Perspective

Misunderstanding the harms of online misinformation

https://doi.org/10.1038/s41586-024-07417-w

Received: 13 October 2021

Accepted: 11 April 2024

Published online: 5 June 2024

Check for updates

Ceren Budak[1], Brendan Nyhan[2], David M. Rothschild[3]✉, Emily Thorson[4] & Duncan J. Watts[5]

The controversy over online misinformation and social media has opened a gap between public discourse and scientific research. Public intellectuals and journalists frequently make sweeping claims about the effects of exposure to false content online that are inconsistent with much of the current empirical evidence. Here we identify three common misperceptions: that average exposure to problematic content is high, that algorithms are largely responsible for this exposure and that social media is a primary cause of broader social problems such as polarization. In our review of behavioural science research on online misinformation, we document a pattern of low exposure to false and inflammatory content that is concentrated among a narrow fringe with strong motivations to seek out such information. In response, we recommend holding platforms accountable for facilitating exposure to false and extreme content in the tails of the distribution, where consumption is highest and the risk of real-world harm is greatest. We also call for increased platform transparency, including collaborations with outside researchers, to better evaluate the effects of online misinformation and the most effective responses to it. Taking these steps is especially important outside the USA and Western Europe, where research and data are scant and harms may be more severe.

Social media platforms have become ubiquitous in modern life, prompting great interest in their effects on society. Debates over social media's impact often include claims that social media platforms and algorithms facilitate high levels of exposure to misinformation and other harmful content[1], which in turn cause social problems ranging from polarization to political violence[2,3]. Broad causal claims like these appear frequently in influential public discourse about social media, including in articles written by prominent social scientists[2]. For example, a widely discussed article in *The Atlantic* argues that "social media amplifies political polarization; foments populism, especially right-wing populism; and is associated with the spread of misinformation"[2]. Similarly, *The New York Times* op-eds have argued that "YouTube may be one of the most powerful radicalizing instruments of the 21st century" and that "The dominant digital platform companies, including Facebook and Google … have created a haven for dangerous misinformation and hate speech that has undermined trust in democratic institutions"[4,5]. These concerns are echoed by both politicians[6] and members of the US public[7–9], who say that social media has a negative impact on society and cite misinformation as a primary reason[10].

Assertions like these suggest that exposure to false content on social media has been shown to cause massive social harms. However, we do not believe that compelling empirical evidence exists supporting these claims. The question of whether and when exposure to this type of content on social media causes harm is complex, a matter of active debate among researchers and extremely difficult to conclusively answer. Although disjunctures between public discourse and scholarly work are a problem in many fields[11–14], we argue that unsupported or disputed claims about widespread exposure to (and effects of) misinformation on social media are particularly concerning because they can shape the actions of platforms, legislators and regulators, diverting attention from other ways in which misinformation can cause harm.

Before proceeding, we note that this Perspective focuses strictly on the potential effects of exposure to false and inflammatory content on social media. We acknowledge that reasons for concern exist about other potential harms from social media such as negative effects on mental health and well-being[15,16] and social trust[17,18] (see Table 1 for a summary or ref. 19 for a review). Although these other potential harms are also important, concerns about misinformation have been a focus of much of the public discourse around social media. In addition, narrowing our focus allows for a more thorough review of the relevant literature.

The public debate that we engage with concerns the effects on mass opinion of exposure to misinformation and extremist (for example, white supremacist, alt-right) content on social media. That debate often centres on three claims that are largely unsupported by academic research: (1) average exposure to misinformation is high and growing; as a result, a substantial fraction of the population is exposed to it frequently (for example, refs. 20–24); (2) exposure to this content is primarily driven by the platforms' algorithms rather than by individual users deliberately seeking out such content (for example, refs. 1,3,25–28); and (3) correlations between exposure to false and extremist content on social media platforms and undesirable

[1]University of Michigan School of Information, Ann Arbor, MI, USA. [2]Department of Government, Dartmouth College, Hanover, NH, USA. [3]Microsoft Research, New York, NY, USA. [4]Maxwell School of Citizenship and Public Affairs, Syracuse University, Syracuse, NY, USA. [5]Department of Computer and Information Science, Annenberg School of Communication, and Operations, Information, and Decisions Department, University of Pennsylvania, Philadelphia, PA, USA. ✉e-mail: david@researchdmr.com

psychological or behavioural outcomes reflect causal effects (for example, refs. 2,29–32).

These claims (often assumed rather than asserted explicitly) influence public debate, diverting attention from other potential sources of harm. For example, following the 2016 US presidential election, public attention (including congressional hearings) quickly centred on the possibility that Russia had used social media platforms to shape the beliefs and opinions of large numbers of US voters, in turn affecting their vote choice[33]. But academic research subsequently showed that exposure to these claims represents a tiny part of people's information diets and is not associated with changes in voter attitudes or behaviour[34]. Moreover, this outsize focus on mass persuasion ignored the skew in on-platform information diets that the Russian influence operations helped to reveal. That is, exposure to tweets from Russian influence accounts was disproportionately higher among Republicans and was concentrated in a small set of users[35]—just 1% were responsible for 70% of exposures, and 10% were responsible for 98% of exposures[34]. Other examples show a similar pattern: exposure to problematic content is rare in general and is heavily concentrated among a small minority of people who already have extreme views (for example, refs. 36,37).

Similarly, public discourse about social media has been shaped by the assumption that platform algorithms are the primary cause of exposure to false and extremist content online. For example, a 2022 article in *The New York Times* begins by stating that "It is well known that social media amplifies misinformation and other harmful content" and then proceeds to cite non-academic work by an advocacy group[1]. Broad assertions like these neglect a growing body of research on consumer demand for false and extremist content[36,38], the role of media and political elites in exposing people to misinformation[39,40], and how platform affordances enable the distribution of such content to subscribers and followers[37,39,41,42]. In short, although algorithms indisputably shape the content people see, we interpret recent empirical evidence as suggesting that, on average, these algorithms tend to push users to more moderate content and to offer extreme content predominantly to those who have sought it out[36,38,43]. Increasing public attention to this research will help to foster a broader conversation about other mechanisms of misinformation dissemination and harm besides algorithms.

Finally, social media is particularly vulnerable to the human tendency to confuse correlation with causation. Surveys show that US citizens blame social media for the spread of misinformation, political incivility and even political violence[7–9]. This tendency may reflect the temporal association between social media usage levels and negative social trends of the past two decades as well as the way in which social media content refracts societal ills. However, much of the research explicitly designed to identify causal effects does not support these claims[15,44–46].

These misunderstandings are often exacerbated by data limitations, such as a reliance on measures of sharing and other forms of engagement rather than direct measures of exposure to social media posts. These proxies can be misleading because engagement is disproportionately concentrated among a subset of social media users and is centred on specific types of content that may be unrepresentative of the larger set of content to which people are exposed[47]. We therefore advocate data sharing by social media companies that will allow for (1) better tracking of aggregate exposure to harmful content as well as the extent to which this exposure is concentrated among small groups of people who consume large amounts of false and extremist content, and (2) experiments designed to measure the role of platform features in facilitating exposure to harmful content, especially among heavy consumers. These steps are especially important to take outside the USA and Western Europe in countries that have different media ecosystems (for example, fewer publishers and more censorship), face technical challenges associated with detecting problematic content in languages other than English (especially those that are less common) and/or suffer from a lack of investment by the platforms in limiting extreme content.

Table 1 | Possible harms from social media

Mechanism of harm	Example effects of concern	In this paper
Misinformation	False beliefs increase[145,146]; vaccination intention decreases[147]	Yes
Extremist content	Hate crimes increase[127,148]	Yes
Social comparisons	Subjective well-being decreases[16]; poor mental health increases[17]	No
Divisive content	Polarization on policy issues increases[16,149]	No
Alarmist/sensational content	Media trust decreases[18]; interpersonal/institutional trust decreases[19]	No

The findings above are provided as examples of findings of concern and do not include contrary findings or correlational evidence. See ref. 150 for a review of evidence on social media effects.

As a result, the concerns addressed in this paper may be magnified. By working collaboratively, scientists and platform companies can reduce misunderstandings about the harms of social media and help to identify and respond to the most serious threats around the world.

Misunderstanding social media harms

Overstating exposure to harmful content

We begin by describing four ways in which exposure to potentially harmful false and extremist content on social media tends to be overstated in public discourse. These are summarized in Table 1 and explained in more detail below. As stated above, our focus is exposure to false and inflammatory content on social media, not other potential mechanisms of harm (Table 2).

First, public commentators regularly assume or assert high levels of exposure to potentially harmful content on social media[21,48]. The tendency to over-emphasize misleading statistics (often with little context) is, of course, not unique to the topic of social media[11,12]; coverage of science is frequently distorted by media sensationalism[13] and hype from academic press releases[14]. This distortion happens for various reasons. One important mechanism is the loss of nuance as academic work gets translated into the public sphere. Indeed, even when scholars carefully state the scope of their work (for example, refs. 49,50), this nuance is often lost in media coverage (for example, ref. 51). In the realm of social media, the problem frequently takes the form of a focus on seemingly impressive statistics that fail to take into account the vast scale of content exposure on platforms, the small share of people's information diets that misinformation and extremist content typically represent, and the way that exposure to such content is concentrated among small minorities of users[20–24,34,36,37,52–66]. Of course, even small percentages can translate to meaningful quantities on platforms with billions of users, but the patterns we document can lead to systematic misunderstandings of the scale and nature of the problem and thus divert attention from more pressing threats.

The statistics invoked in public debates about social media also often fail to take into account the volume of content that platforms serve. For example, a 2020 *The New York Times* story on a deceptively edited video of Joe Biden noted that it was "viewed more than 17 million times on social media platforms … [including] hundreds of thousands of views [on Facebook] … [and] more than 800,000 times [on YouTube]"[52]. Similarly, a recent study of deepfake videos in the wild noted that there were 31 on YouTube with more than 500,000 views[53].

These statistics sound large until one considers the scale of the social media platforms in question. On Facebook, for instance, the 20 most widely viewed posts in the USA in the first quarter of 2023 had 776.3 million views on Facebook alone—far more than the Biden video—and yet, in total, represented just 0.04% of views of content on the platform in the USA during the quarter[54]. Facebook likewise reported that content

Table 2 | Four ways public discourse overstates exposure to potentially harmful content

Overstating mass exposure to potentially harmful content
(1) Statistics reporting aggregate-level exposure to harmful content: provided without appropriate population denominators
(2) Statistics reporting individual-level exposure to harmful content: provided without appropriate denominators for total individual-level exposure
(3) Statistics reporting average individual-level exposure to harmful content: skewed by extreme values in small portion of population
(4) Statistics reporting levels of engagement (a public act) with harmful content: may not be representative of levels of exposure (a private act)

made by Russian trolls from the Internet Research Agency reached as many as 126 million US citizens on Facebook before the 2016 US presidential election—a statistic that was widely cited in the press and in studies on Internet Research Agency operations[55,56]. Far less attention was paid to the estimate that such content represented 0.004% of the content that US citizens saw in the Facebook news feed[57]. We acknowledge that these figures are huge in absolute terms, but in context, we believe that their effects are likely to be small given that they represent a tiny proportion of total information flows on the platforms[58]. Citing these absolute numbers may contribute to misunderstandings about how much of the content on social media is misinformation[59,60]: for example, US citizens estimate that 65% of the news they see on social media is misinformation[61].

In addition, although the recent spike in research and coverage focusing on 'fake news' and misinformation might seem to suggest that this content makes up a large share of the media diets of US citizens, the reality is quite different when we consider another frequently neglected denominator—the total amount of news that people consume. Previous research suggests that sustained repetition may be required to generate even fleeting media effects and that such effects may be limited when people can choose to opt out of news altogether[67–69]. Competition from other frames can also decrease the salience of the messages people receive and diminish any persuasive effects[70,71]. For all of these reasons, it is noteworthy that exposure to false, untrustworthy or other forms of potentially harmful online content is actually quite infrequent compared with other news sources. For example, articles from the 490 websites identified as untrustworthy by ref. 62 made up only 5.9% of US citizens' visits to news sites on average in the period before and immediately after the 2016 US election[63]. Similarly, cross-national data show that untrustworthy websites made up only 0.1–4.4% of web traffic on average across the USA, UK, France and Germany for the 2017–2021 period[64]; untrustworthy sources made up just 6.7% of political URLs seen on Twitter during the 2016 US presidential campaign[62]; fewer than 10% of YouTube viewers ever saw an extremist channel video in autumn 2020[37]; and far-left and far-right sources together made up less than 0.5% of total YouTube watch time from 2016 to 2019[36]. When we expand the denominator to include television in addition to online content, fake news websites represent just 0.1% of US citizens' media diet[65]. The studies cited here use different methods to identify untrustworthy (or extremist) content. For instance, refs. 62,64 used 490 websites that were identified using lists published by fact-checkers as well as a manual inspection of Snopes.com to determine domains that publish questionable claims. Reference 63 relies on NewsGuard's assessment of the trustworthiness of 3,592 web domains. The papers also differ in the denominator used to measure prevalence. For instance, ref. 63 relies on the hard news classification provided by ref. 72, whereas ref. 62 uses all URLs included in tweets labelled as being political. These choices can affect prevalence estimations[73] but the pattern of results is generally consistent.

A third problem is that statistics on average exposure levels like those described above disregard the important reality that exposure is concentrated among a small fraction of the population. For example, the 20% of US citizens with the most conservative information diets were responsible for 62% of visits to the 490 untrustworthy websites described above during the 2016 campaign[63]. Similarly, 6.3% of YouTube users were responsible for 79.8% of exposure to extremist channels from July to December 2020[37], 85% of vaccine-sceptical content was consumed by less than 1% of US citizens in the 2016–2019 period[66], and 1% of Twitter users were responsible for 80% of exposures to links from dubious websites during and immediately after the 2016 US presidential campaign[62]. Although these studies draw on different data, they reach similar conclusions. Across numerous data sources, it appears that the typical social media user is exposed to even less dubious or extreme content than the aggregate exposure statistics above might seem to suggest (a pattern that we expect will hold around the world but should be tested in future research). In addition, given how quickly effects on attitudes and beliefs tend to decay[74–76], such harms are less likely to accumulate and persist for typical news consumers relative to the small minorities of people for whom exposure is more frequent and intense. Finally, many of the statistics that currently inform public debate blur the distinction between exposure to misinformation and engagement with it, often drawing data from 'leaderboards' of content with high engagement rates (such as likes, shares and comments). For example, coverage of the highest-engagement content from Facebook pages is often described as showing what is 'popular'[77–79]. The conflation of engagement with consumption is understandable, as the two sound similar and engagement statistics are easier to access from public sources (for example, CrowdTangle). By contrast, platforms rarely share aggregate statistics about exposure. Engagement is an important metric as it can affect what other people see via algorithmic recommendations, especially for connected nodes in a social graph. However, engagement statistics can provide a biased and often misleading view of the content that people actually consume. Engagement measures generally reflect activity that is public, whereas the vast majority of consumption is private[47]. The decision to publicly engage with content is partly a strategic one that depends on anticipated audience perception[80] and, as a result, the content that people publicly engage with is systematically different to the content that they privately consume. For example, publicly shared URLs on Facebook are more likely to be false news than those not publicly shared, according to one recent study in the USA, which found that 7.0% of clicks on URLs that have been publicly shared 100 or more times are to 'fake news' websites compared with 2.5% of clicks in representative data[47]. Similarly, researchers studying the spread of misinformation have found that partisans' decisions about what information to share are affected by their perception of reputational costs and benefits[81,82]. In other words, whether our goal is to understand exposure to misinformation or belief in it, sharing behaviour is a suboptimal proxy.

Although the studies described above have improved the measurement of important quantities (for example, by focusing on exposure instead of sharing), they still face limitations. Most notably, owing to the difficulty of identifying online misinformation at scale, most studies measuring exposure in behavioural data rely on source-level indicators of trustworthiness or extremity and typically lack data on on-platform exposure, especially in mobile browsers and apps[34,36,37,44,62–65]. Even more granular measures of misinformation exposure are necessarily incomplete owing to the capacity limitations of, for example, Meta's third-party fact-checking partners[44,45,83]. Use of source-level measures can lead to overestimation of total exposure to false content because only a subset of content published by untrustworthy sites includes misinformation[84]. However, the lack of comprehensive content coverage by fact-checkers can lead to underestimation of total exposure (by failing to identify all of the untrustworthy sites and/or missing specific false articles from trustworthy sites).

Despite these limitations, we believe that the conclusions of the academic research are clear—exposure to misinformation is low as a

percentage of people's information diets and concentrated among a small minority. However, this reality is not reflected in public discourse about social media. The prevalence of this 'misinformation about misinformation' can have downstream effects on the steps taken by both the public and the platforms to address the potential harms of social media.

Exaggerating the effects of algorithms

Explanations of exposure to potentially harmful content often focus on algorithms, neglecting the research literature that finds relatively stronger demand effects compared with algorithmic effects on consumption of harmful content[36,38,43].

Algorithms are frequently blamed in public discourse for trapping users in online 'filter bubbles' and promoting extreme content[4,85,86]. The reasoning of claims like those in refs. 4,85,86 is that algorithmic recommendations seek to promote user engagement; like-minded and/or inflammatory content is more engaging than average; and algorithms therefore differentially promote it to users. It is not surprising that such beliefs have become widespread given that social media algorithms are both opaque and poorly understood by the public[87,88].

In reality, the existence of large algorithmic effects on people's information diets and attitudes has not been established[89,90]. The most comprehensive study we know of is a randomized large sample of consenting Facebook and Instagram users to a reverse-chronological feed rather than one that was algorithmically ranked[45]. The results on information exposure were mixed: the algorithmic feed showed people more content from politically like-minded sources than the reverse-chronological feed did but less content from untrustworthy sources. Moreover, being switched out of an algorithmically ranked feed for three months had no measurable effect on a variety of political attitudes. Although it is of course possible that a more sustained intervention may have different effects, these results suggest that algorithms typically show people content from accounts they chose to follow or content in which they indicated interest. As a result, exposure levels may reflect people's preferences more than they shape them[38]. Similarly, a recent study found that bots trained on real people's YouTube video view history who followed the site's algorithmic recommendations viewed less partisan content than the actual humans did[43].

By contrast, behavioural data indicates that people who consume a great deal of false, untrustworthy or otherwise potentially harmful content are often already highly attentive to this content and seek it out across mediums[36]. For example, recent evidence from the deplatforming of Parler, a far-right social media site, in the aftermath of the 6 January 2021 insurrection at the US Capitol found that even this dramatic intervention had no discernible effect on the overall consumption of fringe content, which simply shifted to other, similar sites[91]. Audience demand is also more important than algorithmic amplification in driving exposure to fringe content on YouTube among US audiences[36,37]. For instance, people who consume videos from extremist channels on YouTube in the USA are more likely to have previously expressed high levels of hostile sexism and racial resentment[37]. Many actively seek out extremist content by following links from alternative and fringe social media sites such as 4chan and/or subscribing to the channel in question[37]. Similarly, 41% of views of videos classified in previous research as far-right on YouTube took place after people followed a link from an external URL and 36% were directly preceded by another video on the site[36]. By contrast, only 0.4% of algorithmic recommendations directed users to extremist channel videos on YouTube[37]. These findings highlight the need for scrutiny of social media features besides algorithms as well as the systemic factors that drive people towards extremism.

Of course, none of these findings demonstrate that algorithmic effects do not, or cannot, exist. Indeed, simulation-based studies demonstrate how social media platforms can recommend problematic content[92] and controlled experiments suggest algorithmic recommendations can alter the relative prevalence of conservative versus liberal information[93], limit exposure to counter-attitudinal news[94] and have negative attitudinal effects[95]. In addition, methodological challenges limit our ability to disentangle the effects of audience demand from algorithms given the extent to which algorithmic recommendations themselves are shaped by feedback loops[96,97].

In short, algorithms can have important effects (both directly and in how they interact with human behaviour), but we argue that (1) the evidence of large-scale algorithmic effects on public attitudes and behaviour is more limited than what is reflected in the public discourse and (2) the role of audience demand has been neglected in public discourse relative to the attention paid to algorithms.

Causal claims about social media effects

Public discourse frequently blames social media usage or content for negative social trends based on correlational evidence[2,32]. These kinds of relationships often seem obvious: the increase in social media usage visibly coincides with various shifts over the past 20 years (for example, political polarization), leading many to intuit a causal relationship between them[2,98,99]. Moreover, the content circulating online inevitably reflects those trends in some manner, creating striking and memorable manifestations of phenomena that people deplore ranging from political incivility to anti-vaccine activism[100–104]. The public discourse may confidently claim that social media is the primary cause of these trends. However, evidence from social science research paints a more nuanced picture.

Examples of public discourse blaming social media for larger social problems abound. For instance, a 2022 *The Atlantic* article claimed that over the past decade, "Facebook, Twitter, YouTube, and a few other large platforms unwittingly dissolved the mortar of trust, belief in institutions, and shared stories that had held a large and diverse secular democracy together"[2]. But this claim reflects a series of correlational relationships. Polarization and social media use have both increased over the past few decades[98,99]. It is also true that people who use social media more often express higher levels of affective polarization or out-party animosity, especially when they are embedded in networks with high levels of homophily[31,100,105–107]. Such correlations are often taken as strong evidence for a causal relationship, but the best-designed social media deactivation studies so far find that exogenous decreases in Facebook use had no measurable effect on affective polarization in the USA (although issue polarization decreased) and led to less, not more, positive evaluations of outgroup members in Bosnia and Herzegovina[15,46]. Experiments varying individual features of the platform have also yielded mixed results: although increasing exposure to cross-cutting news decreased affective polarization[94], decreasing exposure to content from like-minded sources had no effect[44]. Panel data similarly suggest that the causality goes in the opposite direction—affective polarization predicts media use, not the other way around[108].

Similarly, the *gilets jaunes* (yellow vests) movement in France was described as "a beast born entirely from Facebook" in the press[101] based on the way the platform was used to organize protests; and a 2021 *Forbes* article claimed that "[a]nti-vaxxers and the misinformation they spread on social media caused vaccination rates to drop in the US and the UK" owing to the correspondence between social media use and vaccination rates[102]. To be sure, the *gilets jaunes* did use Facebook to coordinate their actions and anti-vaccine advocates do use social media to spread misinformation[104]. In a very narrow sense, then, these outcomes could not have occurred exactly as they did without the platforms. But the causal question of interest requires consideration of a more difficult counterfactual: whether social media sites like Facebook cause violent protests or anti-vaccine movements to be more severe or prevalent than they would have been in their absence. The answer to this question is unclear. France has a long history of street protests and demonstrations dating back to the French Revolution[109]. Likewise, the modern anti-vaccine movement long predates social media, dating back centuries to the advent of vaccines themselves[110]. As these

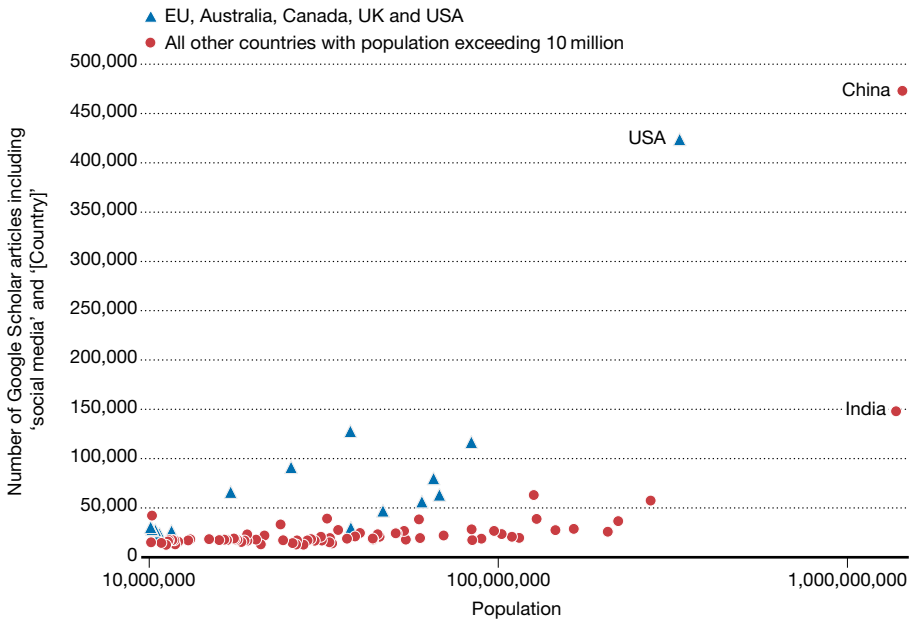

Fig. 1 | Academic research on the use of social media by country. Google Scholar statistics derived from searches of keywords in academic articles. The following searches were conducted for the 1 January 2018 to 9 June 2022 period: 'social media' and '[Country]'. The figure includes all 91 countries with populations of 10 million or more people. Countries in the European Union, Australia, Canada, the UK and the USA are represented as blue triangles; all other countries with a population exceeding 10 million are represented as red circles.

examples demonstrate, although social media may alter the strategic choices of these actors, its existence is not a necessary condition for them to achieve their goals.

Many influential claims in public discourse about the widespread harms of social media rest on correlational evidence (for example, refs. 2,4,86,111). Although some extant studies leverage plausibly exogenous shocks to generate more credible estimates of the causal impacts of social media or its features (for example, refs. 112–114), these opportunities are rare, and thus are the exception rather than the rule. Demonstrating that prominent causal claims about social media harms are true would require new research designs such as experiments in which access to the platform or to specific features of the platform are varied systematically and the results reliably measured. For the same reason, trying to estimate the state of the world absent an entire platform is very difficult; questions about the aggregate effect of social media on society are probably unanswerable (as with, say, television).

None of these inferential challenges mean that social media has no harmful effects. Like any widely used technology, social media platforms are almost certainly responsible for some problems in the world, in part because of their susceptibility to exploitation by bad actors, and should take steps to make positive changes. It is also important to acknowledge the indirect ways in which social media can lead to harms. For instance, the mere availability of misinformation on social media, even if small, could potentially lead to a reduction in public trust, especially when the scale of the problem is exaggerated. If true, however, then it is arguably even more important to correct such misperceptions in public discourse. Similarly, the increasing popularity of social media as a means of news access could theoretically lead to changes in the behaviour of news producers (for example, an increase in sensational or hyper-partisan content). Furthermore, the problems created by social media may have been more severe before the steps taken by platforms such as Facebook (after the 2016 US election) and YouTube (in 2019) to address external criticism. It is, therefore, vital to conduct credible research studies estimating the causal effects of platform features or other interventions on problematic behaviours.

Lack of non-Western data

With a few notable exceptions (for example, refs. 115,116), most research on misinformation focuses on the Global North[117]. Figure 1 shows how dominant the European Union, Australia, Canada, the UK and the USA are in research about social media versus comparable countries by population. (The key exception is China, which has a different set of social media companies.) As such, our discussion has largely focused on the USA and Western Europe. The limitations of the evidence base unfortunately constrain our ability to provide examples of the issues noted above in other regions. Treating the USA and other Western countries as the default may lead to important and often unacknowledged limitations in research findings.

In particular, although few systematic comparisons have been made, it is plausible that social media's effects may be larger in non-Western contexts. First, misleading online content may reach more people in countries that lack reliable mainstream news outlets, put limitations on media freedom and/or have low levels of trust in the media[118,119]. Social media moderation decisions can also be influenced by such regimes, further exacerbating the problem[120], and expert surveys indicate that government disinformation has grown differentially in autocratizing countries versus democratized ones[121]. In addition, the subsidized access provided by Facebook's 'free basics' plan may also increase the prevalence of social media content in users' information diets by limiting exposure to other sources of news[122,123]. Platforms also devote fewer resources to content moderation in the Global South. In 2020, for instance, just 13% of the 3.2 million hours Facebook spent searching out, labelling and taking down false or misleading content was outside the USA[123]. Facebook offers its services in 110 languages but had people reviewing content in just 70 languages and has published its 'community standards' in just 50 as of 2021[124]. Technological approaches to detecting various types of potentially harmful content, such as natural language processing, are also likely to perform less well in low-resource languages[125].

As a result, we expect higher levels of false and extremist content in people's information diets outside of the USA and Western Europe. In these contexts, it seems that the platforms make less of an effort

Box 1

Recommendations for improving public discourse about social media through evidence-based research

Measure exposure and mobilization among extremist fringes. Platforms and academic researchers should identify metrics that capture exposure to false and extremist content not just for the typical news consumer or social media user but also in the fringes of the distribution. Focusing on tail exposure metrics would help to hold platforms accountable for creating tools that allow providers of potentially harmful content to engage with and profit from their audience, including monetization, subscriptions, and the ability to add members and group followers.

Reduce demand for false and extremist content and amplification of it by the media and political elites. Audience demand, not algorithms, is the most important factor in exposure to false and extremist content. It is therefore essential to determine how to reduce, for instance, the negative gender- and race-related attitudes that are associated with the consumption of content from alternative and extremist YouTube channels[37]. We likewise must consider how to discourage the mainstream press and political elites from amplifying misinformation about topics such as COVID-19[40,151] and voter fraud in the 2020 US elections[152].

Increase transparency and conduct experiments to identify causal relationships and mitigate harms. Social media platforms are increasingly limiting data access[153] even as increased researcher data and API access is needed to enable researchers outside the platforms to more effectively detect and study problematic content. Platform-scale data are particularly necessary to study the small groups of extremists who are responsible for both the production and consumption of much of this content. When public data cannot be shared due to privacy concerns, the social media platforms could follow the 'clean room' model used to allow approved researchers to examine, for example, confidential US Census microdata data in secure environments[154]. These initiatives should be complemented by academic–industry collaborations on field experiments, which remain the best way to estimate the causal effects of social media, with protections including review by independent institutional review boards and preregistration to ensure that research is conducted ethically and transparently.

Fund and engage research around the world. It is critical to measure exposure to potentially harmful content in the Global South and in authoritarian countries where content moderation may be more limited and exposure to false and extremist content on social media correspondingly more frequent. Until better data are available to outside researchers, we can only guess at how best to reduce the harms of social media outside the West. Such data can, in turn, be used to enrich fact-checking and content moderation resources and to design experiments testing platform interventions.

to limit and tag potentially harmful content, act less aggressively to counter bad actors and use algorithms that are likely to perform worse due to language differences. The risks of harms are thus concomitantly greater; platform companies can and should increase their efforts to minimize harm in non-Western contexts (although the concerns we raise about counterfactual inference still apply).

Discussion

Our review of research on the harms of online misinformation leads us to conclude that exposure to misinformation and extremist content is not frequent; instead, it is relatively rare and highly concentrated among small groups of extremists. We argue that the many claims to the contrary neglect appropriate denominators, are skewed by high levels of exposure in fringe groups or reflect engagement rather than exposure. In addition, the debate about exposure to potentially harmful content on social media overstates the role of algorithms, neglecting the powerful role of audience demand. Finally, we suggest that public discourse too often asserts causal associations between social media content or exposure and outcomes such as polarization without providing or citing credible evidence of such a relationship. We note, however, that the evidence base is limited and heavily concentrated in the USA and Western Europe—far more research must be conducted in the rest of the world, where there is often greater reason for concern.

Our conclusions have important policy implications for researchers, platforms and civil society (Box 1) First, we should focus more attention on measuring exposure to potentially harmful content in the tails of the distribution and among extremist or fringe groups. Previous research provides worrying evidence linking social media to hate crimes[126,127] and civil unrest[103]. Without better data on misinformation exposure among fringe and extremist groups, designing effective interventions to prevent such effects or holding platforms accountable for failing to do so is impossible. Second, we need to determine how to most effectively limit demand for false and extremist content and the amplification of it by political elites and journalists, who often spread viral falsehoods to far larger audiences than the ones they reach directly online. Traditional news, and in particular television news, still dominates people's news consumption[65] and political elites seek to shape that news coverage[128]. As a result, the mainstream media are a key mechanism for exposing broad audiences to false claims, which often originate with political elites. This exposure can have harmful effects (for example, refs. 129,130). Both scholarly research and civil society should thus recognize the critical role of elites and traditional news media in spreading misinformation. Platforms need to likewise determine how to prevent misinformation from those sources from spreading more effectively[131]. Social media companies must also expand researcher access to platform data and application programming interfaces (APIs) and enable academic–industry collaborations on field experiments that enable assessment of the harms resulting from misinformation on their platforms.

We recognize that greater transparency and data disclosure raises privacy concerns and requires difficult trade-offs[132,133]. However, such concerns must be balanced against the societal need to identify and mitigate platform harms. We are confident that the combination of privacy-protecting approaches to sharing aggregate data and secure data access facilities can manage these risks and address key stakeholder concerns appropriately[132,134]. When possible, independent non-profit organizations and government agencies should help to facilitate data access and researcher partnerships, providing protections to scholars and creating more equitable terms of access. Legislative proposals in the USA (for example, the Platform Accountability and Transparency Act and Social Media Data Act) and existing laws in Europe (for example, article 31 of the Digital Services Act) include provisions for social media data access by researchers. These proposals, as varied as they

are, also demonstrate that a balance can be struck between the need to protect user data and the need for research on social media effects (that is, by enforcing restrictions on who gets access to what types of data and for which types of research goal)[135]. In addition, transparency mechanisms such as preregistration can help to promote trust in the resulting research by reducing the ability of platforms to influence what results are reported. Finally, it is critically important to expand research and monitoring into online misinformation outside the USA and Western Europe. Although platforms and the scientific community claim to serve the world, both research[136–140] and platforms[141–144] have always distributed resources inequitably. Increased investment in content moderation and misinformation research in non-Western contexts could help to offset historical inequalities in the digital world and reduce the threat where the potential harms are greatest[145,146].

1. 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 (13 October 2022).
2. Haidt, J. Why the past 10 years of American life have been uniquely stupid. *The Atlantic*, https://www.theatlantic.com/magazine/archive/2022/05/social-media-democracy-trust-babel/629369/ (11 April 2022).
3. 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/ (28 July 2022).
4. Tufekci, Z. YouTube, the great radicalizer. *The New York Times*, https://www.nytimes.com/2018/03/10/opinion/sunday/youtube-politics-radical.html (10 March 2018).
5. Romer, P. A tax that could fix big tech. *The New York Times*, https://www.nytimes.com/2019/05/06/opinion/tax-facebook-google.html (6 May 2019).
6. Schnell, M. Clyburn blames polarization on the advent of social media. *The Hill*, https://thehill.com/homenews/sunday-talk-shows/580440-clyburn-says-polarization-is-at-its-worst-because-the-advent-of/ (7 November 2021).
7. *Robert F. Kennedy Human Rights/AP-NORC Poll* (AP/NORC, 2023).
8. Goeas, E. & Nienaber, B. *Battleground Poll 65: Civility in Politics: Frustration Driven by Perception* (Tarrance Group, 2019).
9. Murray, M. Poll: Nearly two-thirds of Americans say social media platforms are tearing us apart. *NBC News*, https://www.nbcnews.com/politics/meet-the-press/poll-nearly-two-thirds-americans-say-social-media-platforms-are-n1266773 (2021).
10. Auxier, B. 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).
11. Koomey, J. G. et al. Sorry, wrong number: the use and misuse of numerical facts in analysis and media reporting of energy issues. *Annu. Rev. Energy Env.* **27**, 119–158 (2002).
12. Gonon, F., Bezard, E. & Boraud, T. Misrepresentation of neuroscience data might give rise to misleading conclusions in the media: the case of attention deficit hyperactivity disorder. *PLoS ONE* **6**, e14618 (2011).
13. Copenhaver, A., Mitrofan, O. & Ferguson, C. J. For video games, bad news is good news: news reporting of violent video game studies. *Cyberpsychol. Behav. Soc. Netw.* **20**, 735–739 (2017).
14. Bratton, L. et al. The association between exaggeration in health-related science news and academic press releases: a replication study. *Wellcome Open Res.* **4**, 148 (2019).
15. Allcott, H., Braghieri, L., Eichmeyer, S. & Gentzkow, M. The welfare effects of social media. *Am. Econ. Rev.* **110**, 629–676 (2020).
16. Braghieri, L., Levy, R. & Makarin, A. Social media and mental health. *Am. Econ. Rev.* **112**, 3660–3693 (2022).
17. Guess, A. M., Barberá, P., Munzert, S. & Yang, J. The consequences of online partisan media. *Proc. Natl Acad. Sci. USA* **118**, e2013464118 (2021).
18. Sabatini, F. & Sarracino, F. Online social networks and trust. *Soc. Indic. Res.* **142**, 229–260 (2019).
19. Lorenz-Spreen, P., Lewandowsky, S., Sunstein, C. R. & Hertwig, R. How behavioural sciences can promote truth, autonomy and democratic discourse online. *Nat. Hum. Behav.* **4**, 1102–1109 (2020).
This paper provides a review of possible harms from social media.
20. Lapowsky, I. The mainstream media melted down as fake news festered. *Wired*, https://www.wired.com/2016/12/2016-mainstream-media-melted-fake-news-festered/ (26 December 2016).
21. Lalani, F. & Li, C. *Why So Much Harmful Content Has Proliferated Online—and What We Can Do about It* Technical Report (World Economic Forum, 2020).
22. Stewart, E. America's growing fake news problem, in one chart. *Vox*, https://www.vox.com/policy-and-politics/2020/12/22/22195488/fake-news-social-media-2020 (22 December 2020).
23. Sanchez, G. R., Middlemass, K. & Rodriguez, A. *Misinformation Is Eroding the Public's Confidence in Democracy* (Brookings Institution, 2022).
24. Bond, S. False Information Is Everywhere. 'Pre-bunking' Tries to Head It off Early. *NPR*, https://www.npr.org/2022/10/28/1132021770/false-information-is-everywhere-pre-bunking-tries-to-head-it-off-ear (National Public Radio, 2022).
25. Tufekci, Z. Algorithmic harms beyond Facebook and google: emergent challenges of computational agency. *Colo. Tech. Law J.* **13**, 203 (2015).
26. Cohen, J. N. Exploring echo-systems: how algorithms shape immersive media environments. *J. Media Lit. Educ.* **10**, 139–151 (2018).
27. Shin, J. & Valente, T. Algorithms and health misinformation: a case study of vaccine books on Amazon. *J. Health Commun.* **25**, 394–401 (2020).
28. 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).
29. Pauwels, L., Brion, F. & De Ruyver, B. *Explaining and Understanding the Role of Exposure to New Social Media on Violent Extremism. an Integrative Quantitative and Qualitative Approach* (Belgian Science Policy, 2014).
30. McHugh, B. C., Wisniewski, P., Rosson, M. B. & Carroll, J. M. When social media traumatizes teens: the roles of online risk exposure, coping, and post-traumatic stress. *Internet Res.* **28**, 1169–1188 (2018).
31. Soral, W., Liu, J. & Bilewicz, M. Media of contempt: social media consumption predicts normative acceptance of anti-Muslim hate speech and Islamo-prejudice. *Int. J. Conf. Violence* **14**, 1–13 (2020).
32. Many believe misinformation is increasing extreme political views and behaviors. *AP-NORC* https://apnorc.org/projects/many-believe-misinformation-is-increasing-extreme-political-views-an (2022).
33. Fandos, N., Kang, C. & Isaac, M. Tech executives are contrite about election meddling, but make few promises on Capitol Hill. *The New York Times*, https://www.nytimes.com/2017/10/31/us/politics/facebook-twitter-google-hearings-congress.html (31 October 2017).
34. Eady, G., Paskhalis, T., Zilinsky, J., Bonneau, R., Nagler, J. & Tucker, J. A. 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).
This paper shows that exposure to Russian misinformation on social media in 2016 was a small portion of people's news diets and not associated with shifting attitudes.
35. Badawy, A., Addawood, A., Lerman, K. & Ferrara, E. Characterizing the 2016 Russian IRA influence campaign. *Soc. Netw. Anal. Min.* **9**, 31 (2019).
This paper shows that exposure to and amplification of Russian misinformation on social media in 2016 was concentrated among Republicans (who would have been predisposed to support Donald Trump regardless).
36. Hosseinmardi, H., Ghasemian, A., Clauset, A., Mobius, M., Rothschild, D. M. & Watts, D. J. Examining the consumption of radical content on YouTube. *Proc. Natl Acad. Sci. USA* **118**, e2101967118 (2021).
This paper shows that extreme content is consumed on YouTube by a small portion of the population who tend to consume similar content elsewhere online and that consumption is largely driven by demand, not algorithms.
37. 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).
This paper shows that people who consume extremist content on YouTube have highly resentful attitudes and typically find the content through subscriptions and external links, not algorithmic recommendations to non-subscribers.
38. Munger, K. & Phillips, J. Right-wing YouTube: a supply and demand perspective. *Int. J. Press Polit.* **27**, 186–219 (2022).
39. Lasser, J., Aroyehun, S. T., Simchon, A., Carrella, F., Garcia, D. & Lewandowsky, S. Social media sharing of low-quality news sources by political elites. *PNAS Nexus* **1**, pgac186 (2022).
40. Muddiman, A., Budak, C., Murray, C., Kim, Y. & Stroud, N. J. Indexing theory during an emerging health crisis: how U.S. TV news indexed elite perspectives and amplified COVID-19 misinformation. *Ann. Inte. Commun. Assoc.* **46**, 174–204 (2022).
This paper shows how mainstream media also spreads misinformation through amplification of misleading statements from elites.
41. Pereira, F. B. et al. Detecting misinformation: identifying false news spread by political leaders in the Global South. Preprint at *OSF*, https://doi.org/10.31235/osf.io/hu4qr (2022).
42. Horwitz, J. & Seetharaman, D. Facebook executives shut down efforts to make the site less divisive. *Wall Street Journal*, https://www.wsj.com/articles/facebook-knows-it-encourages-division-top-executives-nixed-solutions-11590507499 (26 May 2020).
43. Hosseinmardi, H., Ghasemian, A., Rivera-Lanas, M., Horta Ribeiro, M., West, R. & Watts, D. J. Causally estimating the effect of YouTube's recommender system using counterfactual bots. *Proc. Natl Acad. Sci. USA* **121**, e2313377121 (2024).
44. Nyhan, B. et al. Like-minded sources on facebook are prevalent but not polarizing. *Nature* **620**, 137–144 (2023).
45. Guess, A. M. et al. How do social media feed algorithms affect attitudes and behavior in an election campaign? *Science* **381**, 398–404 (2023).
This paper shows that algorithms supply less untrustworthy content than reverse chronological feeds.
46. Asimovic, N., Nagler, J., Bonneau, R. & Tucker, J. A. Testing the effects of Facebook usage in an ethnically polarized setting. *Proc. Natl Acad. Sci. USA* **118**, e2022819118 (2021).
47. Allen, J., Mobius, M., Rothschild, D. M. & Watts, D. J. Research note: Examining potential bias in large-scale censored data. *Harv. Kennedy Sch. Misinformation Rev.* **2**, https://doi.org/10.37016/mr-2020-74 (2021).
This paper shows that engagement metrics such as clicks and shares that are regularly used in popular and academic research do not take into account the fact that fake news is clicked and shared at a higher rate relative to exposure and viewing than non-fake news.
48. Scheuerman, M. K., Jiang, J. A., Fiesler, C. & Brubaker, J. R. A framework of severity for harmful content online. *Proc. ACM Hum. Comput. Interact.* **5**, 1–33 (2021).
49. Vosoughi, S., Roy, D. & Aral, S. The spread of true and false news online. *Science* **359**, 1146–1151 (2018).
50. Roy, D. Happy to see the extensive coverage of our science paper on spread of true and false news online, but over-interpretations of the scope of our study prompted me to diagram actual scope (caution, not to scale!). *Twitter*, https://twitter.com/dkroy/status/974251282071474177 (15 March 2018).
51. Greenemeier, L. You can't handle the truth—at least on Twitter. *Scientific American*, https://www.scientificamerican.com/article/you-cant-handle-the-truth-at-least-on-twitter/ (8 March 2018).
52. Frankel, S. Deceptively edited video of Biden proliferates on social media. *The New York Times*, https://www.nytimes.com/2020/11/02/technology/biden-video-edited.html (2 November 2020).

53. Jiameng P. et al. Deepfake videos in the wild: analysis and detection. In *Proc. Web Conference 2021* 981–992 (International World Wide Web Conference Committee, 2021).
54. *Widely Viewed Content Report: What People See on Facebook: Q1 2023 Report* (Facebook, 2023).
55. Mayer, J. How Russia helped swing the election for Trump. *The New Yorker*, https://www.newyorker.com/magazine/2018/10/01/how-russia-helped-to-swing-the-election-for-trump (24 September 2018).
56. Jamieson, K. H. *Cyberwar: How Russian Hackers and Trolls Helped Elect A President: What We Don't, Can't, and Do Know* (Oxford Univ. Press, 2020).
57. Solon, O. & Siddiqui, S. Russia-backed Facebook posts 'reached 126m Americans' during US election. *The Guardian*, https://www.theguardian.com/technology/2017/oct/30/facebook-russia-fake-accounts-126-million (30 October 2017).
58. Watts, D. J. & Rothschild, D. M. Don't blame the election on fake news. Blame it on the media. *Columbia J. Rev.* **5**, https://www.cjr.org/analysis/fake-news-media-election-trump.php (2017).
This paper explores how seemingly large exposure levels to problematic content actually represent a small proportion of total news exposure.
59. Jie, Y. Frequency or total number? A comparison of different presentation formats on risk perception during COVID-19. *Judgm. Decis. Mak.* **17**, 215–236 (2022).
60. Reyna, V. F. & Brainerd, C. J. Numeracy, ratio bias, and denominator neglect in judgments of risk and probability. *Learn. Individ. Differ.* **18**, 89–107 (2008).
This paper details research into how salient numbers can lead to confusion in judgements of risk and probability, such as denominator neglect in which people fixate on a large numerator and do not consider the appropriate denominator.
61. Jones, J. Americans: much misinformation, bias, inaccuracy in news. *Gallup*, https://news.gallup.com/opinion/gallup/235796/americans-misinformation-bias-inaccuracy-news.aspx (2018).
62. 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).
63. Guess, A. M., Nyhan, B. & Reifler, J. Exposure to untrustworthy websites in the 2016 US election. *Nat. Hum. Behav.* **4**, 472–480 (2020).
This paper shows untrustworthy news exposure was relatively rare in US citizens' web browsing in 2016.
64. Altay, S., Nielsen, R. K. & Fletcher, R. Quantifying the "infodemic": people turned to trustworthy news outlets during the 2020 coronavirus pandemic. *J. Quant. Descr. Digit. Media* **2**, 1–30 (2022).
65. 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).
This paper shows that exposure to fake news is a vanishingly small part of people's overall news diets when you take television into account.
66. 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).
This paper shows hows how a small portion of the population accounts for the vast majority of exposure to vaccine-sceptical content.
67. Chong, D. & Druckman, J. N. Framing public opinion in competitive democracies. *Am. Polit. Sci. Rev.* **101**, 637–655 (2007).
68. Arendt, F. Toward a dose-response account of media priming. *Commun. Res.* **42**, 1089–1115 (2015).
This paper shows that people may need repeated exposure to information for it to affect their attitudes.
69. Arceneaux, K., Johnson, M. & Murphy, C. Polarized political communication, oppositional media hostility, and selective exposure. *J. Polit.* **74**, 174–186 (2012).
70. Feldman, L. & Hart, P. Broadening exposure to climate change news? How framing and political orientation interact to influence selective exposure. *J. Commun.* **68**, 503–524 (2018).
71. Druckman, J. N. Political preference formation: competition, deliberation, and the (ir)relevance of framing effects. *Am. Polit. Sci. Rev.* **98**, 671–686 (2004).
72. Bakshy, E., Messing, S. & Adamic, L. A. Exposure to ideologically diverse news and opinion on facebook. *Science* **348**, 1130–1132 (2015).
73. Bozarth, L., Saraf, A. & Budak, C. Higher ground? How groundtruth labeling impacts our understanding of fake news about the 2016 U.S. presidential nominees. In *Proc. International AAAI Conference on Web and Social Media* Vol. 14, 48–59 (Association for the Advancement of Artificial Intelligence, 2020).
74. Gerber, A. S., Gimpel, J. G., Green, D. P. & Shaw, D. R. How large and long-lasting are the persuasive effects of televised campaign ads? Results from a randomized field experiment. *Am. Polit. Sci. Rev.* **105**, 135–150 (2011).
This paper shows that the effect of news decays rapidly; news needs repeated exposure for long-term impact.
75. Hill, S. J., Lo, J., Vavreck, L. & Zaller, J. How quickly we forget: the duration of persuasion effects from mass communication. *Polit. Commun.* **30**, 521–547 (2013).
This paper shows that the effect of persuasive advertising decays rapidly, necessitating repeated exposure for lasting effect.
76. Larsen, M. V. & Olsen, A. L. Reducing bias in citizens' perception of crime rates: evidence from a field experiment on burglary prevalence. *J. Polit.* **82**, 747–752 (2020).
77. Roose, K. What if Facebook is the real 'silent majority'? *The New York Times*, https://www.nytimes.com/2020/08/28/us/elections/what-if-facebook-is-the-real-silent-majority.html (27 August 2020).
78. Breland, A. A new report shows how Trump keeps buying Facebook ads. *Mother Jones*, https://www.motherjones.com/politics/2021/07/real-facebook-oversight-board/ (28 July 2021).
79. Marchal, N., Kollanyi, B., Neudert, L.-M. & Howard, P. N. *Junk News during the EU Parliamentary Elections: Lessons from A Seven-language Study of Twitter and Facebook* (Univ. Oxford, 2019).
80. Ellison, N. B., Trieu, P., Schoenebeck, S., Brewer, R. & Israni, A. Why we don't click: interrogating the relationship between viewing and clicking in social media contexts by exploring the "non-click". *J. Comput. Mediat. Commun.* **25**, 402–426 (2020).
81. Pennycook, G., Epstein, Z., Mosleh, M., Arechar, A. A., Eckles,and, D. & Rand, D. G. Shifting attention to accuracy can reduce misinformation online. *Nature* **592**, 590–595 (2021).
82. Ghezae, I. et al. Partisans neither expect nor receive reputational rewards for sharing falsehoods over truth online. *Open Science Framework* https://osf.io/5jwgd/ (2023).
83. 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).
84. Godel, W. et al. Moderating with the mob: evaluating the efficacy of real-time crowdsourced fact-checking. *J. Online Trust Saf.* **1**, https://doi.org/10.54501/jots.v1i1.15 (2021).
85. Rogers, K. Facebook's algorithm is broken. We collected some suggestion on how to fix it. *FiveThirtyEight*, https://fivethirtyeight.com/features/facebooks-algorithm-is-broken-we-collected-some-spicy-suggestions-on-how-to-fix-it/ (16 November 2021).
86. Roose, K. The making of a YouTube radical. *The New York Times*, https://www.nytimes.com/interactive/2019/06/08/technology/youtube-radical.html (8 June 2019).
87. Eslami, M. et al. First I "like" it, then I hide it: folk theories of social feeds. In *Proc. 2016 CHI Conference on Human Factors in Computing Systems* 2371–2382 (Association for Computing Machinery, 2016).
88. Silva, D. E., Chen, C. & Zhu, Y. Facets of algorithmic literacy: information, experience, and individual factors predict attitudes toward algorithmic systems. *New Media Soc.* https://doi.org/10.1177/14614448221098042 (2022).
89. Eckles, D. *Algorithmic Transparency and Assessing Effects of Algorithmic Ranking. Testimony before the Senate Subcommittee on Communications, Media, and Broadband*, https://www.commerce.senate.gov/services/files/62102355-DC26-4909-BF90-8FB068145F18 (U.S. Senate Committee on Commerce, Science, and Transportation, 2021).
90. Kantrowitz, A. Facebook removed the news feed algorithm in an experiment. Then it gave up. *OneZero*, https://onezero.medium.com/facebook-removed-the-news-feed-algorithm-in-an-experiment-then-it-gave-up-25c8cb0a35a3 (25 October 2021).
91. Ribeiro, M. H., Hosseinmardi, H., West, R. & Watts, D. J. Deplatforming did not decrease parler users' activity on fringe social media. *PNAS Nexus* **2**, pgad035 (2023).
This paper shows that shutting down Parler just displaced user activity to other fringe social media websites.
92. Alfano, M., Fard, A. E., Carter, J. A., Clutton, P. & Klein, C. Technologically scaffolded atypical cognition: the case of YouTube's recommender system. *Synthese* **199**, 835–858 (2021).
93. Huszár, F. et al. Algorithmic amplification of politics on Twitter. *Proc. Natl Acad. Sci. USA* **119**, e2025334119 (2022).
94. Levy, R. Social media, news consumption, and polarization: evidence from a field experiment. *Am. Econ. Rev.* **111**, 831–870 (2021).
95. Cho, J., Ahmed, S., Hilbert, M., Liu, B. & Luu, J. Do search algorithms endanger democracy? An experimental investigation of algorithm effects on political polarization. *J. Broadcast. Electron. Media* **64**, 150–172 (2020).
96. Lewandowsky, S., Robertson, R. E. & DiResta, R. Challenges in understanding human-algorithm entanglement during online information consumption. *Perspect. Psychol. Sci.* https://doi.org/10.1177/17456916231180809 (2023).
97. Narayanan, A. *Understanding Social Media Recommendation Algorithms* (Knight First Amendment Institute at Columbia University, 2023).
98. Finkel, E. J. et al. Political sectarianism in America. *Science* **370**, 533–536 (2020).
99. Auxier, B. & Anderson, M. *Social Media Use in 2021* (Pew Research Center, 2021).
100. Frimer, J. A. et al. Incivility is rising among American politicians on Twitter. *Soc. Psychol. Personal. Sci.* **14**, 259–269 (2023).
101. Broderick, R. & Darmanin, J. The "yellow vest" riots in France are what happens when Facebook gets involved with local news. *Buzzfeed News*, https://www.buzzfeednews.com/article/ryanhatesthis/france-paris-yellow-jackets-facebook (2018).
102. Salzberg, S. De-platform the disinformation dozen. *Forbes*, https://www.forbes.com/sites/stevensalzberg/2021/07/19/de-platform-the-disinformation-dozen/ (2021).
103. Karell, D., Linke, A., Holland, E. & Hendrickson, E. "Born for a storm": hard-right social media and civil unrest. *Am. Soc. Rev.* **88**, 322–349 (2023).
104. Smith, N. & Graham, T. Mapping the anti-vaccination movement on Facebook. *Inf. Commun. Soc.* **22**, 1310–1327 (2019).
105. Brady, W. J., McLoughlin, K., Doan, T. N. & Crockett, M. J. How social learning amplifies moral outrage expression in online social networks. *Sci. Adv.* **7**, eabe5641 (2021).
106. Suhay, E., Bello-Pardo, E. & Maurer, B. The polarizing effects of online partisan criticism: evidence from two experiments. *Int. J. Press Polit.* **23**, 95–115 (2018).
107. Arugute, N., Calvo, E. & Ventura, T. Network activated frames: content sharing and perceived polarization in social media. *J. Commun.* **73**, 14–24 (2023).
108. Nordbrandt, M. Affective polarization in the digital age: testing the direction of the relationship between social media and users' feelings for out-group parties. *New Media Soc.* **25**, 3392–3411 (2023).
This paper shows that affective polarization predicts media use, not the other way around.
109. AFP. Street protests, a French tradition par excellence. *The Local* https://www.thelocal.fr/20181205/revolutionary-tradition-the-story-behind-frances-street-protests (2018).
110. Spier, R. E. Perception of risk of vaccine adverse events: a historical perspective. *Vaccine* **20**, S78–S84 (2001).
This article documents the history of untrustworthy information about vaccines, which long predates social media.
111. Bryant, L. V. The YouTube algorithm and the alt-right filter bubble. *Open Inf. Sci.* **4**, 85–90 (2020).
112. Sismeiro, C. & Mahmood, A. Competitive vs. complementary effects in online social networks and news consumption: a natural experiment. *Manage. Sci.* **64**, 5014–5037 (2018).
113. Fergusson, L. & Molina, C. *Facebook Causes Protests Documento CEDE No. 41*, https://doi.org/10.2139/ssrn.3553514 (2019).
114. Lu, Y., Wu, J., Tan, Y. & Chen, J. Microblogging replies and opinion polarization: a natural experiment. *MIS Q.* **46**, 1901–1936 (2022).
115. Porter, E. & Wood, T. J. The global effectiveness of fact-checking: evidence from simultaneous experiments in Argentina, Nigeria, South Africa, and the United Kingdom. *Proc. Natl Acad. Sci. USA* **118**, e2104235118 (2021).

116. Arechar, A. A. et al. Understanding and combatting misinformation across 16 countries on six continents. *Nat. Hum. Behav.* **7**, 1502–1513 (2023).
117. Blair, R. A. et al. *Interventions to Counter Misinformation: Lessons from the Global North and Applications to the Global South* (USAID Development Experience Clearinghouse, 2023).
118. Haque, M. M. et al. Combating misinformation in Bangladesh: roles and responsibilities as perceived by journalists, fact-checkers, and users. *Proc. ACM Hum. Comput. Interact.* **4**, 1–32 (2020).
119. Humprecht, E., Esser, F. & Van Aelst, P. Resilience to online disinformation: a framework for cross-national comparative research. *Int. J. Press Polit.* **25**, 493–516 (2020).
120. Gillum, J. & Elliott, J. Sheryl Sandberg and top Facebook execs silenced an enemy of Turkey to prevent a hit to the company's business. *ProPublica*,https://www.propublica.org/article/sheryl-sandberg-and-top-facebook-execs-silenced-an-enemy-of-turkey-to-prevent-a-hit-to-their-business (24 February 2021).
121. Nord M. et al. *Democracy Report 2024: Democracy Winning and Losing at the Ballot* V-Dem Report (Univ. Gothenburg V-Dem Institute, 2024).
122. Alba, D. How Duterte used Facebook to fuel the Philippine drug war. *Buzzfeed*, https://www.buzzfeednews.com/article/daveyalba/facebook-philippines-dutertes-drug-war (4 September 2018).
123. Zakrzewski, C., De Vynck, G., Masih, N. a& Mahtani, S. How Facebook neglected the rest of the world, fueling hate speech and violence in India. *Washington Post*, https://www.washingtonpost.com/technology/2021/10/24/india-facebook-misinformation-hate-speech/ (24 October 2021).
124. Simonite, T. Facebook is everywhere; its moderation is nowhere close. *Wired*, https://www.wired.com/story/facebooks-global-reach-exceeds-linguistic-grasp/ (21 October 2021).
125. Cruz, J. C. B. & Cheng, C. Establishing baselines for text classification in low-resource languages. Preprint at https://arxiv.org/abs/2005.02068 (2020).
This paper shows one of the challenges that makes content moderation costlier in less resourced countries.
126. Müller, K. & Schwarz, C. Fanning the flames of hate: social media and hate crime. *J. Eur. Econ. Assoc.* **19**, 2131–2167 (2021).
127. Bursztyn, L., Egorov, G., Enikolopov, R. & Petrova, M. *Social Media and Xenophobia: Evidence from Russia* (National Bureau of Economic Research, 2019).
128. Lewandowsky, S., Jetter, M. & Ecker, U. K. H. Using the President's tweets to understand political diversion in the age of social media. *Nat. Commun.* **11**, 5764 (2020).
129. Bursztyn, L., Rao, A., Roth, C. P. & Yanagizawa-Drott, D. H. *Misinformation During a Pandemic* (National Bureau of Economic Research, 2020).
130. Motta, M. & Stecula, D. Quantifying the effect of Wakefield et al. (1998) on skepticism about MMR vaccine safety in the US. *PLoS ONE* **16**, e0256395 (2021).
131. Sanderson, Z., Brown, M. A., Bonneau, R., Nagler, J. & Tucker, J. A. Twitter flagged Donald Trump's tweets with election misinformation: they continued to spread both on and off the platform. *Harv. Kennedy Sch. Misinformation Rev.* **2**, https://doi.org/10.37016/mr-2020-77 (2021).
132. Anhalt-Depies, C., Stenglein, J. L., Zuckerberg, B., Townsend, P. A. & Rissman, A. R. Tradeoffs and tools for data quality, privacy, transparency, and trust in citizen science. *Biol. Conserv.* **238**, 108195 (2019).
133. Gerber, N., Gerber, P. & Volkamer, M. Explaining the privacy paradox: a systematic review of literature investigating privacy attitude and behavior. *Comput. Secur.* **77**, 226–261 (2018).
This paper explores the trade-offs between privacy and research.
134. Isaak, J. & Hanna, M. J. User data privacy: Facebook, Cambridge Analytica, and privacy protection. *Computer* **51**, 56–59 (2018).
135. Vogus, C. *Independent Researcher Access to Social Media Data: Comparing Legislative Proposals* (Center for Democracy and Technology, 2022).
136. Xie, Y. "Undemocracy": inequalities in science. *Science* **344**, 809–810 (2014).
137. Nielsen, M. W. & Andersen, J. P. Global citation inequality is on the rise. *Proc. Natl Acad. Sci. USA* **118**, e2012208118 (2021).
138. King, D. A. The scientific impact of nations. *Nature* **430**, 311–316 (2004).
139. Zaugg, I. A., Hossain, A. & Molloy, B. Digitally-disadvantaged languages. *Internet Policy Rev.* **11**, 1–11 (2022).
140. Zaugg, I. A. in *Digital Inequalities in the Global South* (eds Ragnedda, M. & Gladkova, A.) 247–267 (Springer, 2020).
141. Sablosky, J. Dangerous organizations: Facebook's content moderation decisions and ethnic visibility in Myanmar. *Media Cult. Soc.* **43**, 1017–1042 (2021).
This paper highlights the challenges of content moderation in the Global South.
142. Warofka, A. An independent assessment of the human rights impact of Facebook in Myanmar. *Facebook Newsroom*, https://about.fb.com/news/2018/11/myanmar-hria/ (2018).
143. Fick, M. & Dave, P. Facebook's flood of languages leave it struggling to monitor content. *Reuters*, https://www.reuters.com/article/idUSKCN1RZ0DL/ (23 April 2019).
144. Newman, N. *Executive Summary and Key Findings of the 2020 Report* (Reuters Institute for the Study of Journalism, 2020).
145. Hilbert, M. The bad news is that the digital access divide is here to stay: domestically installed bandwidths among 172 countries for 1986–2014. *Telecommun. Policy* **40**, 567–581 (2016).
146. Traynor, I. Internet governance too US-centric, says European commission. *The Guardian*, https://www.theguardian.com/technology/2014/feb/12/internet-governance-us-european-commission (12 February 2014).
147. Pennycook, G., Cannon, T. D. & Rand, D. G. Prior exposure increases perceived accuracy of fake news. *J. Exp. Psychol. Gen.* **147**, 1865–1880 (2018).
148. Guess, A. M. et al. "Fake news" may have limited effects beyond increasing beliefs in false claims. *Kennedy Sch. Misinformation Rev.* **1**, https://doi.org/10.37016/mr-2020-004 (2020).
149. Loomba, S., de Figueiredo, A., Piatek, S. J., de Graaf, K. & Larson, H. J. Measuring the impact of COVID-19 vaccine misinformation on vaccination intent in the UK and USA. *Nat. Hum. Behav.* **5**, 337–348 (2021).
150. Lorenz-Spreen, P., Oswald, L., Lewandowsky, S. & Hertwig, R. Digital media and democracy: a systematic review of causal and correlational evidence worldwide. *Nat. Hum. Behav.* **7**, 74–101 (2023).
This paper provides a review of evidence on social media effects.
151. Donato, K. M., Singh, L., Arab, A., Jacobs, E. & Post, D. Misinformation about COVID-19 and Venezuelan migration: trends in Twitter conversation during a pandemic. *Harvard Data Sci. Rev.* **4**, https://doi.org/10.1162/99608f92.a4d9a7c7 (2022).
152. Wieczner, J. Big lies vs. big lawsuits: why Dominion Voting is suing Fox News and a host of Trump allies. *Fortune*, https://fortune.com/longform/dominion-voting-lawsuits-fox-news-trump-allies-2020-election-libel-conspiracy-theories/ (2 April 2021).
153. Calma, J. Twitter just closed the book on academic research. *The Verge* https://www.theverge.com/2023/5/31/23739084/twitter-elon-musk-api-policy-chilling-academic-research (2023).
154. Edelson, L., Graef, I. & Lancieri, F. *Access to Data and Algorithms: for an Effective DMA and DSA Implementation* (Centre on Regulation in Europe, 2023).

Author contributions C.B., B.N., D.M.R., E.T. and D.J.W. wrote and revised the paper. D.M.R. collected the data and prepared Fig. 1.

Competing interests The authors declare no competing interests, but provide the following information in the interests of transparency and full disclosure. C.B. and D.J.W. previously worked for Microsoft Research and D.M.R. currently works for Microsoft Research. B.N. has received grant funding from Meta. B.N. and E.T. are participants in the US 2020 Facebook and Instagram Election Study as independent academic researchers. D.J.W. has received funding from Google Research. D.M.R. and D.J.W. both previously worked at Yahoo!.

Additional information
Correspondence and requests for materials should be addressed to David M. Rothschild.
Peer review information *Nature* thanks Stephan Lewandowsky, David Rand, Emma Spiro and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.
Reprints and permissions information is available at http://www.nature.com/reprints.
Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.