,49,51,52].

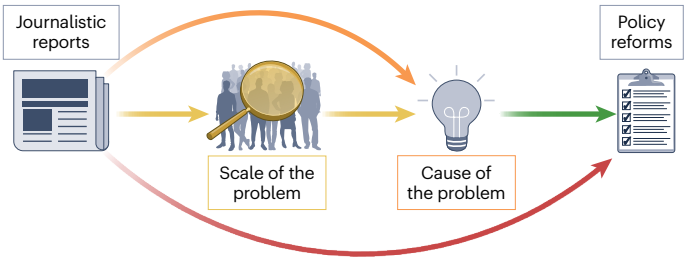

Fig. 1 | Proposed interventions must both reflect the scale of the problem and target the appropriate causal mechanism. Although journalistic reports highlight issues in the online space, a concerning trend is the rush to policy reforms without fully understanding the scale or causes of the problem (bottom pathway). Similarly, focusing solely on causes without assessing the prevalence of the issue or the affected demographics is equally problematic (middle pathway). We argue for a structured approach: first, conducting high-quality descriptive studies to determine the prevalence of issues, followed by resource allocation to investigate all potential causes (top pathway). Only then should mechanism-targeting interventions, algorithmic changes, or policy reforms be implemented. Image courtesy of Amir Ghasemian.

Clarifying this issue requires robust empirical evidence that addresses these two fundamental dimensions of the problem. First, it is important to establish the prevalence of problematic outcomes in real-world online environments. For example, what share of the population actually 'live' in filter bubbles, and to what degree are their information environments homogeneous? Similarly, empirical work must assess the extent of low-quality content consumption across user populations and determine what fraction of users' overall media diets such content represents. Second, when problematic patterns are indeed observed, empirical studies must identify their primary causal mechanisms. For example, in cases where users appear constrained within filter bubbles, research must disentangle the relative contributions of algorithmic curation versus deliberate user selection behaviors.

Both dimensions are important. Should filter bubbles prove neither widespread nor persistent, even effective remedies will have little effect on most users' experiences; thus, resources may be better allocated to more pervasive digital harms. Conversely, if filter bubbles are either widespread or else can be shown to cause significant harm even among smaller groups, intervention may be justified. To be effective, however, proposed interventions must target the appropriate causal mechanism: whereas algorithmically driven patterns might respond to technological solutions, those arising from individual preferences or broader social dynamics would likely require more comprehensive interventions that address underlying behavioral and societal factors (Fig. 1).

In the remainder of this Review, we first synthesize existing evidence on the prevalence and causes of the most widely discussed problematic outcomes associated with online platforms (Table 1). After drawing some preliminary conclusions from these studies, we then outline some important limitations of current research, such as inconsistent methodologies, contradictory findings, and variations in temporal and contextual scope, which collectively impede definitive answers. In the final section, we propose future research directions and address the inherent challenges, alongside potential solutions, for studying human–algorithm interactions at scale.

Current state of understanding

Prevalence of problematic outcomes

Public discourse surrounding algorithmic harms has been heavily influenced by journalistic accounts and popular books arguing that filter bubbles[30,32], platform-mediated radicalization[21,53], and user exposure to misleading or offensive content[54,55] constitute urgent societal problems. However, much of the evidence presented in these accounts is

Table 1 | Representative evidence, approach and findings in research on the social impact of digital platforms

Evidence type	Approach	Main contribution	Common limitations	Support
Anecdotal/journalistic	Media narratives and whistleblower accounts[21,37,191]	Surface plausible risks; generate public awareness and policy attention toward algorithmic harms	Non-representative and speculative; lack of systematic sampling, baselines or counterfactuals; overgeneralize from extreme cases	Problematic outcomes are ubiquitous and are largely driven by algorithms
Descriptive/empirical	Large-scale analyses of exposure and engagement patterns using panel or trace data[63,65,71,74]	Quantify prevalence and concentration of problematic or partisan content	Correlational; limited ability to isolate algorithmic effects from human selection; measurement error and access constraints	Problematic outcomes are less prevalent than expected
Causal/experimental	Feed-randomization, counterfactual-bot and field experiments on recommendation systems[43,92,101]	Disentangle algorithmic influence from user behavior	Short-term or context-bound; cannot capture long-run feedback effects	Problematic outcomes are largely driven by user engagement

anecdotal and impressionistic rather than systematic and empirical. For example, the early prominent concerns over echo chambers[56] were entirely theoretical, while the concept of filter bubbles[32] was based largely on the author's personal observations with search results, drawing criticism for its lack of systematic validation[57,58]. Similarly, prominent reports documenting YouTube promoting problematic videos[21,22] or Facebook's amplification of harmful content have often relied on individual experiences or employee testimonials[59] rather than comprehensive behavioral data. Illustrative of this trend, a 2019 *New York Times Magazine* cover article[53] focused on a single person who had reportedly been radicalized via YouTube. Subsequently, a 2022 article in *The New York Times*[55] confidently asserted that the widespread amplification of misinformation by social media is well-known, but cited for its primary evidence an online report that itself relied on a small, biased sample of posts[60].

Articles such as these are highly visible and widely read and cited, and so give the impression of consensus around the problematic role of algorithms in society[61]. However, scientific evidence on the prevalence of problematic outcomes on online platforms—which tends to accumulate more slowly and receive much less attention—is either mixed or points to a much lower prevalence than popular accounts imply[62]. For example, consumption of 'fake news' on online platforms has been found to be rare on average, comprising less than 1% of overall news consumption and less than one-tenth of one percent of overall media consumption[63]. Correspondingly, although approximately 40% of US adults visited at least one untrustworthy news website during the final weeks of the 2016 election, such sites represented only 5.9% of their total news consumption[64]. Furthermore, analysis of YouTube browsing patterns from 2016 to 2019 found that far-left and far-right communities collectively comprised merely 0.05% of American users—less than one tenth the size of mainstream political communities[65]. These findings underscore that although absolute counts of exposure to problematic content might seem substantial, their prevalence is minimal when contextualized against the vast scale of total platform activity and user engagement[62].

Moreover, these population averages are inflated by the activity of a small, highly engaged subset of users who are actively seeking out the content[14,65–70]. During the 2016 US election, for example, ~1% of X users accounted for 70% of all Russian disinformation exposure[67], while an identical proportion of individuals consumed 80% of fake news sources[68]. Similarly, fewer than 1% of US citizens consumed 85% of vaccine-skeptical content between 2016 and 2019[69]. This concentration pattern extends to extremist content, where 0.6% of YouTube users were responsible for 80% of total viewing time on extremist channels[14], with far-right community members exhibiting more than double the monthly watch time of centrist users[65]. Finally, this pattern of highly concentrated consumption among a small fraction of users is also evident in general web news consumption[66]. In other words, population averages of the consumption of problematic content, although low, probably overstate the extent of the problem for typical users. That being said, we emphasize that rarity does not necessarily imply inconsequentiality. The prevalence of problematic content consumption and the societal importance of its consequences are distinct quantities that should be evaluated separately. Even when exposure is concentrated among a small fraction of users, the resulting harms may be substantial. Our point is therefore not that low prevalence implies low importance, but rather that assessments of platform harms should distinguish between the scale of a phenomenon and the magnitude of its consequences. Measuring one should not be taken as evidence for, or against, the other.

A similar theme has emerged in empirical research on echo chambers: although online consumption of ideologically congruent content is common on many platforms, it only dominates consumption for a small minority of users and is often more diverse than in other contexts[71]. For example, a large-scale collaborative investigation between Meta and independent researchers identified some degree of ideological segregation in political news consumption[72] and demonstrated users' greater likelihood of encountering politically aligned versus cross-cutting content[73]. However, the collaboration also found that extreme echo chamber patterns remained relatively uncommon: only 20.6% of Facebook users received more than 75% of their news exposure from ideologically similar sources, a rate that is comparable to the size of partisan echo chambers in cable TV news consumption[74]. Such results suggest that informational isolation affects only a minority of platform users and occurs similarly with and without algorithmic mediation.

Similarly, recent work combining nationally representative surveys with digital trace data reveals substantial cross-partisan exposure across platforms. On X, ideological distributions of content encountered by users at opposing ends of the political spectrum exhibit 51% overlap[75], while analyses of web browsing histories demonstrate considerable convergence in uniform resource locator (URL) visits between Democratic and Republican users[76]. Large-scale panel studies further challenge the echo chamber narrative: only 4% of online news consumers exhibit strong partisan segregation[74] and, crucially, ideological segregation in digital news consumption remains significantly lower than that observed in face-to-face social interactions[66]. Finally, empirical evidence offers little support for the 'filter bubble' hypothesis either in Google Web search[77–79] or Google News[80].

Thus, despite evidence of concerning consumption patterns among specific subpopulations[12], empirical research has challenged the prevailing narrative that online platforms inherently or generically foster echo chambers, filter bubbles or widespread exposure to harmful content[71,81,82]. Rather than constraining information diets, in fact, search engines and social media platforms may function as serendipitous discovery mechanisms, exposing users to ideologically diverse content they would be unlikely to encounter through traditional media channels or their social networks[83,84]. As Guess et al.[71] conclude, 'critics who decry "echo chambers", "filter bubbles" and "information cocoons" [...] overstate the prevalence and severity of

these patterns, which at most capture the experience of a minority of the public'. This disconnect between popular narrative and empirical reality suggests that extreme information segregation, while arguably problematic where it occurs, characterizes the experience of only a small user subset, with the majority encountering relatively diverse and mainstream content through algorithmic curation.

A further limitation of the evidence reviewed is the concentration of the existing literature on online content consumption, misinformation exposure and political segregation in the United States (and other Western democracies). Media systems, regulatory environments, platform penetration and levels of media literacy may differ substantially across countries, and consequently, the prevalence and drivers of problematic outcomes may not generalize beyond the country under study. Additionally, this Review focuses primarily on a small number of platforms, with limited treatment of platforms such as WhatsApp, Telegram or non-social-media recommendation systems such as news aggregators and streaming platforms, where the supply–demand balance may operate differently. Similarly, because the evidence synthesized in this Review focuses primarily on political information environments, the conclusions drawn should not be assumed to extend to other forms of platform harm, such as eating-disorder content, self-harm content or financial scams.

Causes of problematic outcomes

Although the evidence just surveyed suggests that echo chambers and harmful content are less prevalent than commonly portrayed, the same evidence shows that such concerns do apply to at least some users; thus, understanding the mechanisms driving these phenomena remains an important problem. Do platform algorithms actively steer users toward polarized or harmful content either by design or as an unintended consequence of optimizing for other goals (for instance, maximizing engagement)? Alternately, do these outcomes primarily reflect consumers' intrinsic preferences, on the one hand, and the interests of content producers, on the other? As already noted, this question cannot be resolved simply by observing users' engagement patterns, which almost certainly conflate all these elements. Complicating matters further, content producers, consumers and platforms all affect one another via dynamic feedback loops. User preferences shape algorithmic recommendations, which influence content production, which in turn affects both future user behavior and algorithmic learning. Disentangling these causal pathways is essential for developing targeted interventions that address the root causes rather than the symptoms of information polarization and the proliferation of harmful content (Fig. 2).

As with concerns about the pervasiveness of false or problematic content and echo chambers, concerns about the algorithmic origins of these problems have arisen largely out of theoretical arguments and individual cases rather than out of systematic empirical evidence. In theory, the argument is that in seeking to maximize profits, companies also seek to maximize user engagement. In turn, they deploy algorithms that promote sensationalist, controversial or extreme material, all of which are believed to boost engagement either with the content itself or between users. If this argument is correct, it follows that algorithms are inherently geared to exacerbate polarization and the spread of misinformation. Not only are arguments of this sort plausible; it is also easy to find individual examples that appear to support them. For example, in 2025 Substack sent a push alert for a Nazi newsletter[85] to many of its users, including users who had never expressed any interest in Nazi ideology. Although Substack later claimed the alert was sent in error, it is easy to believe that such events are more likely to occur in systems that value engagement for its own sake.

Events of this sort often receive considerable press coverage, creating the impression that they are highly prevalent. As with misinformation on social media platforms, however, systematic empirical evidence presents a more complicated and often inconsistent picture. On YouTube, for example, research employing sock-puppet audits—automated accounts simulating user behavior—frequently reports that YouTube's recommendation system creates self-reinforcing pathways toward misinformation and pseudoscientific content[86–90]. A recent audit study on TikTok similarly identifies systematic differences in political exposure generated by recommendation systems[91]. However, this pattern does not necessarily indicate that the recommendation system is 'amplifying' problematic content so much as showing users content similar to that with which they have previously engaged from the universe of available content (supply). Indeed, explicitly causal analyses in which logged-in bots that mimic 'organic' user behavior are contrasted with 'counterfactual bots' that click exclusively on algorithmic recommendations find that the latter tend to steer users with extreme tastes to more moderate content, whereas for moderate users the recommendations are essentially neutral[92]. Moreover, exposure to extremist channels occurs almost exclusively among users who have already subscribed to such sources[14] and who seek it out across multiple platforms[65].

Evidence from X tells a similar story. On the one hand, sock-puppet studies suggest that X's algorithms amplify politically focused content clusters, potentially intensifying partisan divides[93]. On the other hand, a randomized experiment involving 58 million X users found no evidence of algorithmic amplification favoring political extremes over moderate content, though the same study did find systematic amplification of right-leaning over left-leaning content across most countries studied[94]. Indeed, algorithmic curation on X appears to improve information quality: users exposed to algorithmic timelines encounter less ideologically congruent, less extreme, and more reliable news sources compared to reverse-chronological feeds[43]. Furthermore, large-scale studies[42] point to the role of user behavior in shaping online engagement patterns, especially low-quality content. An analysis of 10 million posts across seven social media platforms revealed that posts linking to lower-quality news sources systematically received higher engagement than those linking to higher-quality sources. This pattern holds across both algorithmically curated and chronologically ordered platforms, suggesting that higher engagement of such low-quality content cannot be attributed to algorithmic amplification alone, but also reflects users' active engagement choices.

Facebook studies further illuminate how the interplay between algorithmic curation, individual choice and network structure shapes exposure to cross-cutting political viewpoints. Although some early studies found that algorithms reduce cross-cutting political content, these studies also found that users actively choosing like-minded sources exerted substantially greater influence on ideological homogeneity[95–97]. Subsequent work has tried to isolate the causal effect of recommendation algorithms on users' exposure and political attitudes via experimental designs. Most prominently, the US 2020 Facebook and Instagram Election Study employed this strategy in several large-scale interventions. Users randomly assigned to reverse-chronological feeds for three months reduced platform usage but showed no measurable changes in affective or issue polarization, political knowledge or offline political behavior compared to those using algorithmic feeds[98]. Similarly, removing reshared content substantially decreased platform engagement without affecting polarization, institutional trust or perceptions of political division[15]. Even directly downranking like-minded content to increase cross-cutting exposure failed to alter users' political attitudes, affective polarization or susceptibility to election-related misinformation[73]. In contrast, a recent independent field experiment on X found that switching users from a chronological to an algorithmic feed for seven weeks shifted political attitudes in a more conservative direction, including on policy priorities and views on current events, without affecting affective polarization or partisanship[99]. The effect was asymmetric: switching the algorithm on produced measurable opinion change, but switching it off did not reverse it, as users continued following the conservative

https://doi.org/10.1038/s43588-026-01038-1

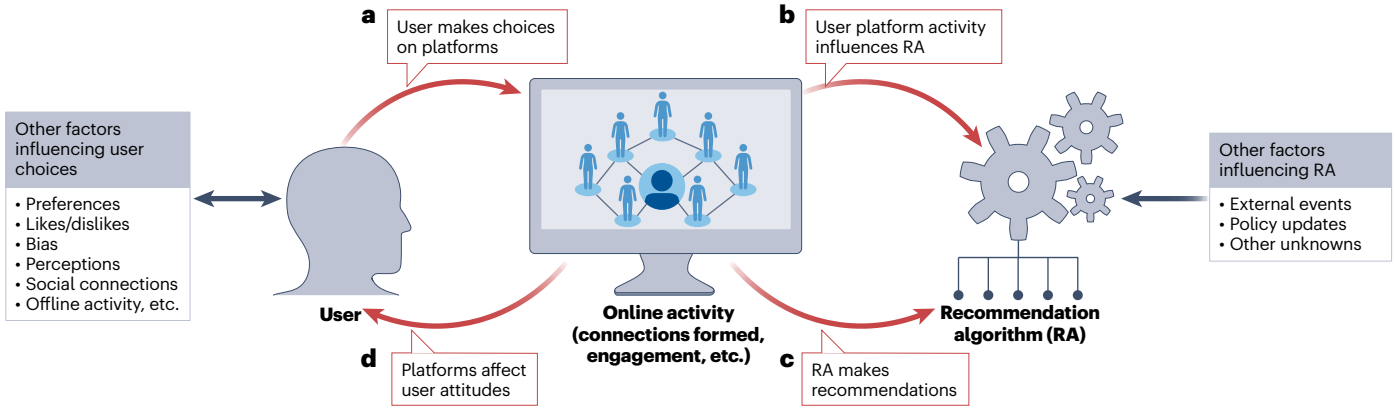

Fig. 2 | Feedback loops between user agency and algorithmic curation. **a**, User characteristics—including intrinsic preferences, cognitive biases and offline social networks—shape platform engagement patterns through content selection, interaction behaviors and response patterns. **b**, These behavioral signals constitute the primary input for recommendation algorithms. **c**, Algorithms process these signals to generate personalized content recommendations, thereby shaping users' information exposure, social connections and engagement trajectories. **d**, Cumulative algorithmic influence potentially modifies users' attitudes, beliefs and offline behaviors over time, which subsequently alter their online preferences and choices, perpetuating the feedback cycle. Although substantial research has examined algorithmic influence on user behavior (pathway **d**), the role of user agency in initiating and perpetuating these dynamics (pathway **a**) remains critically understudied, representing a substantial gap in our understanding of human–algorithm interaction.

activist accounts the algorithm had led them to discover. This asymmetry helps reconcile the apparent contradiction with the Meta findings: switching off an algorithm may show no effect not because the algorithm was inconsequential, but because its influence has already been absorbed into users' organic feed.

Further, critics of the 2020 Election Study[100] have pointed out that it was conducted during a period when Meta was taking other actions to reduce the prevalence of problematic content and hence its findings are unlikely to generalize beyond this specific period. These concerns are indeed valid and reinforce the necessity of multiple studies before drawing sweeping conclusions. We note, however, that the conventional wisdom prior to the Election Study that Meta's algorithms were causing polarization was not based on unbiased generalizable data either. Imperfect as they are, therefore, these null effects are still the best evidence available for Meta platforms. Also, although this evidence does not show that algorithms are irrelevant, consistent with the evidence from other social media platforms just cited, it indicates that the effects are in part driven by user preferences within a platform-designed environment. Moreover, results similar to those observed on YouTube, X and Facebook have been found to apply beyond social media platforms to the web in general. For example, research on Google Search that carefully distinguishes between algorithmic exposure (what users see) and user engagement (what users choose to interact with) finds that partisan bias in search results is minimal, but that users' engagement, conditional on exposure, with partisan and unreliable sources is highly correlated with their political identities[101].

Finally, we note, that null effects in individual-level experiments do not necessarily imply the absence of broader societal effects. As is argued in ref. 102, complex social systems are characterized by feedback loops, emergent dynamics, and interference across individuals, making it difficult to infer system-level consequences from individual-level estimates alone. We therefore interpret these results as evidence about direct exposure effects on individual users, rather than as evidence against population- or system-level effects that the study was not designed to detect. Beyond the limitations of individual-level study designs, algorithmic effects need not operate solely through direct exposure to problematic content. Platform design may shape societal outcomes indirectly by influencing the incentives faced by content producers, affecting which content receives attention and visibility, and interacting with users through feedback loops that unfold over time. For example, politicians, news outlets and influencers may adapt their messaging to maximize algorithmic reach, potentially increasing the production of sensationalist, emotionally charged, or otherwise attention-grabbing content. Algorithms may also shape the information environment more broadly by determining which topics, narratives and voices receive prominence, thereby influencing public attention and perceptions, even among users who are not directly exposed to problematic content. And, cumulative effects on norms, salience and perceived social consensus may emerge gradually over long time horizons and therefore be difficult to detect in short-term individual-level studies. Consequently, the absence of large direct exposure effects should not be interpreted as conclusive evidence that algorithms are irrelevant to broader societal outcomes. Understanding these indirect and system-level pathways remains an important direction for future research.

Summarizing, there now exists a considerable body of evidence across multiple platforms that challenges prevailing narratives of algorithmic determinism. Instead, it supports a more balanced view of problematic online content that has been called the 'supply and demand' framework[17,18,103]. According to this view, platforms—whether social media platforms or search engines—act as 'markets' that match users' demand for content with content producers' willingness to supply it. Importantly, the platform is not a neutral intermediary but an active architect of it. By setting optimization objectives and designing the choice architecture, it contributes to shaping both how content is supplied and how user demand is generated. Consequently, treating demand as one of the drivers of consumption does not imply that demand is exogenous to the platform. As we have pointed out, this matching process is necessarily mediated by ranking and recommendation algorithms, and hence their role merits scrutiny along with other factors. However, the current evidence suggests that many observed consumption patterns cannot be explained by algorithmic amplification alone and often reflect substantial contributions from user preferences and content supply. Yet, because user preferences and algorithmic exposure are entangled through long-term feedback loops, observational evidence alone cannot determine how far today's preferences were themselves shaped by prior algorithmic exposure. If anything, the evidence is that they nudge users towards more politically neutral material than they would otherwise seek[104]. Similarly, the supply

and demand framework allows for the possibility that echo chambers exist, but argues that they again primarily reflect the preferences of users rather than algorithms—a point that we again note is supported by the comparable prevalence of echo chambers on television news viewing, a medium in which algorithmic filtering is absent[74].

We emphasize that the supply and demand framework does not seek to minimize the problems with which critics of social media platforms and 'Big Tech' are concerned, nor to absolve platforms of responsibility to address these problems. Clearly, problematic and highly partisan content does exist on most, if not all, platforms, and at least some users engage with one or both types of content heavily, potentially with negative societal consequences. Moreover, in such an environment, platforms are not powerless to act, any more than market designers anywhere are powerless to act. Although their responsibility does not hinge on algorithms alone, they can shape the environment in which both supply and demand operate, and thus cannot be treated as neutral intermediaries. For example, all platforms already prohibit certain types of content, removing it automatically and penalizing the content creators for posting it. In theory, they can also effectively remove content by 'shadow banning', a practice in which platforms deliberately down-rank to a degree that it does not appear to users unless specifically requested. Finally, platforms can influence content producers by tweaking financial incentives, altering recommendation algorithm parameters to boost or penalize certain types of content, or even banning users outright. As we will discuss later, these remedies are all imperfect and can generate blowback from users as well as other unintended consequences. This added complexity does not diminish platform responsibility, but does suggest that effective interventions depend on accurately identifying the mechanisms involved. For now, the important point to note is that the supply and demand framework posits a fundamentally different origin for problematic consumption patterns than the prevailing algorithmic account—namely, that not everybody agrees that it is problematic.

Conceptual and computational obstacles to advancing the field

As we have just discussed, a small but growing body of empirical research has in recent years addressed widespread concerns about the role of algorithms in causing problematic online content consumption. Although we believe this work has added much needed context and nuance to preexisting arguments based largely on theory and anecdotes, we emphasize that our understanding of these complex issues remains incomplete and somewhat fragmentary. Addressing these gaps is critical for advancing the field in a rigorous and meaningful direction, but will require substantial conceptual and computational challenges to be overcome. In the remainder of this section, we briefly describe some of these challenges, though we note that this list is also incomplete.

Conceptual ambiguity of key definitions

Perhaps one of the most fundamental challenges is the lack of consensus around, and clarity of, basic terminology. For example, the boundaries of what constitutes 'problematic' content—whether misinformation, bias or extremity—are rarely specified explicitly and hence are often ambiguous[105–108]. Similarly, constructs like 'filter bubble' and 'echo chamber' are inconsistently defined, or not defined at all, leading studies to adopt various measurement criteria, which in turn produce inconsistent results[81,82,109,110]. To illustrate, some empirical work finds that although there is evidence of ideological segregation, it is accompanied by cross-partisan exposure, raising the question of at what point one can determine the presence of a filter bubble[57]. Similarly, labeling content as extreme or partisan is inherently subjective and often influenced by the annotator's perspective[111,112]. Given that the most salient patterns of problematic content consumption tend to arise in the long tail—driven by small subsets of users and outlets—these definitional ambiguities can have outsized effects on measurement and interpretation. There is no 'correct' answer, but transparency and clarity are necessary to compare different research findings.

Inadequate measurement practices

A second source of conceptual and computational difficulty is that standard practices for measuring quantities of interest—whether on the production side (for example, partisanship, factuality or extremity) or the consumption side (for instance, amount of news consumed, dominant sources of information and so on)—fail to capture the underlying construct. On the production side, this problem arises most obviously in the practice of classifying content at the level of an entire domain (for example, publisher, channel owner, TV program) rather than at the content level itself[42,65,83,101]. Scholars often utilize third-party ratings (for instance, Media Bias/Fact Check, AllSides), which assign scores to entire publications based on editorial practices, audience perceptions or expert review[5]. Alternately, scholars may infer source attributes from the political composition of their audiences—for example, labeling domains by the partisan affiliations of users who share them[20,77,83]. Although convenient, both methods of assigning domain-level labels necessarily overlook heterogeneity of content within a single source[113]. For many sources, not all content from the same source exhibits the same partisan bias or is even the same type of content[114,115]: for example, many news publishers also publish a large amount of entertainment and lifestyle content, while in other cases straight news and editorial content can differ substantially in partisanship. Likewise, there is no guarantee that the partisanship of the consumer base mirrors the bias of the content they consume[113], especially—as is often the case—partisanship is inferred from geographic and demographic metadata rather than elicited directly.

Because either of these sources of conflation could either increase or decrease estimates of problematic content consumption, depending on the specifics, it is unclear how current findings would shift if content were annotated at a more granular level, such as articles[15,20,64,77,101], segments[74] or videos[65,116]. Regardless, it is likely that any such changes would impact downstream inferences regarding the kinds of macro-social phenomenon that we have discussed above[117]. Ideally, therefore, characterizations of content should be grounded in the content itself, not who is producing or consuming it. Fortunately, the advent of large language models (LLMs) promises to alleviate some of these technical barriers; however, transitioning to comprehensive content-level analysis also introduces new complexities, necessitating operations at an unprecedented scale where researchers must process and evaluate millions of individual content items rather than thousands of sources.

On the consumption side, a parallel limitation arises from the common practice of using self-reported data to quantify user consumption patterns, particularly in news. As has been widely discussed, self-reported data are highly susceptible to recall bias and social desirability bias, among other errors[118]—which, in the case of news consumption, often yields large overestimates of respondents' actual usage and exposure[119,120]. Comparative studies have shown stark disparities. For instance, self-reported news exposure can be inflated by a factor of three on average, and up to eightfold for certain demographics, when compared to digital tracking data[121]. These methodological shortcomings are particularly concerning for examining incidental or passive news consumption[84,122], as consistent evidence points to substantial discrepancies between reported and actual media consumption[123].

Inadequate validation

Social scientists in recent years have adopted various computational tools to study digital content (for instance, detecting fake news or measuring extremity)[124,125]. However, researchers often overlook the need to validate these algorithms on the specific domain to which they are being applied, which is often different from the applications for which the methods were originally developed[126,127]. For example, methods such

as the structural topic model (STM) or latent Dirichlet allocation (LDA), which have been widely used in social[128], communication[125] and political science[129–131], are frequently employed without adequate scrutiny of their fitness for the specific inferential tasks at hand[117]. Although these models are powerful for extracting insights from large corpora, their application to questions involving, for example, selection bias or the representation of extreme political content, demands rigorous assessment of their error rates[126,130] and how they handle long-tail content. Conventional metrics of model fit, such as topic coherence or maximum likelihood, are often insufficient to validate interpretations of substantive social scientific interest[117,132].

Recently, advances in transformer methods have redirected attention from conventional tools to LLMs and other generative AI tools, both for measuring features such as topic[133], partisanship[134] or sentiment[135], as well as for more complex applications such as developing virtual agents or predicting election outcomes[124,136]. Although this new generation of methods is demonstrably superior to the previous generation for many applications, the same principles of rigorous validation apply. When employing techniques such as zero-shot or few-shot learning with LLMs for inference, assuming error-free labels—even in cases of high reported accuracy—can lead to biased and invalid conclusions[137]. Therefore, comprehensive and context-aware validation procedures remain essential to ensure the scientific integrity and interpretability of findings derived from these powerful new tools. Moreover, because many of the findings based on earlier methods continue to influence current thinking, the shortcomings of these methods remain relevant even as the methods themselves have become outdated.

Limited access to platforms

A notable barrier to understanding online consumption dynamics stems from the limited and in some cases diminishing access to comprehensive platform data. In most fields, only a small number of external researchers have internal access to platform data, typically through one-off academic collaborations[95,138], which greatly facilitates large-scale observation of user behavior in natural settings, real-time monitoring of algorithm–user interactions, and controlled A/B testing[15,72,73,98]. In contrast, most studies of platforms rely on data collected from application programming interfaces (API)[104,139], which are usually limited both in the type and quantity of data that can be requested. For example, data available through X API cannot answer many user-level questions, such as what a user has clicked on, or recommended to them in their feed[140], and is limited instead to, for instance, likes or reshares, and so on. Unfortunately, API access to platform data has, if anything, become more restricted over time, in some cases because newer platforms (for example, TikTok) have less researcher-friendly policies than older platforms, and in other cases because platforms (for example, X) have proactively changed their policies to become less researcher-friendly[141–143]. Another set of works relies on scraping public data[116], which comes with other sets of risks and complications, such as violating the terms and service of platforms.

Because of limited data, researchers have been forced to rely on some combination of: small sample sizes[43,80,93,144,145], which is a problem as problematic content tends to cluster in a small fraction of users[62]; user watch histories that reflect the combined effect of user preferences and algorithmic features[65,116]; data donations from platform users[146,147], which, while capturing real browsing behavior, suffer from self-selection bias and limited sample representativeness; or, synthetically scraped social media platforms using 'sock puppets'[148], ignoring the personalized recommendations that real users receive based on their past browsing behavior[93,104] and the subscriptions they have opted into consuming[14,86]. A related and increasingly promising approach uses browser extensions to observe[14,101] or experimentally modify users' real online experiences[43]. These methods can enable platform-independent field experiments that manipulate content exposure while preserving many aspects of naturalistic platform use, but they remain subject to many of the same limitations as data-donation approaches.

Cross-platform limitations

Most existing studies are limited to a single platform under study (such as Facebook[95] or YouTube[104]), often due to reliance on an API or internal access[149], and therefore fail to account for the users' off-platform consumption, which could paint a different picture of the overall consumption pattern. For instance, one study of the effectiveness of moderation policies found that shutting down the right-wing social media site Parler had little impact on individuals' consumption of conspiratorial content: users simply replaced their diets of such content via other sources on the web[150]. Leveraging cross-platform consumptions, researchers found that banning Instagram in China pushed users towards using tools to bypass restrictions and end up not only accessing this platform, but also many more, such as Twitter, Facebook and blocked political pages on Wikipedia[151]. Yet another study found that users who consume far-right content on YouTube also consume it directly on far-right websites[65], a finding that has subsequently been replicated on data from a different panel[14]. Finally, a recent study examining seven platforms[42] found that despite differences in moderation policies, algorithms, user demographics and political preferences, users consistently exhibited higher engagement with lower-quality content. Although these findings are not causal, they point to a broader, systematic pattern in user behavior rather than platform-specific effects.

Feedback loop effect

A final challenge is that recommendation systems and user preferences form feedback loops, where recommendations affect the content that users consume, and can impact their future preferences, which in turn is used as feedback for making subsequent recommendations to the user[48,152,153]. Importantly, such loops imply that observed preferences should not be treated as exogenous to the algorithm itself: what appears as user demand may in part reflect the preferences that earlier cycles of recommendation has contributed in shaping. As a result, demand should be interpreted only as a description of current consumption patterns rather than a claim about how those preferences were historically formed. These feedback loops introduce considerable additional complexity because the effects of algorithmic recommendations or user preferences may become evident only in the long run after several cycles of feedback[47,49,154,155], and are therefore not captured by studies that only span a short period of time. Thus far, a handful of studies have attempted to explore these dynamics by simulation, and in doing so have demonstrated the theoretical potential of algorithms to create such loops[46,47] or to design algorithms that ameliorate them[48,156]. As we have noted earlier, however, the theoretical possibility of a problem—however plausible it may seem—does not necessarily imply empirical importance. Entirely new types of empirical study and research design are therefore needed to establish the actual prevalence and causal effects of feedback loops between algorithmic systems and their users.

Recommendations

Thus far, we have argued that the current empirical literature does not provide strong support for claims that algorithms are the dominant driver of many commonly discussed social and political outcomes. Instead, existing evidence suggests that user preferences, content supply and algorithmic curation all contribute to observed consumption patterns. As we have also pointed out, however, the empirical literature is not in a state of high consensus on this important question and is in any case far from conclusive, in part due to the challenges just highlighted. In this section, therefore, we offer some recommendations for future research that we believe would, if adopted, improve both our collective conception of the problem as well as our empirical understanding of it (Box 1).

BOX 1

Key recommendations for platform research

Prioritize problem scale before investigating causes

- Establish prevalence through descriptive studies before pursuing intervention mechanisms.
- Ensure research focuses on problems that exist at a meaningful scale.
- Example: cross-platform analysis.

Match research design to research question

- Clarify the intended audience (policymakers, public or platform owners).
- Design studies appropriate for stakeholder needs.
- Limit research questions to the power of the design and data.
- Example: optimizing for feedback loop.

Pursue causal identification through complementary approaches

- Clearly define the content of interest, the comparison baseline, the exposure outcome, the role of user behavior, and the time horizon over which effects are expected to stabilize.
- Combine observational analyses, controlled experiments, and synthetic-user or exposure-engagement designs to separate algorithmic supply from user demand and capture feedback over time.
- Provide careful methodological validation.

Interpret findings carefully and transparently

- Restrict conclusions to the studied populations and time periods.
- Emphasize meaning alongside statistical significance.
- Ensure complete transparency and replicability of methods to build consensus.

Address platform and data-access challenges

- Rebalance collaboration toward independent, parallel replications rather than large, single-team designs.
- Secure sustainable, timely researcher access to platform data and APIs to facilitate longitudinal audits.
- Ameliorate cross-platform limitations.

Incorporate economic and behavioral perspectives

- Examine platforms as two-sided markets shaped by supply and demand.
- Reevaluate engagement metrics by empirically testing when observed behavior diverges from users' long-term welfare.
- Define measurable dimensions of societal value and develop empirical frameworks for integrating them into algorithm evaluation and optimization.

Broadening the scope of inquiry

At the highest level, we urge researchers to formulate the problem using a broader and more integrated framing. Rather than adopting an excessively narrow focus on either recommendation algorithms or user preferences, researchers should holistically consider the complex interplay between three forces: (1) the increasing ease with which online content can be produced today (with producers steering their content towards what consumers may want to consume), (2) the convenience of discovering niche content on online platforms due to the efficiency of their recommendation algorithms, and (3) the intrinsic human tendency to seek like-minded information and the discomfort dealing with the unfamiliar[157,158].

Understanding how institutional, cultural and regulatory conditions shape the interplay between content producers, users and algorithmic systems is also an understudied area. The relative importance of these forces may vary substantially across contexts characterized by different media systems, levels of platform adoption, media literacy, regulatory environments and availability of alternate information sources. Future work should therefore prioritize comparative and cross-national designs that can identify how platform effects, and the balance between user demand, content supply and algorithmic curation, vary across institutional and cultural settings.

Investigate prevalence before causes

Next, we argue that descriptive studies that quantify the scope of problematic outcomes—such as the share of users in filter bubbles or the prevalence of low-quality content consumption—are necessary first steps. These findings prevent overgeneralization from anecdotal examples[21,53] and ensure research focuses on problems that actually exist at a meaningful scale[62,63]. After establishing that a problem affects substantial populations or causes significant harm, resources can be directed accordingly to investigate causal mechanisms and develop interventions. Further, quantitative research based on observational data provides valuable groundwork for developing hypotheses: although correlation should not be confused with causation, systematic correlational patterns can help uncover plausible hypotheses to test experimentally.

Cross-platform analyses provide one such example. Studies comparing user behavior across social media ecosystems reveal that many engagement patterns—such as preference for low-quality or partisan content—persist across platforms with very different algorithms, policies and audiences[14,42,65]. Although these findings are largely correlational, comparative studies over the same population and time frames can provide supported evidence on whether observed amplification arises from algorithmic design or from stable user demand that manifests wherever people go online.

Match research design to research question

As many observed patterns on online platforms—such as content exposure, attention and downstream behavioral effects—arise from the interplay of user demand, platform design and external events, researchers studying deployed algorithmic systems should ensure that their data, methods and analytic strategies correspond adequately to the questions they seek to answer. A key challenge for researchers is that many societally consequential outcomes, such as harmful content exposure and echo chambers, are rare at the population level and disproportionately concentrated among extreme users and content. A critical first step is identifying that population of interest and ensuring the strategy is capable of capturing it (meaning, quality measures for the stakeholder needs).

Similarly, researchers should also consider how data limitations and methodological choices affect the validity and generalizability of their conclusions. For instance, sock-puppet accounts are frequently used to audit whether recommendation algorithms amplify exposure to problematic content or steer users into 'rabbit holes'[89,90]. Yet such synthetic agents fail to capture real human usage patterns and may therefore distort findings, as they measure the algorithmic supply of recommendations rather than the causal role of algorithms in shaping actual user experience within human–algorithm interactions. Other challenges are more complex, and sometimes even impossible to test empirically with current methods. For example, when studying the feedback loop effect between human behavior and algorithmic recommendations, a critical design decision is how long an experiment should be run, as the effects of algorithmic exposure can take weeks, months or even years to manifest. Short-term experiments may fail to capture

the cumulative or delayed nature of these dynamics, raising questions about whether prolonged exposure to like-minded content may solidify opinions to the point where short-term exposure to diverse content has little observed effect[98]. Addressing the effect of exposure on users or the efficacy of interventions such as diversified or cross-partisan news requires longitudinal studies that tracks user behavior over extended periods, which is lacking in many recent studies[15,43,98,159].

Pursue causal identification through complementary approaches

A key challenge for researchers remains their ability to provide causal answers when often the available evidence is dominated by descriptive studies and correlational findings. This challenge becomes particularly acute when moving beyond the causal role of algorithms alone and uncovering the role of human behavior and societal demand. Humans and platforms must be studied in context, over time, to properly determine whether their interaction has societal implications or suggest proper interventions or policy reforms[49,51]. Auditing algorithms in isolation does not answer the key socio-technical questions researchers want to know, nor do one-shot descriptive studies with limited access[49,51,52]. Both perspectives capture partial truths, and causal research must integrate them[160]. This necessity highlights the lack of access to platforms, which places a hard-to-overcome limit on the extent to which researchers can examine these critical interactions in depth. As a result, and as noted earlier, the majority of causal explorations of socio-technical systems can be grouped into either causal interpretations from correlational observations or causal studies that examine the role of algorithms in isolation.

Observed information consumption reflects the joint product of algorithmic curation and user preferences, and disentangling these two forces is among the field's central challenges. We argue that observational data—when coupled with careful, validated causal designs—can address this challenge through two experimental strategies. The first strategy isolates algorithmic amplification by varying the recommendation system while holding user behavior constant. Building on the concept of relative algorithmic amplification[160], researchers can quantify how content distribution changes under one algorithm versus an alternative. This approach requires defining clear baselines and outcomes: amplified relative to what, and in terms of what exposure measure. For example, reverse-chronological feeds, random recommendation orders, or historical versions of ranking systems can serve as counterfactual baselines[43,98]. The second strategy isolates behavioral amplification by fixing the platform algorithm and experimentally varying user behavior. Counterfactual-bot designs on YouTube[92] exemplify this approach: bots trained on real users' histories are later assigned rule-based viewing policies with no agency versus the actual user experience, which preserves user agency. Comparing their subsequent exposures quantifies the causal effect of changing behavioral policy while keeping the recommender unchanged, revealing how much of the observed bias arises from human choice rather than platform logic.

With the recent advancements in the availability of large-scale simulation tools and agent-based modeling environments, these earlier designs can now be substantially extended. Digital twins—computational replicas of real users built on realistic behavioral patterns—allow for controlled tests for various platform safety features and feedback-loop formation at scale[161]. Such digital-twin systems can incorporate decision rules as the 'brain' and automated accounts as the 'arm', allowing repeated interaction with recommendation systems over extended simulated horizons. A natural further extension is to replace rule-based behavioral policies with LLM-driven agents, whose ability to generate contextually coherent, heterogeneous responses more closely approximates the diversity of real user behavior[162–164]. Where earlier bot designs assigned fixed viewing policies, LLM-based agents can dynamically adapt their content selection in response to recommendations—simulating users with different political priors, susceptibility to sensationalism, or cross-cutting curiosity—while still allowing the researcher to hold the recommendation algorithm constant. This approach makes it possible to study not just whether algorithms amplify problematic content, but for whom and under what behavioral conditions such amplification occurs. These synthetic yet human-like environments create realistic proxies of engagement rhythms, time spent and responsiveness, allowing researchers to systematically test how various platform features influence human–algorithm co-evolution, to effectively bridge the gap between short-term laboratory experiments and sustained real-world behavior. Moreover, they make it possible to investigate sensitive topics that cannot be ethically tested with human participants. In the context of digital platforms, such agents could systematically explore counterfactual scenarios[136], such as testing how ranking objectives, moderation strategies or interface changes affect the emergence of echo chambers or the spread of low-quality information[165,166], while complementing observational and field-experimental approaches. That said, LLM-based agents cannot fully substitute for human subjects, and their behavioral fidelity at the level of click patterns, dwell time and content selection under realistic algorithmic pressure remains largely unvalidated[167,168]. One promising approach is to evaluate synthetic agents against behavioral traces collected from data-donation projects, browser panels or other sources of observed online behavior. Such benchmarks could assess whether synthetic agents reproduce not only average patterns of exposure and engagement, but also the behavior of the small subpopulations that often account for a disproportionate share of problematic content consumption. Developing and validating such benchmarks therefore represents an important direction for future research.

Interpret findings carefully and transparently

We suggest careful interpretation of the studies for such complex systems, while acknowledging limitations. For example, conclusions from experimental research should be carefully restricted to the populations accurately represented by the study sample. Extrapolating findings to broader populations without robust evidence, particularly when the sample characteristics differ from the general population[73,116], risks overgeneralization and undermines the validity and reliability of the research. Similarly, it is important to keep in mind that conclusions from the study are valid only for the duration of the study, and therefore the results should not be extrapolated to other time periods. For instance, it is possible that the studies finding null effect on users' attitudes in studies conducted specifically during election time periods[98] would have found some significant effects had the studies been conducted during different time periods, perhaps, because it is difficult for people to be persuaded by opposing media content during a contentious election campaign[169]. Finally, when interpreting their findings, researchers should not conflate statistical significance with practical significance. A small effect size, even if statistically significant, may be less consequential in domains like education or psychology, where larger effects are often necessary to achieve meaningful change. Conversely, even small effects can have substantial implications in fields like public policy or healthcare when they impact large populations or involve critical outcomes. Therefore, studies should provide a context-dependent interpretation of effect sizes to offer a more nuanced understanding of the real-world implications of their findings.

Transparency in situating the findings, especially given the many moving parts involved in studying platforms, is essential at every step of the research pipeline. Another researcher should be able to replicate the study based on the released materials. For example, in many social or political science constructs, there is ongoing debate about whether humans can generate high-quality annotations. Transparency in steps such as data sampling, survey design and providing demographic information about annotators will facilitate replication and enable building upon previous work. Such an approach also ensures that small

observed effects are scrutinized, for example, to find whether they might disappear if a different metric were chosen or if the quality of the measurements was properly validated. Similarly, a persistent problem that limits the impact and reproducibility of platform audit studies is the lack of transparency and insufficient methodological detail. Too often, key aspects of the design, such as data-collection protocols, model parameters or the full pipeline for simulation and labeling, are not made publicly available. This prevents independent verification and undermines cumulative progress. Recent research has demonstrated how sensitive algorithmic findings can be to seemingly minor decisions at multiple stages of the pipeline—ranging from the choice of seed content to the authentication state of the auditing account and the window length of the experiment[170–172]. Even small variations in these parameters can reverse or erase observed effects[173], highlighting the critical need for detailed documentation and reproducibility standards, as expected in any social science research.

Address platform and data-access challenges

Robust causal and systemic research on socio-technical systems requires sustained and transparent access on two fronts: platform data, and the rare subpopulations that occupy the tail of consumption. Yet the current research environment is marked by declining access, fragmented APIs, and opaque and rare collaboration models. To make meaningful progress, the field needs a collective push toward institutional mechanisms that make platform access both feasible and trustworthy. Industry partnerships can contribute to platform research in two critical ways: by improving the structure of collaborations and by expanding researcher access through stable, auditable interfaces.

First, we call for a new generation of transparent and replicable academic–industry collaborations, modeled after open-science principles. Although large-scale partnerships like Facebook's 2020 election studies[15,73,98] provided unprecedented access to platform data, they also revealed weaknesses in transparency, replicability and independent verification[72,98,174]. Future collaborations should emphasize open-science models in which multiple independent research groups can reproduce analyses using controlled data-access systems (for instance, census-style secure repositories). Emerging initiatives such as the National Internet Observatory (NIO) are promising steps toward this vision, providing independent, privacy-preserving infrastructure for studying platform effects at scale[175]. Over time, concerns about reproducibility may be more effectively addressed if, instead of large multi-author collaborations producing a single study, multiple research groups conduct the study independently on the same research questions.

Second, a stronger commitment to sustained data access is equally critical. The erosion of open APIs—for example, X's termination of free API access in 2023[176] and Meta's replacement of CrowdTangle[177] with the Meta Content Library[178]—has severely constrained academic inquiry, with the impact falling hardest on early-career and independent scholars who lack direct industry ties or funding for proprietary data agreements. Meta's decision not to replicate its 2020 collaboration in 2024 underscores the fragility of ad hoc arrangements.

Ideally, platforms will move beyond ad hoc data-sharing and establish formal researcher-whitelisting programs and dedicated research APIs that guarantee safe, auditable and sustained access to core endpoints. These infrastructures would preserve privacy while enabling reproducible science, effectively translating the benefits of large-scale collaborations into a more distributed and equitable model of access. Such mechanisms would lower barriers to innovation, broaden participation, and reduce the duplication of effort that currently forces researchers to engineer alternate data-collection pipelines or develop synthetic environments to study real-world systems. They would also allow independent teams to pursue original designs without relying on individualized data agreements that limit transparency and scalability. For example, allowing verified researchers to deploy automated agents that simulate human behavior under controlled conditions would open new avenues for studying user agency, exposure mechanisms and content dynamics. Current restrictions have forced scholars to invest extraordinary effort into developing pipelines to pursue core questions about socio-technical dynamics that cannot be addressed with existing data access[92,159,179]. The European Union's Digital Services Act (DSA) represents an attempt to institutionalize such access[180]. Under Article 40, the DSA provides legal guarantees for vetted and independent researchers to access both internal and publicly available platform data under privacy- and security-conscious frameworks[181]. The United States could adopt similar legal or policy mechanisms to ensure that data accessibility for public-interest research becomes a durable institutional norm rather than a temporary corporate favor. Securing platform data, however, addresses only the first front. Complementary initiatives such as data-donation programs or browser extensions can further broaden visibility into users' online experiences, but their reach is limited. Obtaining data on a specific tail population is as much an access problem as a platform-data one, and the approach depends on how that rare population is constituted. Some tails can be targeted through institutions: younger adults for bullying or companion-AI exposure within a school, vulnerable populations for eating disorders, self-harm or suicidal content within clinics, or for health misinformation within oncology populations. For these, cross-disciplinary collaboration supplies both a list frame for an otherwise unframed population and a clinically or contextually validated definition that no behavioral proxy can match[182], together with the consent, monitoring and oversight without which the most vulnerable of these populations cannot be studied responsibly at all. Other tails are defined purely by behavior and have no institutional home—consumers of conspiratorial or far-right content, users with unimodal or otherwise extreme diets—and can be reached instead by screening. Either way, the resulting samples are non-random: institutional and self-selected sources alike characterize *a* tail rather than *the* tail, and reweighting against a population frame remains necessary.

Whether the binding constraint is platform data or the populations themselves, a sustainable ecosystem for platform research ultimately depends on a genuine institutional commitment to transparency and academic independence.

Incorporate economic and behavioral perspectives into platform research

Finally, the business models of online platforms merit closer scrutiny. A growing body of work conceptualizes platforms as two-sided markets, where content supply and user demand are jointly shaped by advertising incentives and engagement metrics[18]. From this perspective, the rise of online platforms has not necessarily introduced new behavioral tendencies so much as it has made preexisting social and psychological dynamics visible at unprecedented scale[50]. The observed consumption patterns may therefore reflect not only algorithmic influence but also the underlying structure of societal demand, a reality that may be uncomfortable to acknowledge[103].

If recent studies show that users' own choices contribute to the consumption of problematic content[65,101], an important question follows: should algorithms simply mirror user preferences? Recently, scholars have begun questioning the long-standing assumption that engagement is a reliable proxy for what the user wants. Platforms often equate clicks, watch time or session length with satisfaction, yet behavioral evidence suggests users frequently make choices inconsistent with their long-term preferences[183]. Such inconsistencies expose the limits of revealed-preference modeling and call for new theoretical and empirical work to distinguish momentary engagement from enduring well-being[184–187].

More broadly, researchers should explore ways to redefine the objective functions of recommender systems to reflect societal rather than purely individual or commercial values[185]. Rather than optimizing

for stimulation or retention, platforms could explicitly encode goals such as promoting civic participation, community health or informed discourse. Achieving this vision requires both empirical measurement—linking content design, exposure and outcomes—and theoretical clarity about what constitutes welfare in digital environments, which has stayed largely hypothetical. For example, recent studies have begun to move beyond conventional accuracy-based metrics, introducing new optimization frameworks aimed at promoting diversity[188] or fairness[189]. However, most of these approaches have been developed and evaluated in commercial or entertainment benchmarks, such as product (for example, Amazon) or movie recommendations (for example, Netflix, Spotify), rather than in the context of public information or news exposure. Future research could extend multi-objective optimization frameworks to domains where societal values—such as exposure diversity and informational integrity—are central to democratic outcomes.

Outlook

In this Review we raise the question of why so much public and academic attention is focused on algorithms when the available evidence strongly suggests the significant role of demand in concerning online consumption patterns such as echo chambers and engagement with problematic content. One possible reason is that algorithms offer a seemingly simple solution: 'fix the algorithms',[190] whereas addressing demand is far more complex and may not even have a clear solution (or any at all). Correspondingly, focusing on algorithms can therefore feel more tractable than confronting the demand-side dynamics or the business models and design incentives that shape them. As we have argued, however, the reality necessitates that we confront the deeper, more challenging aspects of the issue, even if it is tricky and uncomfortable for us to address. Fortunately, such investigations are possible, and recent years have witnessed meaningful progress, but much more progress is needed. Investigating socio-technical systems demands a rigorous, context-sensitive approach that respects both the technical intricacies of algorithmic platforms and the complex realities of human behavior. Greater transparency, improved data sharing, careful attention to study design, and innovative methods that combine observational data with automated user simulations can all help move the field forward. Above all, researchers should resist the temptation to reduce the complexity of socio-technical problems to purely technical ones. Expanding the state space of analysis to include the social, behavioral and institutional dimensions will enable a more complete account of how platforms and societies mutually shape one another.

We also emphasize that our interpretation of the current literature reflects the current balance of evidence rather than a settled conclusion. Although the current evidence does not generally support strong claims that algorithms are the dominant driver of problematic outcomes, future evidence may substantially revise this assessment. Particularly informative would be studies that identify persistent effects of algorithmic interventions across multiple contexts, demonstrate cumulative effects that emerge over longer time horizons, or show that otherwise similar users experience systematically different outcomes under different recommendation architectures. Likewise, evidence that population-level consequences arise from algorithmic systems in ways that cannot be explained by user selection, content supply or other competing mechanisms would strengthen algorithmic explanations. We therefore view the relative contributions of these interacting forces as an open empirical question that requires continued investigation as stronger evidence accumulates.

Finally, we recognize that debates about algorithmic effects and an emphasis on the limits of current evidence are not without implications for ongoing regulatory debates. Our argument is not that platforms should face less scrutiny, but that accountability efforts should be informed by an understanding of the mechanisms involved. Indeed, greater transparency, independent auditing, and researcher access to platform data are our recommendations to improve that understanding.

Data availability

No new data were generated or analyzed in this Review. All studies discussed are available in the cited sources.

Code availability

No new code was generated or analyzed in this Review. All studies discussed are available in the cited sources.

References

1. Eckles, D. *Testimony before the Senate Subcommittee on Communications, Media, and Broadband: Algorithmic Transparency and Assessing Effects of Algorithmic Ranking* (US Senate Committee on Commerce, Science and Transportation, 2022).
2. Milli, S. et al. Engagement, user satisfaction, and the amplification of divisive content on social media. *Proc. Natl Acad. Sci. Nexus* **4**, pgaf062 (2025).
3. Knight, M. Explainer: how Facebook has become the world's largest echo chamber. *The Conversation* https://theconversation.com/explainer-how-facebook-has-become-the-worlds-largest-echo-chamber-91024 (2018).
4. Wojcieszak, M., Casas, A., Yu, X., Nagler, J. & Tucker, J. A. Echo chambers revisited: the (overwhelming) sharing of in-group politicians, pundits and media on Twitter. Preprint at https://doi.org/10.31219/osf.io/xwc79_v1 (2021).
5. Cinelli, M., De Francisci Morales, G., Galeazzi, A., Quattrociocchi, W. & Starnini, M. The echo chamber effect on social media. *Proc. Natl Acad. Sci. USA* **118**, e2023301118 (2021).
6. Hswen, Y., Xu, X., Hing, A., Hawkins, J. B., Brownstein, J. S. & Gee, G. C. Association of "# covid19" versus "# chinesevirus" with anti-Asian sentiments on Twitter: March 9–23, 2020. *Am. J. Public Health* **111**, 956–964 (2021).
7. Noble, S. U. *Algorithms of Oppression* (New York Univ. Press, 2018).
8. Shmargad, Y. & Klar, S. Sorting the news: how ranking by popularity polarizes our politics. *Polit. Commun.* **37**, 423–446 (2020).
9. Van Bavel, J. J., Rathje, S., Harris, E., Robertson, C. & Sternisko, A. How social media shapes polarization. *Trends Cogn. Sci.* **25**, 913–916 (2021).
10. Levy, R. Social media, news consumption, and polarization: evidence from a field experiment. *Am. Econ. Rev.* **111**, 831–870 (2021).
11. Allcott, H., Braghieri, L., Eichmeyer, S. & Gentzkow, M. The welfare effects of social media. *Am. Econ. Rev.* **110**, 629–676 (2020).
12. Strümke, I., Slavkovik, M. & Stachl, C. Against algorithmic exploitation of human vulnerabilities. Preprint at https://arxiv.org/abs/2301.04993 (2023).
13. Ceylan, G., Anderson, I. A. & Wood, W. Sharing of misinformation is habitual, not just lazy or biased. *Proc. Natl Acad. Sci. USA* **120**, e2216614120 (2023).
14. Chen, A. Y., Nyhan, B., Reifler, J., Robertson, R. E. & Wilson, C. Subscriptions and external links help drive resentful users to alternative and extremist YouTube channels. *Sci. Adv.* **9**, eadd8080 (2023).
15. Guess, A. M. et al. Reshares on social media amplify political news but do not detectably affect beliefs or opinions. *Science* **381**, 404–408 (2023).
16. Munger, K. All the news that's fit to click: the economics of clickbait media. *Polit. Commun.* **37**, 376–397 (2020).
17. Munger, K. & Phillips, J. Right-wing YouTube: a supply and demand perspective. *Int. J. Press Polit.* **27**, 186–219 (2022).
18. Aridor, G., Jiménez-Durán, R., Levy, R. & Song, L. The economics of social media. *J. Econ. Lit.* **62**, 1422–1474 (2024).

19. DiResta, R. et al. *The Tactics & Tropes of the Internet Research Agency* (UN Digital Commons, 2019).
20. Golovchenko, Y., Buntain, C., Eady, G., Brown, M. A. & Tucker, J. A. Cross-platform state propaganda: Russian trolls on Twitter and YouTube during the 2016 US presidential election. *Int. J. Press Polit.* **25**, 357–389 (2020).
21. Tufekci, Z. YouTube, the great radicalizer. *The New York Times* https://www.nytimes.com/2018/03/10/opinion/sunday/youtube-politics-radical.html (2018).
22. Lewis, P. 'Fiction is outperforming reality': how YouTube's algorithm distorts truth. *The Guardian* https://www.theguardian.com/technology/2018/feb/02/how-youtubes-algorithm-distorts-truth (2018).
23. Levin, S. Las Vegas survivors furious as YouTube promotes clips calling shooting a hoax. *The Guardian* https://www.theguardian.com/us-news/2017/oct/04/las-vegas-shooting-youtube-hoax-conspiracy-theories/ (2017).
24. Twenge, J. M. Have smartphones destroyed a generation. *The Atlantic* https://www.theatlantic.com/magazine/archive/2017/09/has-the-smartphone-destroyed-a-generation/534198/ (2017).
25. Haidt, J. & Twenge, J. M. This is our chance to pull teenagers out of the smartphone trap. *The New York Times* https://www.nytimes.com/2021/07/31/opinion/smartphone-iphone-social-media-isolation.html (2021).
26. Waterson, J. & Milmo, D. Facebook whistleblower Frances Haugen calls for urgent external regulation. *The Guardian* https://www.theguardian.com/technology/2021/oct/25/facebook-whistleblower-frances-haugen-calls-for-urgent-external-regulation (2021).
27. Stokel-Walker, C. The EU wants to label 'addictive design' a systemic risk under the DSA https://www.techpolicy.press/the-eu-wants-to-label-addictive-design-a-systemic-risk-under-the-dsa/ (2026).
28. Tiffany, K. Can Instagram ruin your life? The jury will decide. *The Atlantic* https://www.theatlantic.com/technology/2026/02/instagram-meta-addiction-lawsuits/685947/ (2026).
29. Sunstein, C. *Republic.com* (Princeton Univ. Press, 2001).
30. Sunstein, C. *Republic.com 2.0* (Princeton Univ. Press, 2007).
31. Johnson, S. L., Kitchens, B. & Gray, P. Facebook serves as an echo chamber, especially for conservatives. blame its algorithm. *The Washington Post* https://www.washingtonpost.com/opinions/2020/10/26/facebook-algorithm-conservative-liberal-extremes/ (2020).
32. Pariser, E. *The Filter Bubble: How the New Personalized Web is Changing What We Read and How We Think* (Penguin Books, 2012).
33. Shapiro, A., Levitt, M. & Intagliata, C. How the polarizing effect of social media is speeding up. *NPR* (9 September, 2022).
34. Fisher, M. *The Chaos Machine: the Inside Story of How Social Media Rewired our Minds and Our World* (Little Brown and Company, 2022).
35. Whitaker, B. Social media's role in America's polarized political climate. *CBS News* https://www.cbsnews.com/news/social-media-political-polarization-60-minutes-2022-11-06/ (2022).
36. Lorenz-Spreen, P., Oswald, L., Lewandowsky, S. & Hertwig, R. A systematic review of worldwide causal and correlational evidence on digital media and democracy. *Nat. Hum. Behav.* **7**, 74–101 (2023).
37. Haidt, J. Yes, social media really is undermining democracy. *The Atlantic* https://www.theatlantic.com/ideas/archive/2022/07/social-media-harm-facebook-meta-response/670975/ (2022).
38. El-Bermawy, M. M. Your filter bubble is destroying democracy. *Wired* https://www.wired.com/2016/11/filter-bubble-destroying-democracy/ (2016).
39. Sunstein, C. *#Republic: Divided Democracy in the Age of Social Media* (Princeton Univ. Press, 2018).
40. Benkler, Y. *Network Propaganda: Manipulation, Disinformation, and Radicalization in American Politics* (Oxford Academic, 2018).
41. Vaidhyanathan, S. *Antisocial Media: How Facebook Disconnects Us and Undermines Democracy* (Oxford Univ. Press, 2018).
42. Mosleh, M., Allen, J. & Rand, D. G. Divergent patterns of engagement with partisan and low-quality news across seven social media platforms. *Proc. Natl Acad. Sci. USA* **122**, e2425739122 (2025).
43. Wang, S., Huang, S., Zhou, A. & Metaxa, D. Lower quantity, higher quality: auditing news content and user perceptions on Twitter/X algorithmic versus chronological timelines. In *Proc. ACM on Human Computer Interaction* Vol. 8 (ed. Nichols, J.) 57 (ACM, 2024).
44. Anti-Defamation League, Avaaz, Decode Democracy, Mozilla, & New America's Open Technology Institute Ranking and recommendation systems https://www.mozillafoundation.org/en/campaigns/trained-for-deception-how-artificial-intelligence-fuels-online-disinformation/ranking-and-recommendation-systems/ (Mozilla Foundation, 2023).
45. Narayanan, A. Understanding social media recommendation algorithms https://knightcolumbia.org/content/understanding-social-media-recommendation-algorithms (The Knight First Amendment Institute, 2023).
46. Ribeiro, M. H., Veselovsky, V. & West, R. The amplification paradox in recommender systems. In *Proc. International AAAI Conference on Web and Social Media* (eds Lin, Y.-R. et al.) Vol. 17, 1138–1142 https://doi.org/10.1609/icwsm.v17i1.22223 (AAAI Press, 2023).
47. Mansoury, M., Abdollahpouri, H., Pechenizkiy, M., Mobasher, B. & Burke, R. Feedback loop and bias amplification in recommender systems. In *Proc. 29th ACM International Conference on Information and Knowledge Management* 2145–2148 (ACM, 2020).
48. Jiang, R., Chiappa, S., Lattimore, T., György, A. & Kohli, P. Degenerate feedback loops in recommender systems. In *Proc. 2019 AAAI/ACM Conference on AI, Ethics and Society* 383–390 (ACM, 2019).
49. Pedreschi, D. et al. Human-AI coevolution. *Artif. Intell.* **339**, 104244 (2024).
50. Wagner, C. et al. Measuring algorithmically infused societies. *Nature* **595**, 197–204 (2021).
51. Matias, J. N. Humans and algorithms work together-so study them together. *Nature* **617**, 248–251 (2023).
52. Lewandowsky, S., Robertson, R. E. & DiResta, R. Challenges in understanding human-algorithm entanglement during online information consumption. *Perspect. Psychol. Sci.* **19**, 758–766 (2024).
53. Roose, K. The making of a YouTube radical. *The New York Times* https://www.nytimes.com/interactive/2019/06/08/technology/youtube-radical.html (2019).
54. Mozilla. YouTube regrets: a crowdsourced investigation into YouTube's recommendation algorithm https://www.mozillafoundation.org/en/youtube/findings/ (Mozilla Foundation, 2022).
55. Myers, S. L. How social media amplifies misinformation more than information. *The New York Times* https://www.nytimes.com/2022/10/13/technology/misinformation-integrity-institute-report.html (2022).
56. Sunstein, C. R. *Echo Chambers: Bush v. Gore, Impeachment, and Beyond* (Princeton Univ. Press, 2001).
57. Bruns, A. It's not the technology, stupid: how the 'echo chamber' and 'filter bubble' metaphors have failed us. In *Proc. IAMCR Conference* 19771 (International Association for Media and Communication Research, 2019).
58. Dahlgren, P. M. A critical review of filter bubbles and a comparison with selective exposure. *Nordicom Rev.* **42**, 15–33 (2021).
59. Paul, K. Facebook harms children, damaging democracy, claims whistleblower. *The Guardian* https://www.theguardian.com/technology/2021/oct/05/facebook-harms-children-damaging-democracy-claims-whistleblower (2021).

60. Allen, J. Misinformation amplification analysis and tracking dashboard. *The Integrity Institute* https://www.integrityinstitute.org/blog/misinformation-amplification-tracking-dashboard
61. Auxier, B. 64% of Americans say social media have a mostly negative effect on the way things are going in the U.S. today. https://www.pewresearch.org/short-reads/2020/10/15/64-of-americans-say-social-media-have-a-mostly-negative-effect-on-the-way-things-are-going-in-the-u-s-today/ (Pew Research Center, 2020).
62. Budak, C., Nyhan, B., Rothschild, D. M., Thorson, E. & Watts, D. J. Misunderstanding the harms of online misinformation. *Nature* **630**, 45–53 (2024).
63. Allen, J., Howland, B., Mobius, M., Rothschild, D. & Watts, D. J. Evaluating the fake news problem at the scale of the information ecosystem. *Sci. Adv.* **6**, eaay3539 (2020).
64. Guess, A. M., Nyhan, B. & Reifler, J. Exposure to untrustworthy websites in the 2016 US election. *Nat. Hum. Behav.* **4**, 472–480 (2020).
65. Hosseinmardi, H. et al. Examining the consumption of radical content on YouTube. *Proc. Natl Acad. Sci. USA* **118**, e2101967118 (2021).
66. Gentzkow, M. & Shapiro, J. M. Ideological segregation online and offline. *Q. J. Econ.* **126**, 1799–1839 (2011).
67. Eady, G. et al. Exposure to the Russian Internet Research Agency foreign influence campaign on Twitter in the 2016 US election and its relationship to attitudes and voting behavior. *Nat. Commun.* **14**, 62 (2023).
68. Grinberg, N., Joseph, K., Friedland, L., Swire-Thompson, B. & Lazer, D. Fake news on Twitter during the 2016 US presidential election. *Science* **363**, 374–378 (2019).
69. Guess, A. M., Nyhan, B., O'Keeffe, Z. & Reifler, J. The sources and correlates of exposure to vaccine-related (mis)information online. *Vaccine* **38**, 7799–7805 (2020).
70. Bergeron-Boutin, O. et al. Untrustworthy sources on Facebook and Instagram in 2020: concentrated exposure but no attitudinal effects. *Sci. Adv.* **12**, eadz6502 (2026).
71. Guess, A., Nyhan, B., Lyons, B. & Reifler, J. Avoiding the echo chamber about echo chambers https://kf-site-production.s3.amazonaws.com/media_elements/files/000/000/133/original/Topos_KF_White-Paper_Nyhan_V1.pdf (Knight Foundation, 2018).
72. González-Bailón, S. et al. Asymmetric ideological segregation in exposure to political news on Facebook. *Science* **381**, 392–398 (2023).
73. Nyhan, B. et al. Like-minded sources on Facebook are prevalent but not polarizing. *Nature* **620**, 137–144 (2023).
74. Muise, D. et al. Quantifying partisan news diets in Web and TV audiences. *Sci. Adv.* **8**, eabn0083 (2022).
75. Eady, G., Nagler, J., Guess, A., Zilinsky, J. & Tucker, J. A. How many people live in political bubbles on social media? Evidence from linked survey and Twitter data. *Sage Open* https://doi.org/10.1177/2158244019832705 (2019).
76. Guess, A. M. (Almost) everything in moderation: new evidence on Americans' online media diets. *Am. J. Polit. Sci.* **65**, 1007–1022 (2021).
77. Robertson, R. E. et al. Auditing partisan audience bias within google search. *Proc. ACM. Hum. Comput. Interact.* **2**, 1–22 (2018).
78. Courtois, C., Slechten, L. & Coenen, L. Challenging google search filter bubbles in social and political information: disconforming evidence from a digital methods case study. *Telemat. Inform.* **35**, 2006–2015 (2018).
79. Hannak, A. et al. Measuring personalization of web search. In *Proc. 22nd International Conference on World Wide Web* 527–538 (ACM, 2013).
80. Haim, M., Graefe, A. & Brosius, H.-B. Burst of the filter bubble? Effects of personalization on the diversity of Google News. *Digit. Journal.* **6**, 330–343 (2018).
81. Zuiderveen Borgesius, F. J. et al. Should we worry about filter bubbles? *Internet Policy Rev.* https://doi.org/10.14763/2016.1.401 (2016).
82. Bruns, A. Filter bubble. *Internet Policy Rev.* https://doi.org/10.14763/2019.4.1426 (2019).
83. Flaxman, S., Goel, S. & Rao, J. M. Filter bubbles, echo chambers, and online news consumption. *Public Opin. Q.* **80**, 298–320 (2016).
84. Fletcher, R. & Nielsen, R. K. Automated serendipity: the effect of using search engines on news repertoire balance and diversity. *Digit. Journal.* **6**, 976–989 (2018).
85. Lorenz, T. Substack sent a push alert promoting a Nazi blog. *User Magazine* https://www.usermag.co/p/substack-sent-a-push-alert-promoting-nazi-white-supremacist-blog (2025).
86. Whittaker, J., Looney, S., Reed, A. & Votta, F. Recommender systems and the amplification of extremist content. *Internet Policy Rev.* https://doi.org/10.14763/2021.2.1565 (2021).
87. Hussein, E., Juneja, P. & Mitra, T. Measuring misinformation in video search platforms: an audit study on YouTube. *Proc. ACM Hum. Comput. Interact.* **4**, 1–27 (2020).
88. Papadamou, K. et al. 'It is just a flu': assessing the effect of watch history on YouTube's pseudoscientific video recommendations. In *Proc. International AAAI Conference on Web and Social Media* Vol. 16, 723–734 (2022).
89. Haroon, M. et al. Auditing YouTube's recommendation system for ideologically congenial, extreme, and problematic recommendations. *Proc. Natl Acad. Sci. USA* **120**, e2213020120 (2023).
90. Ibrahim, H., AlDahoul, N., Lee, S., Rahwan, T. & Zaki, Y. YouTube's recommendation algorithm is left-leaning in the United States. *PNAS Nexus* **2**, pgad264 (2023).
91. Ibrahim, H. et al. Systematic partisan content skews in TikTok during the 2024 US elections. *Nature* **65**, 1004–1011 (2026).
92. Hosseinmardi, H. et al. Causally estimating the effect of YouTube's recommender system using counterfactual bots. *Proc. Natl Acad. Sci. USA* **121**, e2313377121 (2024).
93. Bandy, J. & Diakopoulos, N. More accounts, fewer links: how algorithmic curation impacts media exposure in Twitter timelines. *Proc. ACM Hum. Comput. Interact.* **5**, 1–28 (2021).
94. Huszár, F. et al. Algorithmic amplification of politics on Twitter. *Proc. Natl Acad. Sci. USA* **119**, e2025334119 (2022).
95. Bakshy, E., Messing, S. & Adamic, L. A. Exposure to ideologically diverse news and opinion on Facebook. *Science* **348**, 1130–1132 (2015).
96. Hosanagar, K. Blame the echo chamber on Facebook. But blame yourself, too. *Wired* https://www.wired.com/2016/11/facebook-echo-chamber/ (2016).
97. Hosanagar, K., Fleder, D., Lee, D. & Buja, A. Will the global village fracture into tribes? Recommender systems and their effects on consumer fragmentation. *Manage. Sci.* **60**, 805–823 (2013).
98. Guess, A. M. et al. How do social media feed algorithms affect attitudes and behavior in an election campaign? *Science* **381**, 398–404 (2023).
99. Gauthier, G., Hodler, R., Widmer, P. & Zhuravskaya, E. The political effects of X's feed algorithm. *Nature* **652**, 416–423 (2026).
100. Bagchi, C. et al. Social media algorithms can curb misinformation, but do they? Preprint at https://arxiv.org/abs/2409.18393 (2024).
101. Robertson, R. E. et al. Users choose to engage with more partisan news than they are exposed to on Google Search. *Nature* **618**, 342–348 (2023).
102. Bak-Coleman, J. B. et al. Moving towards informative and actionable social media research. Preprint at https://doi.org/10.48550/arXiv.2505.09254 (2025).
103. Williams, D. Is social media destroying democracy or giving it to us good and hard? *Conspicuous Cognization* https://www.conspicuouscognition.com/p/is-social-media-destroying-democracyor/ (2025).

104. Ledwich, M. & Zaitsev, A. Algorithmic extremism: examining YouTube's rabbit hole of radicalization. Preprint at https://arxiv.org/abs/1912.11211 (2019).
105. Mele, N. et al. Combating fake news: an agenda for research and action https://www.hks.harvard.edu/publications/combating-fake-news-agenda-research-and-action (Harvard Kennedy School, 2017).
106. Scheufele, D. A. & Krause, N. M. Science audiences, misinformation, and fake news. *Proc. Natl Acad. Sci. USA* **116**, 7662–7669 (2019).
107. Sellars, A. *Defining Hate Speech*. Research Publication 16 (Berkman Klein Center, 2016).
108. Kim, B., Xiong, A., Lee, D. & Han, K. A systematic review on fake news research through the lens of news creation and consumption: research efforts, challenges, and future directions. *PLoS ONE* **16**, e0260080 (2021).
109. Michiels, L., Leysen, J., Smets, A. & Goethals, B. What are filter bubbles really? A review of the conceptual and empirical work. In *Adjunct Proc. 30th ACM Conference on User Modeling, Adaptation and Personalization* 274–279 (ACM, 2022).
110. Ross Arguedas, A., Robertson, C., Fletcher, R. & Nielsen, R. Echo chambers, filter bubbles, and polarisation: a literature review https://reutersinstitute.politics.ox.ac.uk/echo-chambers-filter-bubbles-and-polarisation-literature-review (Reuters Institute for the Study of Journalism, 2022).
111. Sachdeva, P. S., Barreto, R., von Vacano, C. & Kennedy, C. J. Assessing annotator identity sensitivity via item response theory: a case study in a hate speech corpus. In *Proc. 2022 ACM Conference on Fairness, Accountability and Transparency* 1585–1603 (ACM, 2022).
112. Lin, W.-H. & Hauptmann, A. Vox Populi Annotation: measuring intensity of ideological perspectives by aggregating group judgments. In *Proc. Sixth International Conference on Language Resources and Evaluation* (LREC, 2008).
113. Green, J. et al. Curation bubbles. *Am. Polit. Sci. Rev.* **119**, 1704–1722 (2025).
114. Hosseinmardi, H., Wolken, S., Rothschild, D. M. & Watts, D. J. Unpacking media bias in the growing divide between cable and network news. *Sci. Rep.* **15**, 17607 (2025).
115. Kim, E., Lelkes, Y. & McCrain, J. Measuring dynamic media bias. *Proc. Natl Acad. Sci. USA* **119**, e2202197119 (2022).
116. Ribeiro, M. H., Ottoni, R., West, R., Almeida, V. A. & Meira, W. Jr. Auditing radicalization pathways on YouTube. In *Proc. 2020 Conference on Fairness, Accountability and Transparency* 131–141 (ACM, 2020).
117. Ying, L., Montgomery, J. M. & Stewart, B. M. Topics, concepts and measurement: a crowdsourced procedure for validating topics as measures. *Polit. Anal.* **30**, 570–589 (2021).
118. Bernard, H. R., Killworth, P., Kronenfeld, D. & Sailer, L. The problem of informant accuracy: the validity of retrospective data. *Annu. Rev. Anthropol.* **13**, 495–517 (1984).
119. Scharkow, M. The accuracy of self-reported internet use—a validation study using client log data. *Commun. Methods Measures* **10**, 13–27 (2016).
120. Konitzer, T. et al. Comparing estimates of news consumption from survey and passively collected behavioral data. *Public Opin. Q.* **85**, 347–370 (2021).
121. Prior, M. The immensely inflated news audience: assessing bias in self-reported news exposure. *Public Opin. Q.* **73**, 130–143 (2009).
122. Fletcher, R. & Nielsen, R. K. Are people incidentally exposed to news on social media? A comparative analysis. *New Media Soc.* **20**, 2450–2468 (2018).
123. Parry, D. A. et al. A systematic review and meta-analysis of discrepancies between logged and self-reported digital media use. *Nat. Hum. Behav.* **5**, 1535–1547 (2021).
124. Li, L. et al. Political-LLM: large language models in political science. Preprint at https://arxiv.org/abs/2412.06864 (2024).
125. Maier, D. et al. in *Computational Methods for Communication Science* (eds van Atteveldt, W. & Peng, T.-Q.) 13–38 (Routledge, 2021).
126. Birkenmaier, L., Lechner, C. M. & Wagner, C. The search for solid ground in text as data: a systematic review of validation practices and practical recommendations for validation. *Commun. Methods Measures* **18**, 249–277 (2023).
127. DeVerna, M. R., Yan, H. Y., Yang, K.-C. & Menczer, F. Fact-checking information from large language models can decrease headline discernment. *Proc. Natl Acad. Sci. USA* **121**, e2322823121 (2024).
128. Mohr, J. W. & Bogdanov, P. Special issue title: topic models and the cultural sciences. *Poetics* **41**, 545–770 (2013).
129. O'Callaghan, D., Greene, D., Conway, M., Carthy, J. & Cunningham, P. Down the (white) rabbit hole: the extreme right and online recommender systems. *Soc. Sci. Comput. Rev.* **33**, 459–478 (2015).
130. Grimmer, J. & Stewart, B. M. Text as data: the promise and pitfalls of automatic content analysis methods for political texts. *Polit. Anal.* **21**, 267–297 (2013).
131. Isoaho, K., Gritsenko, D. & Mäkelä, E. Topic modeling and text analysis for qualitative policy research. *Policy Stud. J.* **49**, 300–324 (2021).
132. Hoyle, A. et al. Is automated topic model evaluation broken? The incoherence of coherence. *Adv. Neural Inf. Process. Syst.* **34**, 2018–2033 (2021).
133. Heseltine, M. & Clemm von Hohenberg, B. Large language models as a substitute for human experts in annotating political text. *Res. Polit.* https://doi.org/10.1177/20531680241236239 (2024).
134. Le Mens, G. & Gallego, A. Positioning political texts with large language models by asking and averaging. *Polit. Anal.* **33**, 274–282 (2025).
135. Le Mens, G., Kovács, B., Hannan, M. T. & Pros, G. Uncovering the semantics of concepts using GPT-4. *Proc. Natl Acad. Sci. USA* **120**, e2309350120 (2023).
136. Törnberg, P., Valeeva, D., Uitermark, J. & Bail, C. Simulating social media using large language models to evaluate alternative news feed algorithms. Preprint at https://arxiv.org/abs/2310.05984 (2023).
137. Egami, N., Hinck, M., Stewart,B. & Wei H. Using large language model annotations for the social sciences: a general framework of using predicted variables in downstream analyses (2024).
138. Allcott, H. et al. The effects of Facebook and Instagram on the 2020 election: a deactivation experiment. *Proc. Natl Acad. Sci. USA* **121**, e2321584121 (2024).
139. Schmitt, J. B., Rieger, D., Rutkowski, O. & Ernst, J. Counter-messages as prevention or promotion of extremism?! The potential role of YouTube: recommendation algorithms. *J. Commun.* **68**, 780–808 (2018).
140. X developer platform https://developer.x.com/en/docs/x-api (X, 2024).
141. Stokel-Walker, C. Twitter's $42,000-per-month API prices out nearly everyone. *Wired* https://www.wired.com/story/twitter-data-api-prices-out-nearly-everyone/ (2023).
142. Geurkink, B. & Gilbert, S. Why Reddit's decision to cut off researchers is bad for its business—and humanity. *Fast Company* https://www.fastcompany.com/91014116/reddit-researchers-bad-for-business (2024).
143. Peters, J. Reddit is making sitewide protests basically impossible. *The Verge* https://www.theverge.com/2024/9/30/24253727/reddit-communities-subreddits-request-protests (2024).
144. Le, H. et al. Measuring political personalization of Google News Search. In *Proc. The World Wide Web Conference* 2957–2963 (ACM, 2019).

145. Bartley, N., Abeliuk, A., Ferrara, E. & Lerman, K. Auditing algorithmic bias on Twitter. In *Proc. 13th ACM Web Science Conference 2021* 65–73 (ACM, 2021).
146. Ohme, J. et al. Digital trace data collection for social media effects research: Apis, data donation, and (screen) tracking. *Commun. Methods. Measures* **18**, 124–141 (2024).
147. Gómez Ortega, A., Bourgeois, J. & Kortuem, G. Participation in data donation: co-creative, collaborative, and contributory engagements with athletes and their intimate data. In *Proc. 2024 ACM Designing Interactive Systems Conference* 2388–2402 (ACM, 2024).
148. Sandvig, C., Hamilton, K., Karahalios, K. & Langbort, C. Auditing algorithms: research methods for detecting discrimination on internet platforms. In *Proc. 64th Annual Meeting of the International Communication Association* 4349 (International Communication Association, 2014).
149. Özkula, S. M., Reilly, P. J. & Hayes, J. Easy data, same old platforms? A systematic review of digital activism methodologies. *Inf. Commun. Soc.* **26**, 1470–1489 (2022).
150. Horta Ribeiro, M., Hosseinmardi, H., West, R. & Watts, D. J. Deplatforming did not decrease Parler users' activity on fringe social media. *PNAS Nexus* **2**, pgad035 (2023).
151. Hobbs, W. R. & Roberts, M. E. How sudden censorship can increase access to information. *Am. Polit. Sci. Rev.* **112**, 621–636 (2018).
152. D'Amour, A. et al. Fairness is not static: deeper understanding of long term fairness via simulation studies. In *Proc. 2020 Conference on Fairness, Accountability and Transparency* 525–534 (ACM, 2020).
153. Sinha, A., Gleich, D. F. & Ramani, K. Deconvolving feedback loops in recommender systems. In *Proc. Advances in Neural Information Processing Systems* 29 (NeurIPS, 2016).
154. Lucherini, E., Sun, M., Winecoff, A. & Narayanan, A. T-RECS: a simulation tool to study the societal impact of recommender systems. Preprint at https://arxiv.org/abs/2107.08959 (2021).
155. Chaney, A. J., Stewart, B. M. & Engelhardt, B. E. How algorithmic confounding in recommendation systems increases homogeneity and decreases utility. In *Proc. 12th ACM Conference on Recommender Systems* 224–232 (ACM, 2018).
156. Zhang, Y. et al. Causal intervention for leveraging popularity bias in recommendation. In *Proc. 44th International ACM SIGIR Conference on Research and Development in Information Retrieval* 11–20 (ACM, 2021).
157. Figà Talamanca, G. & Arfini, S. Through the newsfeed glass: rethinking filter bubbles and echo chambers. *Philos. Technol.* **35**, 20 (2022).
158. Hart, W. et al. Feeling validated versus being correct: a meta-analysis of selective exposure to information. *Psychol. Bull.* **135**, 555–588 (2009).
159. Liu, N. et al. Short-term exposure to filter-bubble recommendation systems has limited polarization effects: naturalistic experiments on YouTube. *Proc. Natl Acad. Sci. USA* **122**, e2318127122 (2025).
160. Stray, J., Thorburn, L. & Bengani, P. Making amplification measurable. *Tech Policy Press* https://www.techpolicy.press/making-amplification-measurable (2023).
161. Chan, A. et al. Redefining research crowdsourcing: incorporating human feedback with LLM-powered digital twins: incorporating human feedback with LLM-powered digital twins. In *Proc. Extended Abstracts of the CHI Conference on Human Factors in Computing Systems* 1–10 (ACM, 2025).
162. Park, J. S. et al. Generative agents: interactive simulacra of human behavior. In *Proc. 36th Annual ACM Symposium on User Interface Software and Technology* 1–22 (ACM, 2023).
163. Argyle, L. P. et al. Out of one, many: using language models to simulate human samples. *Polit. Anal.* **31**, 337–351 (2023).
164. Horton, J. J., Apostolos, F. & Manning, B. S. *Large Language Models as Simulated Economic Agents: What Can We Learn from Homo Silicus*? Working Paper 31122 https://doi.org/10.3386/w31122 (National Bureau of Economic Research, 2023).
165. Wang, C., Liu, Z., Yang, D. & Chen, X. Decoding echo chambers: LLM-powered simulations revealing polarization in social networks. In *Proc. 31st International Conference on Computational Linguistics* 3913–3923 (ACL, 2025).
166. Larooij, M. & Törnberg, P. Can we fix social media? Testing prosocial interventions using generative social simulation. Preprint at https://arxiv.org/abs/2508.03385 (2025).
167. Bisbee, J., Clinton, J. D., Dorff, C., Kenkel, B. & Larson, J. M. Synthetic replacements for human survey data? The perils of large language models. *Polit. Anal.* **32**, 401–416 (2024).
168. Wang, Z. Z. et al. How well does agent development reflect real-world work? Preprint at https://doi.org/10.48550/arXiv.2603.01203 (2026).
169. Guess, A. M., Barberá, P., Munzert, S. & Yang, J. The consequences of online partisan media. *Proc. Natl Acad. Sci. USA* **118**, e2013464118 (2021).
170. Chandio, S., Dar, M. D. P. & Nithyanand, R. How audit methods impact our understanding of YouTube's recommendation systems. In *Proc. International AAAI Conference on Web and Social Media* (eds Lin, Y.-R. et al.) Vol. 18, 241–253 https://doi.org/10.1609/icwsm.v18i1.31311 (AAAI Press, 2024).
171. Metaxa, D. et al. Auditing algorithms: understanding algorithmic systems from the outside in. *Found. Trends Hum. Comput. Interact.* **14**, 272–344 (2021).
172. Bandy, J. Problematic machine behavior: a systematic literature review of algorithm audits. *Proc. ACM Hum. Comput. Interact.* **5**, 1–34 (2021).
173. Mosnar, M. et al. Revisiting algorithmic audits of TikTok: poor reproducibility and short-term validity of findings. In *Proc. 48th International ACM SIGIR Conference on Research and Development in Information Retrieval* 3357–3366 (ACM, 2025).
174. Messing, S. Are algorithmic bias claims supported? *Science* **381**, 1420 (2023).
175. National Internet Observatory https://nationalinternetobservatory.org/
176. @XDevelopers X, https://x.com/XDevelopers/status/1621026986784337922 (2023).
177. Meta Crowdtangle https://transparency.meta.com/researchtools/other-data-catalogue/crowdtangle/ (2024).
178. Meta Meta content library and api https://transparency.meta.com/en-gb/researchtools/meta-content-library/ (2025).
179. Brown, M. A. et al. Echo chambers, rabbit holes and algorithmic bias: how YouTube recommends content to real users https://doi.org/10.2139/ssrn.4114905 (2022).
180. Regulation (EU) 2022/2065 of the European Parliament and of the council on a single market for digital services (digital services act) https://eur-lex.europa.eu/eli/reg/2022/2065/oj/eng (2022).
181. Keller, D. Determining which researchers can collect public data under the DSA. *Tech Policy Press* https://www.techpolicy.press/determining-which-researchers-can-collect-public-data-under-the-dsa/ (2025).
182. Kalton, G. Methods for oversampling rare subpopulations in social surveys. *Survey Methodol.* **35**, 125–141 (2009).
183. Kleinberg, J., Mullainathan, S. & Raghavan, M. The challenge of understanding what users want: inconsistent preferences and engagement optimization. *Manage. Sci.* **70**, 6336–6355 (2024).
184. Ekstrand, M. D. et al. Recommending with, not for: co-designing recommender systems for social good. *ACM Trans. Recommender Syst.* https://doi.org/10.1145/3759261 (2025).
185. Bernstein, M. et al. Embedding societal values into social media algorithms. *J. Online Trust Safety* https://doi.org/10.54501/jots.v2i1.148 (2023).

186. Jia, C., Lam, M. S., Mai, M. C., Hancock, J. T. & Bernstein, M. S. Embedding democratic values into social media AIs via societal objective functions. *Proc. ACM Hum. Comput. Interact.* **8**, 1–36 (2024).
187. Jahanbakhsh, F. et al. Value alignment of social media ranking algorithms. In *Proc. of the 2026 CHI Conference on Human Factors in Computing Systems* article no. 1401, 1–26 https://doi.org/10.1145/3772318.3791281 (ACM, 2026).
188. Kaminskas, M. & Bridge, D. Diversity, serendipity, novelty and coverage: a survey and empirical analysis of beyond-accuracy objectives in recommender systems. *ACM Trans. Interact. Intell. Syst. (TiiS)* **7**, 1–42 (2016).
189. Patro, G. K., Biswas, A., Ganguly, N., Gummadi, K. P. & Chakraborty, A. FairRec: two-sided fairness for personalized recommendations in two-sided platforms. In *Proc. Web Conference 2020* 1194–1204 (ACM, 2020).
190. Williams, D. Scapegoating the algorithm. *Asterisk* https://asteriskmag.com/issues/11/scapegoating-the-algorithm (2025).
191. Solsman, J. E. YouTube's AI is the puppet master over most of what you watch. *CNET* https://www.cnet.com/tech/services-and-software/youtube-ces-2018-neal-mohan (2018).

Acknowledgements

D.J.W. acknowledges support from R. J. Mack, the Knight Foundation (grant ID G-2023-67859) and the Defense Advanced Research Projects Agency's (DARPA) SciFy program (agreement no. HR00112520300). We thank A. Ghasemian for helpful discussions, feedback on the paper, and for producing Fig. 1.

Author contributions

H.H. conceived the paper's structure, wrote the manuscript, and reviewed the literature. D.J.W. conceived the paper's structure and wrote the manuscript. U.D. wrote the manuscript and reviewed the literature. D.R. provided edits and feedback.

Competing interests

The authors declare no competing interests.

Additional information

Correspondence and requests for materials should be addressed to Homa Hosseinmardi or Duncan J. Watts.

Peer review information *Nature Computational Science* thanks Benjamin Lyons, Sacha Altay and Andrew M. Guess for their contribution to the peer review of this work.