# Social Media and Political Dysfunction: A Collaborative Review

This Google doc is an open-source working document that contains the citations and abstracts of published articles that shed light on a question that is currently being debated within many democratic nations: **Is social media a major contributor to the rise of political dysfunction seen in the USA and some other democracies since the early 2010s?** This is too broad a question to be answered, so we break it down into seven more specific questions for which there is a substantial research literature.

This document is curated by Jonathan Haidt (NYU-Stern) and Chris Bail (Duke), with research assistance from Zach Rausch. If you are a researcher or industry insider and have studies or comments to add, please click the “Request Access” button above (while signed in to a Google account), tell us who you are, and Zach will give you commenter status. We especially welcome critical comments: What studies have we missed, or misinterpreted?

You can cite this document as: Haidt, J., & Bail, C. (ongoing). *Social media and political dysfunction: A collaborative review*. Unpublished manuscript, New York University.

First posted: November 2, 2021. Last updated: August 28, 2023.
You can always find this document at: https://tinyurl.com/PoliticalDysfunctionReview
To see Haidt’s other Collaborative Review docs:

- Adolescent mood disorders since 2010: A collaborative review [with Jean Twenge]
- Social Media and Mental Health: A Collaborative Review [with Jean Twenge]

==========

## Clickable Table of Contents

* * * * * * * * * * * * * * *

# INTRODUCTION

In the 1990s, it seemed that liberal democracy had triumphed over all other forms of government as the best way to run a modern, prosperous, diverse nation. When the Internet became widespread, in the late 1990s and early 2000s, it seemed to be a gift to democracy; what dictator could stand up to the people, empowered? How could any nation keep the internet out? Techno-democratic optimism arguably reached a high point in 2011, a year that began with the Arab Spring, followed by mass protests in Israel and Spain, and culminating with the Occupy movement that began in New York City and then spread globally.

The 2010s did not turn out as many of us expected. Democracy is now on the back foot, with more countries becoming less democratic, and the decline begins or accelerates in the 2010s (see Appendix D). The United States in particular has veered into deep political dysfunction, intense affective polarization, and televised political violence. Alternatives to liberal democracy are more numerous— and in some ways more stable—including illiberal democracies such as Hungary, and the one-party authoritarian system developing in China.

What happened? Why is the outlook for democracy so much darker in 2022 than it was in 2011?

Among the most widely discussed causes of recent political dysfunction is social media, which transformed social connections, mass movements, news consumption, and avenues for electoral interference, manipulation, and misinformation. The two unexpected successes of the Brexit referendum and the Trump campaign, both in 2016, turned attention to Facebook in particular, but also to Twitter and YouTube. A number of popular books in recent years have made the case that Facebook, in particular, was a danger to democracy. Reporting by the Wall Street Journal (The Facebook Files), and by the New York Times and Washington Post also pointed to democracy-disrupting effects of Facebook and other platforms.

Is it true? Are Facebook and other social media platforms damaging democracies? Documents brought out by whistleblower Frances Haugen, along with her Congressional testimony, suggest the answer may be "yes." Facebook denies the charge, and points to several studies published by social scientists in its defense. *A systematic review of the literature is therefore needed to communicate the findings of this rapidly evolving literature to the public*. Unfortunately, there is now so much research published (or circulating as working papers) that it is impossible for anyone who does not study this question full time to know what is out there, and what it all adds up to. Hence this document.

We (Haidt & Bail) have organized the document into the major questions that extant research has addressed. For each question, we list all the published studies we can find (along with working papers from established researchers), grouped into those that support the proposition that social media is harming democracies, and those that do not support the proposition. After we created the initial framework for this document we invited other researchers to add other studies we had missed, and to critique the relevance or interpretation provided in the text below.

We thank these researchers for offering their ideas and constructive criticisms: Kevin Munger (Penn State U), David Rand (MIT), Andy Guess (Princeton), Will Blakey (UNC), Richard Fletcher (University of Oxford), Sacha Altay (University of Oxford), Olivia Fischer (University of Zurich), Tim Samples (University of Georgia) [more to come]... And we thank Gideon Lewis-Kraus for exploring this collaborative review, and criticisms of it, in an essay in The New Yorker.

# NOTES AND CAVEATS

**1. What do we mean by "Political dysfunction"?**

A comprehensive overview of the many effects of social media on politics is beyond the scope of this review. We acknowledge that there is evidence that social media has created positive outcomes on issues such as voter registration, mobilization within authoritarian regimes, and others, but this review focuses on evidence of harm (see Lorenz-Spreen et al. 2022 [study 9.1.13] for evidence that the benefits of social media are mostly found in less developed democracies, while the harms are more frequently found in advanced democracies). We review the literature on social media and political dysfunction. Our definition of political dysfunction includes political polarization—including not only increasing disagreement about substantive issues but also the rise in negative feelings and attitudes between partisans (often referred to as "affective polarization"). Our definition of dysfunction also includes a broader set of behavioral and attitudinal outcomes including a) support for the use of violence to achieve political ends; b) alienation from the democratic process (through voter suppression or general apathy about government); c) declining trust in government, politicians, and key institutions; d) decreased willingness to listen to or work with those from other groups/parties, and e) the spread of misinformation and misleading claims about politics within the broader information environment.

**2. What do we mean by "Social Media"?**

We do not examine the impact of "The Internet" writ large on politics— a topic which would also require a much broader effort. Instead, we focus upon the impact of *social media* alone. We define social media as *communications technology that allows people to create an online social network where they agree to receive updates in text or audio-visual format from other users that are delivered to them within a "news feed" or ordered list of information. The list may or may not be determined by an algorithm*. We thus focus primarily upon the impact of large platforms such as Facebook, Twitter, Instagram, TikTok, Reddit, and YouTube. Our review does not include chat platforms such as WhatsApp, SnapChat, Telegram, or Discord that do not involve an ordered timeline and that primarily serve peer-to-peer conversations.

**3. What time period and what countries are we covering?**

We include only articles published in or after 2014. Haidt wrote an essay with Tobias Rose-Stockwell in The Atlantic in 2019 making the case that social media -- especially Facebook and Twitter-- changed fundamentally in the years 2009 through 2012, after Facebook added the Like button and Twitter added the Retweet button. For this reason, **research on social media and democracy drawing on data before 2013 is not as relevant**. We can't cover everything, so we limit ourselves to studies that were published in 2014 or later.

Although we focus on social media and political dysfunction, we acknowledge that the latter has many complex causes that predate the emergence of platforms such as Facebook, Twitter, and YouTube. Among other things, these include an array of historical forces from the realignment of strategies of political parties, the rise of cable news in the 1990s, increases in negative campaigning, the rise of social and economic inequality, enduring racial prejudice, and many other factors.

Because the bulk of empirical research on social media and political dysfunction has been conducted in Western democracies, our review focuses upon this area. Wherever possible, however, we included emerging evidence from other regions of the world as well and we invite readers to suggest relevant work from these areas that we may have missed as well.

**4. This is not a formal meta-analysis**

A search on Google Scholar for "social media" and "democracy" from 2014 to 2022 produces 214,000 hits. We did not begin with such a search. Rather, we tried to identify the articles that are being cited and discussed from the years 2014 through 2020, while trying to do a more comprehensive job of capturing new research published in and after 2021. We are particularly interested in experimental and quasi-experimental research. We invite researchers to add links to any studies they believe should be included, in the relevant sections, after the text that says, in green, What have we missed? We also do not attempt to weight studies by their sample size or quality, as would be done in a formal meta-analysis. Rather, our goal here is to help researchers and members of the public get an overview of the kinds of research that are out there, structured so that readers can quickly see evidence on both sides of each question. We caution readers not to simply add up the number of studies on each side and declare one side the winner.

**5. A note about this Google doc**

We are not unbiased. Haidt has written two Atlantic articles arguing that social media is damaging democratic and epistemic institutions. Bail wrote an entire book explaining how to Make our platforms less polarizing. We therefore began the project with prior beliefs and a bias toward confirming the "yes" answer to each of the seven questions below. As scholars, however, we want to be right in the long run, not the short run. We are great fans of John Stuart Mill, who wrote that "the only way in which a human being can make some approach to knowing the whole of a subject, is by hearing what can be said about it by persons of every variety of opinion, and studying all modes in which it can be looked at by every character of mind." And: "The steady habit of correcting and completing his own opinion by collating it with those of others… is the only stable foundation for a just reliance on it." We therefore created this Google doc to invite researchers who have different opinions and confirmation biases to collate their views with ours. Social media platforms and democratic difficulties are changing so fast that the normal academic cycle of data collection, publication, and meta-analysis, often spanning five to ten years, is just too slow to keep up. A living document where researchers can add their own in-press publications and their own critiques, may be a helpful supplement to the normal academic process.

* * * * * * * * * * * * * * *

# QUESTION 1: DOES SOCIAL MEDIA MAKE PEOPLE MORE ANGRY OR AFFECTIVELY POLARIZED?

This section of our review covers multiple outcomes studied by social scientists—and political scientists in particular—to capture anger and "affective polarization," which refers to animosity between members of different political parties independent of the content of their beliefs.

## 1.1 STUDIES INDICATING YES

**1.1.1** Banks, Calvo, Karol, & Telhami (2020). #PolarizedFeeds: Three experiments on polarization, framing, and social media. *The International Journal of Press/Politics.*

ABSTRACT: Does exposure to social media polarize users or simply sort out like-minded voters based on their preexisting beliefs? In this paper, we conduct three survey experiments to assess the direct and unconditioned effect of exposure to tweets on perceived ideological polarization of candidates and parties. We show that **subjects treated with negative tweets see greater ideological distance between presidential nominees and between their parties. We also demonstrate that polarization increases with processing time. We demonstrate a social media effect on perceived polarization beyond that due to the self-selection of like-minded users into different media communities.** We explain our results as the result of social media frames that increase *contrast* effects between voters and candidates.

**1.1.2** Cho, Ahmed, Hilbert, Liu, & Luu (2020). Do search algorithms endanger democracy? An experimental investigation of algorithm effects on political polarization. *Journal of Broadcasting & Electronic Media.*

ABSTRACT: This study examines algorithm effects on user opinion, utilizing a real-world recommender algorithm of a highly popular video-sharing platform, YouTube. We experimentally manipulate user search/watch history by our custom programming. A controlled laboratory experiment is then conducted to examine whether exposure to algorithmically recommended content reinforces and polarizes political opinions. Results suggest that **political self-reinforcement, as indicated by the political emotion-ideology alignment, and affective polarization are heightened by political videos – selected by the YouTube recommender algorithm – based on participants' own search preferences.** Suggestions for how to reduce algorithm-induced political polarization and implications of algorithmic personalization for democracy are discussed.

**1.1.3** Barnidge, M. (2017). Exposure to political disagreement in social media versus face-to-face and anonymous online settings. *Political Communication*, *34*(2), 302–321.

ABSTRACT: This article investigates political disagreement on social media in comparison to face-to-face and anonymous online settings. Because of the structure of

social relationships and the social norms that influence expression, it is hypothesized that people perceive more political disagreement in social media settings versus face-to-face and anonymous online settings. **Analyses of an online survey of adults in the United States show that (a) social media users perceive more political disagreement than non-users, (b) they perceive more of it on social media than in other communication settings, and (c) news use on social media is positively related to perceived disagreement on social media.** Results are discussed in light of their implications for current debates about the contemporary public sphere and directions for future research.

**1.1.4** Rathje, Van Bavel, & van der Linden (2021). Out-group animosity drives engagement on social media. *Proceedings of the National Academy of Sciences*.

ABSTRACT: There has been growing concern about the role social media plays in political polarization. We investigated whether out-group animosity was particularly successful at generating engagement on two of the largest social media platforms: Facebook and Twitter. Analyzing posts from news media accounts and US congressional members (n = 2,730,215), we found that posts about the political out-group were shared or retweeted about twice as often as posts about the in-group. **Each individual term referring to the political out-group increased the odds of a social media post being shared by 67%. Out-group language consistently emerged as the strongest predictor of shares and retweets**: the average effect size of out-group language was about 4.8 times as strong as that of negative affect language and about 6.7 times as strong as that of moral-emotional language—both established predictors of social media engagement. Language about the out-group was a very strong predictor of “angry” reactions (the most popular reactions across all datasets), and language about the in-group was a strong predictor of “love” reactions, reflecting in-group favoritism and out-group derogation. This out-group effect was not moderated by political orientation or social media platform, but stronger effects were found among political leaders than among news media accounts. In sum, **out-group language is the strongest predictor of social media engagement across all relevant predictors measured, suggesting that social media may be creating perverse incentives for content expressing out-group animosity.**

**1.1.5** Cho, Ahmed, Keum, Choi, & Lee (2018). Influencing myself: Self-reinforcement through online political expression. *Communication Research*.

ABSTRACT: Over the past decade, various online communication platforms have empowered citizens to express themselves politically. Although the political impact of online citizen expression has drawn considerable attention, research has largely focused on whether and how citizen-generated messages influence the public as an information alternative to traditional news outlets. The present study aims to provide a new perspective on understanding citizen expression by examining its political implications for the expressers themselves rather than those exposed to the expressed ideas. Data from a national survey and an online discussion forum study suggest that expressing oneself about politics provides self-reinforcing feedback. **Political expressions on social media and the online forum were found to (a) reinforce the expressers' partisan thought process and (b) harden their pre-existing political preferences.** Implications for the role the Internet plays in democracy will be discussed.

**1.1.6** Suhay, Bello-Pardo, & Maurer (2018). The polarizing effects of online partisan criticism: Evidence from two experiments. *The International Journal of Press/Politics*.

ABSTRACT: Affective and social political polarization—a dislike of political opponents and a desire to avoid their company—are increasingly salient and pervasive features of politics in many Western democracies, particularly the United States. One contributor to these related phenomena may be increasing exposure to online political disagreements in which ordinary citizens criticize, and sometimes explicitly demean, opponents. This article presents two experimental studies that assessed whether U.S. partisans' attitudes became more prejudiced in favor of the in-party after exposure to online partisan criticism. In the first study, **we draw on an online convenience sample to establish that partisan criticism that derogates political opponents increases affective polarization**. In the second, we replicate these findings with a quasi-representative sample and extend the pattern of findings to social polarization. **We conclude that online partisan criticism likely has contributed to rising affective and social polarization in recent years between Democrats and Republicans in the United States, and perhaps between partisan and ideological group members in other developed democracies as well.** We close by discussing the troubling implications of these findings in light of continuing attempts by autocratic regimes and other actors to influence democratic elections via false identities on social media.

**1.1.7** Goyanes, Borah, & Gil de Zúñiga (2021). Social media filtering and democracy: Effects of social media news use and uncivil political discussions on social media unfriending. *Computers in Human Behavior*.

ABSTRACT: In todays' progressively polarized society, social media users are increasingly exposed to blatant uncivil comments, dissonant views, and controversial news contents, both from their peers and the media organizations they follow. Recent scholarship on selective avoidance suggests that citizens when exposed to contentious stimuli tend to either neglect, avoid, or by-pass such content, a practice scholarly known as users' filtration tactics or *unfriending*. Drawing upon a nationally representative panel survey from the United States (W1 = 1338/W2 = 511) fielded in 2019/2020, this study seeks to a) examine whether social media news use is associated to exposure to uncivil political discussions, and 2) explore the ways in which both constructs causally affect users' unfriending behavior. Finally, the study investigates the contingent moderating role of uncivil political discussion in energizing the relationship between social media use for news and unfriending. **Our findings first find support for the idea that social media news use directly activates citizens' uncivil discussions and unfriending, while uncivil political discussion directly triggers unfriending behavior and significantly contributes to intensify the effect of social media news use over citizens' unfriending levels.** These findings add to current conversations about the potential motivations and deleterious effects of social media filtering in contemporary democracies.

**1.1.8** Brady, McLoughlin, Doan, & Crockett (2021). How social learning amplifies moral outrage expression in online social networks. *Science Advances*.

ABSTRACT: Moral outrage shapes fundamental aspects of social life and is now widespread in online social networks. Here, we show how social learning processes amplify online moral outrage expressions over time. In two preregistered observational studies on Twitter (7331 users and 12.7 million total tweets) and two preregistered behavioral experiments (*N* = 240), we find that **positive social feedback for outrage expressions increases the likelihood of future outrage expressions, consistent with principles of reinforcement learning.** In addition, **users conform their outrage expressions to the expressive norms of their social networks, suggesting norm learning also guides online outrage expressions.** Norm learning overshadows reinforcement learning when normative information is readily observable: in ideologically extreme networks, where outrage expression is more common, users are less sensitive to social feedback when deciding whether to express outrage. Our findings highlight

how platform design interacts with human learning mechanisms to affect moral discourse in digital public spaces.

**1.1.9** Soral, Liu, & Bilewicz (2020). Media of contempt: Social media consumption predicts normative acceptance of anti-muslim hate speech and Islamoprejudice. *International Journal of Conflict and Violence*.

ABSTRACT: The new era of information technology brings new opportunities but also poses new threats. In our paper, we examine whether a shift from traditional print and broadcasting to new online media results in the increased normalization of hate speech towards minorities, and whether this change can subsequently increase prejudice towards minorities. Our research uses data from a representative two-wave longitudinal survey of Polish adults. In wave 1 (N = 1060), data on respondents' primary sources of information about the world (TV, newspapers, radio, online, social media, blogs) was collected. Wave 2 (N = 628), conducted six months later, included measures of perceived normativity of anti-Muslim hate speech and Islamophobia. **We found that respondents who were frequent social media users expressed higher levels of Islamoprejudice and perceived higher normativity of anti-Muslim hate speech than the respondents who got their news from traditional mass media.** We also found that an increase in perceived normativity of anti-Muslim hate speech can act as one of the mechanisms through which use of social media is linked to higher Islamoprejudice.

**1.1.10** Thiel & McCain (2022). Gabufacturing Dissent: An in-depth analysis of Gab. Stanford Internet Observatory.

ABSTRACT: Gab is a small but growing social media ecosystem catering primarily to far-right communities who believe they are unwelcome—rightly or not—on more mainstream social media platforms. Unlike the more mainstream platforms it hopes to replace, Gab makes very few efforts to moderate the content on its platform. As more mainstream platforms crack down on far-right extremism, that content has been welcomed on Gab. In this report, we provide an in-depth qualitative and qualitative analysis of Gab users and content. We find that after years of slow growth and financial difficulties, Gab was invigorated by new users and money following the January 6th insurrection. We also find that content on Gab can be just as toxic as that on sites previously deplatformed by companies such as Cloudflare and Epik; overtly Nazi content gets significant engagement. More analysis is needed to understand the impact

of deplatforming, and whether it may lead to increased funding for extreme platforms and further radicalization.

[NOTE from Haidt: It is an open question whether the giant open platforms like Facebook and Twitter create echo chambers within the user base. But smaller platforms such as Gab, created specifically to welcome users of a particular ideology (usually right wing) are pretty close to the platonic form of an echo chamber]

**1.1.11** Frimer, Aujla, Feinberg, Skitka, Aquino, Eichstaedt, & Willer (2022). Incivility is rising among American politicians on Twitter. *Social Psychological and Personality Science*. (h/t Robb Willer)

ABSTRACT: We provide the first systematic investigation of trends in the incivility of American politicians on Twitter, a dominant platform for political communication in the United States. Applying a validated artificial intelligence classifier to all 1.3 million tweets made by members of Congress since 2009, **we observe a 23% increase in incivility over a decade on Twitter.** Further analyses suggest that the **rise was partly driven by reinforcement learning in which politicians engaged in greater incivility following positive feedback.** Uncivil tweets tended to receive more approval and attention, publicly indexed by large quantities of "likes" and "retweets" on the platform. **Mediational and longitudinal analyses show that the greater this feedback for uncivil tweets, the more uncivil tweets were thereafter.** We conclude by discussing how the structure of social media platforms might facilitate this incivility-reinforcing dynamic between politicians and their followers.

**1.1.12** Bavel, Rathje, Harris, Robertson, & Sternisko (2021). How social media shapes polarization.

QUOTE: "Social media shapes polarization through the following social, cognitive, and technological processes: partisan selection, message content, and platform design and algorithms."

**1.1.13** Lajevardi, Oskooii, & Walker (2022). Hate, amplified? Social media news consumption and support for anti-Muslim policies. *Journal of Public Policy*.

ABSTRACT: Research finds that social media platforms' peer-to-peer structures shape the public discourse and increase citizens' likelihood of exposure to unregulated, false, and prejudicial content. Here, we test whether self-reported reliance on social media as a primary news source is linked to racialised policy support, taking the case of United States Muslims, a publicly visible but understudied group about whom significant false and prejudicial content is abundant on these platforms. Drawing on three original surveys and the Nationscape dataset, **we find a strong and consistent association between reliance on social media and support for a range of anti-Muslim policies. Importantly, reliance on social media is linked to policy attitudes across the partisan divide and for individuals who reported holding positive or negative feelings towards Muslims.** These findings highlight the need for further investigation into the political ramification of information presented on contemporary social media outlets, particularly information related to stigmatised groups.

**1.1.14** Brady, McLoughlin, Torres, Luo, Gendron, & Crockett (pre-print). *Overperception of moral outrage in online social networks inflates beliefs about intergroup hostility*. *OSF.*

ABSTRACT: As individuals and political leaders increasingly interact in online social networks, it is important to understand how the affordances of social media shape social knowledge of morality and politics. Here, we propose that social media users overperceive levels of moral outrage felt by individuals and groups, inflating beliefs about intergroup hostility. Utilizing a Twitter field survey, we measured authors' moral outrage in real time and compared authors' reports to observers' judgments of the authors' moral outrage. We find that observers systematically overperceive moral outrage in authors, inferring more intense moral outrage experiences from messages than the authors of those messages actually reported. This effect was stronger in participants who spent more time on social media to learn about politics. Pre-registered confirmatory behavioral experiments found that **overperception of individuals' moral outrage causes overperception of collective moral outrage and inflates beliefs about hostile communication norms, group affective polarization and ideological extremity.** Together, these results highlight how individual-level overperceptions of online moral outrage produce collective overperceptions that have the potential to warp our social knowledge of moral and political attitudes.

**1.1.15** Kim, Guess, Nyhan, & Reifler (2021). The Distorting Prism of Social Media: How Self-Selection and Exposure to Incivility Fuel Online Comment Toxicity. *Journal of Communication*.

ABSTRACT: Though prior studies have analyzed the textual characteristics of online comments about politics, less is known about how selection into commenting behavior and exposure to other people's comments changes the tone and content of political discourse. This article makes three contributions. First, **we show that frequent commenters on Facebook are more likely to be interested in politics, to have more polarized opinions, and to use toxic language in comments in an elicitation task. Second, we find that people who comment on articles in the real world use more toxic language on average than the public as a whole; levels of toxicity in comments scraped from media outlet Facebook pages greatly exceed what is observed in comments we elicit on the same articles from a nationally representative sample. Finally, we demonstrate experimentally that exposure to toxic language in comments increases the toxicity of subsequent comments.**

**1.1.16** Brady, McLoughlin, Torres, Luo, Gendron, & Crockett (2023). Overperception of moral outrage in online social networks inflates beliefs about intergroup hostility. *Nature Human Behaviour.*

ABSTRACT: As individuals and political leaders increasingly interact in online social networks, it is important to understand the dynamics of emotion perception online. Here, we propose that social media users overperceive levels of moral outrage felt by individuals and groups, inflating beliefs about intergroup hostility. Using a Twitter field survey, we measured authors' moral outrage in real time and compared authors' reports to observers' judgements of the authors' moral outrage. **We find that observers systematically overperceive moral outrage in authors, inferring more intense moral outrage experiences from messages than the authors of those messages actually reported. This effect was stronger in participants who spent more time on social media to learn about politics.** Preregister**ed confirmatory behavioural experiments found that overperception of individuals' moral outrage causes overperception of collective moral outrage and inflates beliefs about hostile communication norms, group affective polarization and ideological extremity.** Together, these results highlight how individual-level overperceptions of online moral outrage produce collective overperceptions that have the potential to warp our social knowledge of moral and political attitudes.

**1. 1. 17** Oldemburgo de Mello, Cheung and Inzlicht (2024). Twitter (X) use predicts substantial changes in well-being, polarization, sense of belonging, and outrage. *Communication Psychology*.

In public debate, Twitter (now X) is often said to cause detrimental effects on users and society. Here we address this research question by querying 252 participants from a representative sample of U.S. Twitter users 5 times per day over 7 days (6,218 observations). Results revealed that **Twitter use is related to decreases in well-being, and increases in political polarization, outrage, and sense of belonging over the course of the following 30 minute**s. Effect sizes were comparable to the effect of social interactions on well-being. These effects remained consistent even when accounting for demographic and personality traits. Different inferred uses of Twitter were linked to different outcomes: **passive usage was associated with lower well-being**, social usage with a higher sense of belonging, and **information-seeking usage with increased outrage** and most effects were driven by within-person changes.

[Other studies? What have we missed?]

## 1.2 STUDIES INDICATING NO

**1.2.1*** Boxell, Gentzkow, & Shapiro (2017). Greater Internet use is not associated with faster growth in political polarization among US demographic groups. *Proceedings of the National Academy of Sciences (PNAS).*

ABSTRACT: We combine eight previously proposed measures to construct an index of political polarization among US adults. **We find that polarization has increased the most among the demographic groups least likely to use the Internet and social media. Our overall index and all but one of the individual measures show greater increases for those older than 65 than for those aged 18–39.** A linear model estimated at the age-group level implies that the Internet explains a small share of the recent growth in polarization.

[**NOTE from JH**: This study makes the important point that the oldest generations show the highest levels of polarization, including affective polarization. This suggests that partisan cable TV, which is consumed most heavily by older Americans, may

be playing a substantial role in causing political polarization; we should not just be looking at "the internet" and social media]

[**NOTE from CB**: This study cannot completely disentangle "age" vs. "period" and "cohort" effects— meaning that we cannot know whether the effects are driven by age, or the political socialization of older generations— as well as current political conditions.

**1.2.2** Boxell, Gentzkow, & Shapiro (2021). Cross-country trends in affective polarization. *National Bureau of Economic Research.*

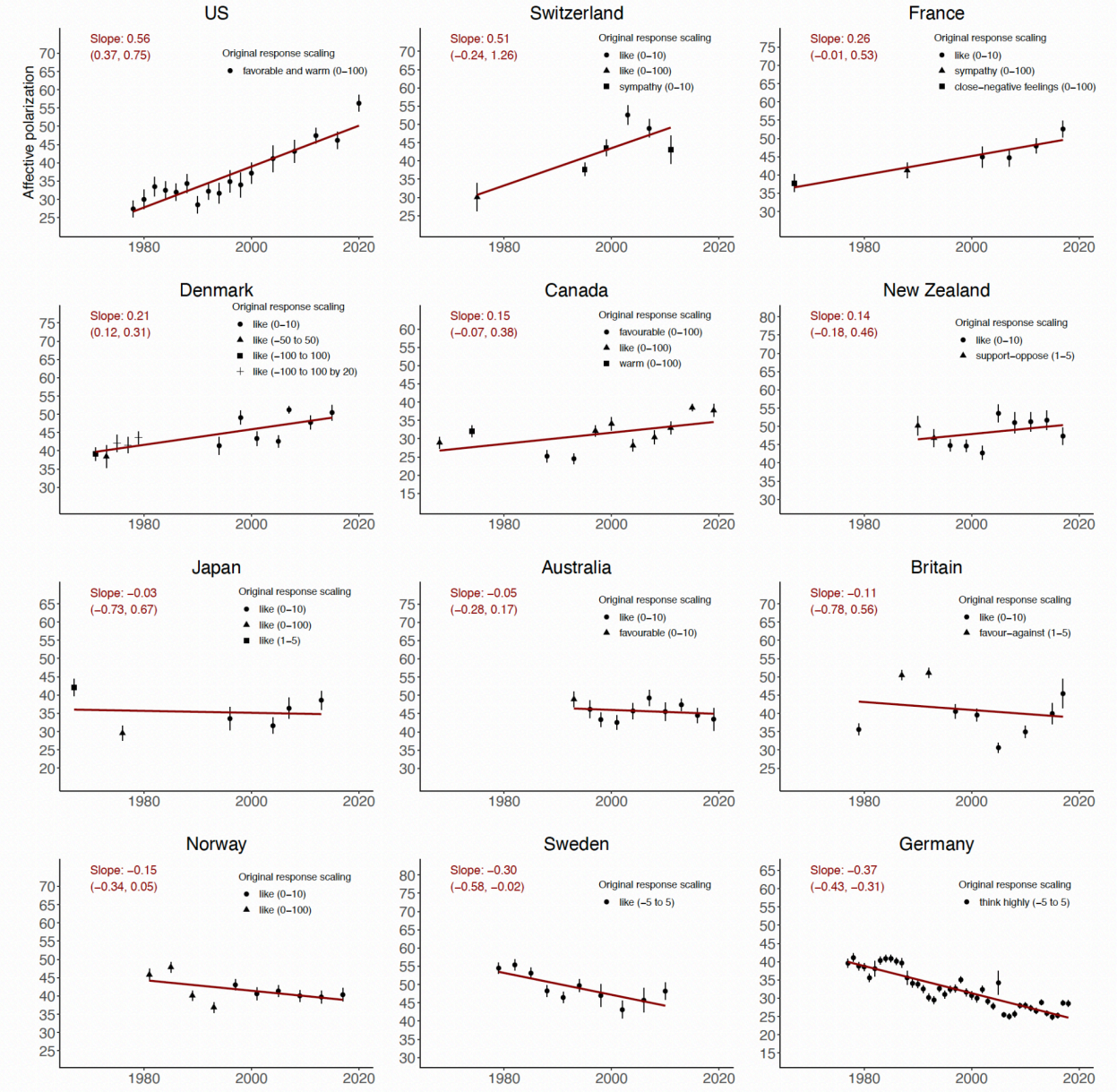
ABSTRACT: We measure trends in affective polarization in twelve OECD countries over the past four decades. **According to our baseline estimates, the US experienced the largest increase in polarization over this period.** Five countries experienced a smaller increase in polarization. Six countries experienced a decrease in polarization. We relate trends in polarization to trends in potential explanatory factors.

**[Note from JH:** This is an important paper. Early drafts only had data up through 2012, but the most recent revision, in 2021, includes a number of data points after 2016, which is much more informative for the questions we ask in this review. However, the question under examination is whether social media became a destructive force only after around 2012, so the long term trend line, since the 1980s, does not help us answer that question. What we need is a hinge at 2012 or 2014. I asked Matt Gentzkow if he could put a "hinge" in the data in the early 2010s, and he said there is not enough data after that to make the analysis reliable.]

[**Note from CB**: It is extremely difficult to determine whether social media drives political polarization by analyzing correlations between the two factors in just twelve countries. As we noted in our introduction, there are myriad factors that shape polarization beyond social media— and these may be responsible for the trends depicted in the figures below, particularly insofar as many of the downward or upward trends pre-date the rise of social media. It is also important to note that this article only examines one type of polarization (affective polarization).]

Figure 1, on p. 20 of the 2021 revision:

Figure 1: Trends in Affective Polarization by Country




**1.2.3** Beam, Hutchens, & Hmielowski (2018). Facebook news and (de)polarization: Reinforcing spirals in the 2016 US election. *Information, Communication & Society*.

ABSTRACT: The rise of social media, and specifically Facebook, as a dominant force in the flow of news in the United States has led to concern that people incur greater isolation from diverse perspectives through filter bubbles (from algorithmic filtering) and

echo chambers (from an information environment populated by social recommendations coming from overwhelmingly like-minded others). This evolution in news diffusion comes at a time when Americans report increased affective partisan polarization. In particular, evidence shows increasingly negative attitudes about out-party members. Based on selective exposure and reinforcing spirals model perspectives, we examined the reciprocal relationship between Facebook news use and polarization using national 3-wave panel data collected during the 2016 US Presidential Election. **Over the course of the campaign, we found media use and attitudes remained relatively stable. Our results also showed that Facebook news use was related to a modest over-time spiral of depolarization. Furthermore, we found that people who use Facebook for news were more likely to view both pro- and counter-attitudinal news in each wave. Our results indicated that counter-attitudinal news exposure increased over time, which resulted in depolarization.** We found no evidence of a parallel model, where pro-attitudinal exposure stemming from Facebook news use resulted in greater affective polarization.

**1.2.4** Nordbrandt (2021). Affective polarization in the digital age: Testing the direction of the relationship between social media and users' feelings for out-group parties. *New Media & Society*.

ABSTRACT: There is considerable disagreement among scholars as to whether social media fuels polarization in society. However, a few have considered the possibility that polarization may instead affect social media usage. To address this gap, the study uses Dutch panel data to test directionality in the relationship between social media use and affective polarization. **No support was found for the hypothesis that social media use contributed to the level of affective polarization. Instead, the results lend support to the hypothesis that it was the level of affective polarization that affected subsequent use of social media.** *The results furthermore reveal heterogeneous patterns among individuals, depending on their previous level of social media usage, and across different social media platforms*. The study gives reason to call into question the predominating assumption in previous research that social media is a major driver of polarization in society.

**1.2.5** Waller, & Anderson (2021). Quantifying social organization and political polarization in online platforms. *Nature*.

ABSTRACT: Mass selection into groups of like-minded individuals may be fragmenting and polarizing online society, particularly with respect to partisan differences. However, our ability to measure the social makeup of online communities and in turn, to understand the social organization of online platforms, is limited by the pseudonymous, unstructured and large-scale nature of digital discussion. Here we develop a neural-embedding methodology to quantify the positioning of online communities along social dimensions by leveraging large-scale patterns of aggregate behaviour. Applying our methodology to 5.1 billion comments made in 10,000 communities over 14 years on Reddit, we measure how the macroscale community structure is organized with respect to age, gender and US political partisanship. **Examining political content, we find that Reddit underwent a significant polarization event around the 2016 US presidential election. Contrary to conventional wisdom, however, individual-level polarization is rare; the system-level shift in 2016 was disproportionately driven by the arrival of new users. Political polarization on Reddit is unrelated to previous activity on the platform and is instead temporally aligned with external events.** We also observe a stark ideological asymmetry, with the sharp increase in polarization in 2016 being entirely attributable to changes in right-wing activity. This methodology is broadly applicable to the study of online interaction, and our findings have implications for the design of online platforms, understanding the social contexts of online behaviour, and quantifying the dynamics and mechanisms of online polarization. [NOTE: this study is also posted in section 2.1, because it shows that Reddit facilitated the creation of politically homogeneous subreddits on the right]

**1.2.6** Munger, Luca, Nagler, & Tucker (2020). The (null) effects of clickbait headlines on polarization, trust, and learning. *Public Opinion Quarterly*.

ABSTRACT: “Clickbait” headlines designed to entice people to click are frequently used by both legitimate and less-than-legitimate news sources. Contemporary clickbait headlines tend to use emotional partisan appeals, raising concerns about their impact on consumers of online news. This article reports the results of a pair of experiments with different sets of subject pools: one conducted using Facebook ads that explicitly target people with a high preference for clickbait, the other using a sample recruited from Amazon’s Mechanical Turk. We estimate subjects’ individual-level preference for clickbait, and randomly assign sets of subjects to read either clickbait or traditional headlines. **Findings show that older people and non-Democrats have a higher “preference for clickbait,” but reading clickbait headlines does not drive affective polarization, information retention, or trust in media.**

**1.2.7** Mukerjee, Jaidka, & Lelkes (2022). The political landscape of the U.S. Twitterverse. *Political Communication*.

ABSTRACT: Prior research suggests that Twitter users in the United States are more politically engaged and more partisan than the American citizenry, who are generally characterized by low levels of political knowledge and disinterest in political affairs. This study seeks to understand this disconnect by conducting an observational analysis of the most popular accounts on American Twitter. We identify opinion leaders by drawing random samples of ordinary American Twitter users and observing whom they follow. We estimate the ideological leaning and political relevance of these opinion leaders and crowdsource estimates of perceived ideology. We find little evidence that American Twitter is as politicized as it is made out to be, with politics and hard news outlets constituting a small subset of these opinion leaders. **Ordinary Americans are significantly more likely to follow nonpolitical opinion leaders on Twitter than political opinion leaders. We find no evidence of polarization among these opinion leaders either.** While a few political professional categories are more polarized than others, the overall polarization dissipates when we factor in the rate at which the opinion leaders tweet: a large number of vocal nonpartisan opinion leaders drowns out the partisan voices on the platform. Our results suggest that **the degree to which Twitter is political has likely been overstated in the past.** Our findings have implications about how we use Twitter and social media, in general, to represent public opinion in the United States.

**1.2.8** Smith, Piwek, Hinds, Brown, & Joinson (2023). Digital traces of offline mobilization.Journal of Personality and Social Psychology. Advance online publication.

ABSTRACT: Since 2009, there has been an increase in global protests and related online activity. Yet, it is unclear how and why online activity is related to the mobilization of offline collective action. One proposition is that online polarization (or a relative change in intensity of posting mobilizing content around a salient grievance) can mobilize people offline. The identity-norm nexus and normative alignment models of collective action further argue that to be mobilizing, these posts need to be socially validated. To test these propositions, across two analyses, we used digital traces of online behavior and data science techniques to model people's online and offline behavior around a mass protest. In Study 1a, we used Twitter behavior posted on the day of the protest by attendees or nonattendees (759 users; 7,592 tweets) to train and

test a classifier that predicted, with 80% accuracy, who participated in offline collective action. Attendees used their mobile devices to plan logistics and broadcast their presence at the protest. In Study 1b, using the longitudinal Twitter data and metadata of a subset of users from Study 1a (209 users; 277,556 tweets), **we found that participation in the protest was not associated with an individual's online polarization over the year prior to the protest, but it was positively associated with the validation ("likes") they received on their relevant posts.** These two studies demonstrate that rather than being low cost or trivial, socially validated online interactions about a grievance are actually key to the mobilization and enactment of collective action.

**1.2.9** Nyhan… & Tucker (2023). Like-minded sources on Facebook are prevalent but not polarizing. *Nature*.

ABSTRACT: Many critics raise concerns about the prevalence of 'echo chambers' on social media and their potential role in increasing political polarization. However, the lack of available data and the challenges of conducting large-scale field experiments have made it difficult to assess the scope of the problem. Here we present data from 2020 for the entire population of active adult Facebook users in the USA showing that content from 'like-minded' sources constitutes the majority of what people see on the platform, although political information and news represent only a small fraction of these exposures. To evaluate a potential response to concerns about the effects of echo chambers, we conducted a multi-wave field experiment on Facebook among 23,377 users for whom we reduced exposure to content from like-minded sources during the 2020 US presidential election by about one-third. **We found that the intervention increased their exposure to content from cross-cutting sources and decreased exposure to uncivil language, but had no measurable effects on eight preregistered attitudinal measures such as affective polarization, ideological extremity, candidate evaluations and belief in false claims.** These precisely estimated results suggest that although exposure to content from like-minded sources on social media is common, reducing its prevalence during the 2020 US presidential election did not correspondingly reduce polarization in beliefs or attitudes.

**1.2.10** Guess… & Tucker (2023). Reshares on social media amplify political news but do not detectably affect beliefs or opinions. *Science*.

ABSTRACT: We studied the effects of exposure to reshared content on Facebook during the 2020 US election by assigning a random set of consenting, US-based users to feeds that did not contain any reshares over a 3-month period. We find that removing reshared content substantially decreases the amount of political news, including content from untrustworthy sources, to which users are exposed; **decreases overall clicks and reactions; and reduces partisan news clicks.** Further, we observe that removing reshared content produces clear decreases in news knowledge within the sample, although there is some uncertainty about how this would generalize to all users. **Contrary to expectations, the treatment does not significantly affect political polarization or any measure of individual-level political attitudes.**

**1.2.11** Guess… & Tucker (2023). How do social media feed algorithms affect attitudes and behavior in an election campaign? *Science*.

ABSTRACT: We investigated the effects of Facebook's and Instagram's feed algorithms during the 2020 US election. We assigned a sample of consenting users to reverse-chronologically-ordered feeds instead of the default algorithms. Moving users out of algorithmic feeds substantially decreased the time they spent on the platforms and their activity. The chronological feed also affected exposure to content: The amount of political and untrustworthy content they saw increased on both platforms, the amount of content classified as uncivil or containing slur words they saw decreased on Facebook, and the amount of content from moderate friends and sources with ideologically mixed audiences they saw increased on Facebook. **Despite these substantial changes in users' on-platform experience, the chronological feed did not significantly alter levels of issue polarization, affective polarization, political knowledge, or other key attitudes during the 3-month study period.**

**1.2.12** Hosseinmardi et al. (2024). Causally estimating the effect of YouTube's recommender system using counterfactual bots, Proceedings of the National Academy of Sciences

ABSTRACT: In recent years, critics of online platforms have raised concerns about the ability of recommendation algorithms to amplify problematic content, with potentially radicalizing consequences. However, attempts to evaluate the effect of recommenders have suffered from a lack of appropriate counterfactuals -- what a user would have viewed in the absence of algorithmic recommendations -- and hence cannot disentangle the effects of the algorithm from a user's intentions. Here we propose a method that we call "counterfactual bots" to causally estimate the role of algorithmic recommendations on the consumption of highly partisan content on YouTube. **By comparing bots that replicate real users' consumption patterns with**

**"counterfactual" bots that follow rule-based trajectories, we show that, on average, relying exclusively on the YouTube recommender results in less partisan consumption**, where the effect is most pronounced for heavy partisan consumers. Following a similar method, we also show that **if partisan consumers switch to moderate content, YouTube's sidebar recommender "forgets" their partisan preference within roughly 30 videos regardless of their prior history, while homepage recommendations shift more gradually toward moderate content**. Overall, our findings indicate that, at least since the algorithm changes that YouTube implemented in 2019, **individual consumption patterns mostly reflect individual preferences, where algorithmic recommendations play, if anything, a moderating role**.

[Other studies? What have we missed?]

## 1.3 MIXED RESULTS OR UNCLASSIFIED

**1.3.1** Bail, Argyle, Brown, Bumpus, Chen, Hunzaker, … Volfovsky (2018). Exposure to opposing views on social media can increase political polarization. *Proceedings of the National Academy of Sciences.*

ABSTRACT: There is mounting concern that social media sites contribute to political polarization by creating “echo chambers” that insulate people from opposing views about current events. We surveyed a large sample of Democrats and Republicans who visit Twitter at least three times each week about a range of social policy issues. One week later, we randomly assigned respondents to a treatment condition in which they were offered financial incentives to follow a Twitter bot for 1 month that exposed them to messages from those with opposing political ideologies (e.g., elected officials, opinion leaders, media organizations, and nonprofit groups). Respondents were resurveyed at the end of the month to measure the effect of this treatment, and at regular intervals throughout the study period to monitor treatment compliance. **We find that Republicans who followed a liberal Twitter bot became substantially more conservative posttreatment. Democrats exhibited slight increases in liberal attitudes after following a conservative Twitter bot, although these effects are not statistically significant.** Notwithstanding important limitations of our study, these findings have significant implications for the interdisciplinary literature on political polarization and the emerging field of computational social science.

**1.3.2** Bor & Petersen (2021). The psychology of online political hostility: A comprehensive, cross-national test of the mismatch hypothesis. *American Political Science Review*.

ABSTRACT: Why are online discussions about politics more hostile than offline discussions? A popular answer argues that human psychology is tailored for face-to-face interaction and people's behavior therefore changes for the worse in impersonal online discussions. We provide a theoretical formalization and empirical test of this explanation: the mismatch hypothesis. We argue that mismatches between human psychology and novel features of online environments could (a) change people's behavior, (b) create adverse selection effects, and (c) bias people's perceptions. Across eight studies, **leveraging cross-national surveys and behavioral experiments** (total *N* = 8,434), **we test the mismatch hypothesis but only find evidence for limited selection effects. Instead, hostile political discussions are the result of status-driven individuals who are drawn to politics and are equally hostile both online and offline. Finally, we offer initial evidence that online discussions feel more hostile, in part, because the behavior of such individuals is more visible online than offline.**

**1.3.3** Yarchi, Baden, & Kligler-Vilenchik (2021). Political polarization on the digital sphere: A cross-platform, over-time analysis of interactional, positional, and affective polarization on social media. *Political Communication*.

ABSTRACT: Political polarization on the digital sphere poses a real challenge to many democracies around the world. Although the issue has received some scholarly attention, there is a need to improve the conceptual precision in the increasingly blurry debate. The use of computational communication science approaches allows us to track political conversations in a fine-grained manner within their natural settings – the realm of interactive social media. The present study combines different algorithmic approaches to studying social media data in order to capture both the interactional structure and content of dynamic political talk online. We conducted an analysis of political polarization across social media platforms (analyzing Facebook, Twitter, and WhatsApp) over 16 months, with close to a quarter million online contributions regarding a political controversy in Israel. Our comprehensive measurement of interactive political talk enables us to address three key aspects of political polarization: (1) interactional polarization – homophilic versus heterophilic user interactions; (2) positional polarization – the positions expressed, and (3) affective polarization – the emotions and attitudes expressed. **Our findings indicate that political polarization on social media cannot**

**be conceptualized as a unified phenomenon, as there are significant cross-platform differences. While interactions on Twitter largely conform to established expectations (homophilic interaction patterns, aggravating positional polarization, pronounced inter-group hostility), on WhatsApp, de-polarization occurred over time. Surprisingly, Facebook was found to be the least homophilic platform in terms of interactions, positions, and emotions expressed.** Our analysis points to key conceptual distinctions and raises important questions about the drivers and dynamics of political polarization online.

**1.3.4** Allcott, Braghieri, Eichmeyer, & Gentzkow (2020). The welfare effects of social media. *American Economic Review*.

ABSTRACT: The rise of social media has provoked both optimism about potential societal benefits and concern about harms such as addiction, depression, and political polarization. In a randomized experiment, we find that **deactivating Facebook for the four weeks before the 2018 US midterm election** (i) reduced online activity, while increasing offline activities such as watching TV alone and socializing with family and friends; (ii) **reduced both factual news knowledge and political polarization; (iii) increased subjective well-being;** and (iv) caused a large persistent reduction in post-experiment Facebook use. Deactivation reduced post-experiment valuations of Facebook, suggesting that traditional metrics may overstate consumer surplus.

ADDITIONAL EXCERPT: **Deactivation [of Facebook] significantly reduced polarization of views on policy issues and a measure of exposure to polarizing news. Deactivation did not statistically significantly reduce affective polarization (i.e., negative feelings about the other political party) or polarization in factual beliefs about current events, although the coefficient estimates also point in that direction.**

[**Note from JH**: Nick Clegg refers to this article as evidence in his Medium essay to show that facebook is not as problematic/polarizing as many argue. Yes, de-activating FB didn't make people dislike the other side less, but it did reduce other measures of polarization, along with increasing well-being]

[**Note from CB**: One additional issue with this study is that it employs a rather unusual measure of polarization that is related to news consumption (and not more conventional attitudinal measures)]

**1.3.5** Lee, Shin, & Hong (2018). Does social media use really make people politically polarized? Direct and indirect effects of social media use on political polarization in South Korea. *Telematics and Informatics*.

ABSTRACT: To help inform the debate over whether social media is related to political polarization, we investigated the effects of social media use on changes in political view using panel data collected in South Korea (N = 6411) between 2012 and 2016. We found that, although there were no direct effects of social media use, social media indirectly contributed to polarization through increased political engagement. **Those who actively used social network sites were more likely to engage in political processes, which led them to develop more extreme political attitudes over time than those who did not use social network sites.** In particular, we observed a clear trend toward a more liberal direction among both politically neutral users and moderately liberal users. **In this study, we highlight the role of social media in activating political participation, which eventually pushes the users toward the ideological poles. The implications of these findings are discussed.**

[Note: because the polarization effect is not direct, but is a result of political "engagement," we put this study into the "mixed results" category]

**1.3.6** Tella, Gálvez, & Schargrodsky (2021, Working Paper). Does social media cause polarization? Evidence from access to Twitter echo chambers during the 2019 Argentine presidential debate. *National Bureau of Economic Research.*

ABSTRACT: We study how two groups, those inside vs those outside echo chambers, react to a political event when we vary social media status (Twitter). Our treatments mimic two strategies often suggested as a way to limit polarization on social media: they expose people to counter-attitudinal data, and they get people to switch off social media. **Our main result is that subjects that started inside echo chambers became more polarized when these two strategies were implemented.** The only scenario where they did not become more polarized is when they did not even experience the political event. Interestingly, **subjects that were outside echo chambers before our study began experienced no change (or a reduction) in polarization.** We also study a group of non-Twitter users in order to have a simple, offline benchmark of the debate's impact on polarization.

**1.3.7** Feezell, Wagner, & Conroy (2021). Exploring the effects of algorithm-driven news sources on political behavior and polarization. *Computers in Human Behavior*. (h/t Jessica Feezell)

ABSTRACT: Do algorithm-driven news sources have different effects on political behavior when compared to non-algorithmic news sources? Media companies compete for our scarce time and attention; one way they do this is by leveraging algorithms to select the most appealing content for each user. While algorithm-driven sites are increasingly popular sources of information, we know very little about the effects of algorithmically determined news at the individual level. The objective of this paper is to define and measure the impact of algorithmically generated news. We begin by developing a taxonomy of news delivery by distinguishing between two types of algorithmically generated news, socially driven and user-driven, and contrasting these with non-algorithmic news. We follow with an exploratory analysis of the consequences of these news delivery modes on political behavior, specifically political participation and polarization. Using two nationally representative surveys, one of young adults and one of the general population, we find that getting news from sites that use socially driven or user-driven algorithms to generate content corresponds with higher levels of political participation, but that getting news from non-algorithmic sources does not. **We also find that neither non-algorithmic nor algorithmically determined news contribute to higher levels of partisan polarization.** This research helps identify important variation in the consequences of news consumption contingent on the mode of delivery.

[Other studies? What have we missed?]

## 1.4 DISCUSSION OF QUESTION #1

[To come: We will add a discussion section at the end of each of our 7 questions, where Jon, Chris, and other researchers will weigh in on what can be concluded from the preponderance of the evidence about this question. If you are a researcher and want to offer your thoughts in brief form, please request edit access]

* * * * * * * * * * * * * * *

# QUESTION 2: DOES SOCIAL MEDIA CREATE ECHO CHAMBERS?

There is widespread concern among journalists, policy makers, and others that social media encourages people to surround themselves with people who share their political views. In this section, we scrutinize the available evidence testing this claim across multiple social media platforms. We do not review studies that look at the impact of the internet overall on the creation of social media echo chambers (but see Guess and Goel et al.)

## 2.1 STUDIES INDICATING YES

**2.1.1** Cinelli, Morales, Galeazzi, Quattrociocchi, & Starnini (2021). The echo chamber effect on social media. *Proceedings of the National Academy of Sciences*.

ABSTRACT: Social media may limit the exposure to diverse perspectives and favor the formation of groups of like-minded users framing and reinforcing a shared narrative, that is, echo chambers. However, the interaction paradigms among users and feed algorithms greatly vary across social media platforms. This paper explores the key differences between the main social media platforms and how they are likely to influence information spreading and echo chambers' formation. We perform a comparative analysis of more than 100 million pieces of content concerning several controversial topics (e.g., gun control, vaccination, abortion) **from Gab, Facebook, Reddit, and Twitter.** We quantify echo chambers over social media by two main ingredients: 1) homophily in the interaction networks and 2) bias in the information diffusion toward like-minded peers. **Our results show that the aggregation of users in homophilic clusters dominate online interactions on Facebook and Twitter. We conclude the paper by directly comparing news consumption on Facebook and Reddit, finding higher segregation on Facebook.**

**2.1.2** Barberá, P. (2015). Birds of the same feather tweet together: Bayesian ideal point estimation using twitter data. *Political Analysis*.

ABSTRACT: Political actors and citizens increasingly engage in political conversations on social media outlets such as Twitter. In this paper I show that the structure of the social networks in which they are embedded has the potential to become a source of information about policy positions. Under the assumption that social networks are homophilic, I develop a Bayesian Spatial Following model that scales Twitter users along a common ideological dimension based on who they follow. I apply this network-based method to estimate ideal points for a large sample of Twitter users in the US, the UK, Spain, Germany, Italy, and the Netherlands. The resulting positions of the party accounts on Twitter are highly correlated with offline measures based on their voting records and their manifestos. Similarly, this method is able to successfully classify individuals who state their political orientation publicly, and a sample of users from the state of Ohio whose Twitter accounts are matched with their voter registration history. To illustrate the potential contribution of these estimates, **I examine the extent to which online behavior is polarized along ideological lines. Using the 2012 US presidential election campaign as a case study, I find that public exchanges on Twitter take place predominantly among users with similar viewpoints.**

**2.1.3** Hong & Kim (2016). Political polarization on twitter: Implications for the use of social media in digital governments. *Government Information Quarterly*.

ABSTRACT: This study investigates two competing opinions regarding the role of social media platforms in partisan polarization. The “echo chambers” view focuses on the highly fragmented, customized, and niche-oriented aspects of social media and suggests these venues foster greater political polarization of public opinion. An alternative, which we term the “crosscutting interactions” view, focuses on the openness of the Internet and social media, with different opinions just a click away. This view thus argues that polarization would not be especially problematic on these outlets. Exploiting the variation among members of the U.S. House of Representatives in measured positions of political ideology, this study estimates the association between politicians' ideological positions and the size of their Twitter readership. **The evidence shows a strong polarization on Twitter readership, which supports the echo chambers view.** Lastly, we discuss the implications of this evidence for governments' use of social media in collecting new ideas and opinions from the public.

**2.1.4** Mosleh, Martel, Eckles, & Rand (2021). Shared partisanship dramatically increases social tie formation in a Twitter field experiment. *Proceedings of the National Academy of Sciences*.

ABSTRACT: Americans are much more likely to be socially connected to copartisans, both in daily life and on social media. However, this observation does not necessarily mean that shared partisanship per se drives social tie formation, because partisanship is confounded with many other factors. Here, we test the causal effect of shared partisanship on the formation of social ties in a field experiment on Twitter. We created bot accounts that self-identified as people who favored the Democratic or Republican party and that varied in the strength of that identification. We then randomly assigned 842 Twitter users to be followed by one of our accounts. **Users were roughly three times more likely to reciprocally follow-back bots whose partisanship matched their own, and this was true regardless of the bot's strength of identification. Interestingly, there was no partisan asymmetry in this preferential follow-back behavior: Democrats and Republicans alike were much more likely to reciprocate follows from copartisans.** These results demonstrate a strong causal effect of shared partisanship on the formation of social ties in an ecologically valid field setting and have important implications for political psychology, social media, and the politically polarized state of the American public.

**2.1.5** Halberstam, & Knight (2016). Homophily, group size, and the diffusion of political information in social networks: Evidence from Twitter. *Journal of Public Economics*.

ABSTRACT: We investigate the role of homophily – a tendency to interact with similar individuals–in the diffusion of political information in social networks. We develop a model predicting disproportionate exposure to likeminded information and that larger groups have more connections and are exposed to more information. To test these hypotheses, we use data on links and communications between politically-engaged Twitter users. We find that users affiliated with majority political groups, relative to the minority group, have more connections, are exposed to more information, and are exposed to information more quickly. **Likewise, we find that users are disproportionately exposed to like-minded information and that information reaches like-minded users more quickly.**

**2.1.6** Waller, & Anderson (2021). Quantifying social organization and political polarization in online platforms. *Nature*.

ABSTRACT: Mass selection into groups of like-minded individuals may be fragmenting and polarizing online society, particularly with respect to partisan differences. However, our ability to measure the social makeup of online communities and in turn, to understand the social organization of online platforms, is limited by the pseudonymous, unstructured and large-scale nature of digital discussion. Here we develop a neural-embedding methodology to quantify the positioning of online communities along social dimensions by leveraging large-scale patterns of aggregate behaviour. Applying our methodology to 5.1 billion comments made in 10,000 communities over 14 years on Reddit, we measure how the macroscale community structure is organized with respect to age, gender and US political partisanship. **Examining political content, we find that Reddit underwent a significant polarization event around the 2016 US presidential election. Contrary to conventional wisdom, however, individual-level polarization is rare; the system-level shift in 2016 was disproportionately driven by the arrival of new users. Political polarization on Reddit is unrelated to previous activity on the platform and is instead temporally aligned with external events.** We also observe a stark ideological asymmetry, with the sharp increase in polarization in 2016 being entirely attributable to changes in right-wing activity. This methodology is broadly applicable to the study of online interaction, and our findings have implications for the design of online platforms, understanding the social contexts of online behaviour, and quantifying the dynamics and mechanisms of online polarization.

[NOTE from ZR: this study is also posted in section 1.2, because it shows that Reddit did not make INDIVIDUALS more polarized, it shifted with new uses so that subreddits on the right became more homogeneous]

**2.1.7** Levy (2021). Social media, news consumption, and polarization: Evidence from a field experiment. *American Economic Review*.

ABSTRACT: Does the consumption of ideologically congruent news on social media exacerbate polarization? I estimate the effects of social media news exposure by conducting a large field experiment randomly offering participants subscriptions to conservative or liberal news outlets on Facebook. I collect data on the causal chain of media effects: subscriptions to outlets, exposure to news on Facebook, visits to online news sites, and sharing of posts, as well as changes in political opinions and attitudes. Four main findings emerge. First, random variation in exposure to news on social media

substantially affects the slant of news sites that individuals visit. **Second, exposure to counter-attitudinal news decreases negative attitudes toward the opposing political party. Third, in contrast to the effect on attitudes, I find no evidence that the political leanings of news outlets affect political opinions. Fourth, Facebook's algorithm is less likely to supply individuals with posts from counter-attitudinal outlets, conditional on individuals subscribing to them.** Together, the **results suggest that social media algorithms may limit exposure to counter-attitudinal news and thus increase polarization.**

[NOTE from JH: this one is complicated. Note the positive effect of counter-attitudinal news, when it happens. But the authors conclude that the overall effect is to limit exposure and increase polarization]

**2.1.8** Sasahara, Chen, Peng, Ciampaglia, Flammini, & Menczer (2021). Social influence and unfollowing accelerate the emergence of echo chambers. *Journal of Computational Social Science*. [h/t Fil Menczer]

ABSTRACT: While social media make it easy to connect with and access information from anyone, they also facilitate basic influence and unfriending mechanisms that may lead to segregated and polarized clusters known as "echo chambers." Here we study the conditions in which such echo chambers emerge by introducing a simple model of information sharing in online social networks with the two ingredients of influence and unfriending. Users can change both their opinions and social connections based on the information to which they are exposed through sharing. **The model dynamics show that even with minimal amounts of influence and unfriending, the social network rapidly devolves into segregated, homogeneous communities.** These predictions are consistent with empirical data from Twitter. Although our findings suggest that echo chambers are somewhat inevitable given the mechanisms at play in online social media, they also provide insights into possible mitigation strategies.

**2.1.9** Shahrezaye, Papakyriakopoulos, Medina Serrano, & Hegelich (2019). Measuring the ease of communication in bipartite social endorsement networks: a proxy to study the dynamics of political polarization. *Proceedings of the 10th International Conference on Social Media and Society*. [h/t Orestis Papkyriakopoulos]

ABSTRACT: In this work, complex weighted bipartite social networks are developed to efficiently analyze, project and extract network knowledge. Specifically, to assess the overall ease of communication between the different network sub-clusters, a proper

projection and measurement method is developed in which the defined measurement is a function of the network structure and preserves maximum relevant information. Using simulations, it is shown how the introduced measurement correlates with the concept of political polarization, after which the proposed method is applied to Facebook networks to demonstrate its ability to capture the polarization dynamics over time. **The method successfully captured the increasing political polarization between the Alternative für Deutschland's (AfD) supporters and the supporters of other political parties, which is in line with previous studies on the rise of the AfD in Germany's political sphere.**

ADDITIONAL EXCERPT: The search information index between the AfD sub-cluster and all other sub-clusters, from which it can be seen that the average search information index between the AfD Facebook posts and the Facebook posts of the other parties was increasing over time. **This implies that the AfD and non-AfD supporters had increased their endorsement activities on the pages connected to their own political orientation, and had decreased their activities on the pages connected to opposite political views.**

**2.1.10.** Boutyline & Willer (2017). The social structure of political echo chambers: Variation in ideological homophily in online networks. *Political Psychology*. (h/t Robb Willer)

ABSTRACT: We predict that people with different political orientations will exhibit systematically different levels of political homophily, the tendency to associate with others similar to oneself in political ideology. Research on personality differences across the political spectrum finds that both more conservative and more politically extreme individuals tend to exhibit greater orientations towards cognitive stability, clarity, and familiarity. We reason that such a "preference for certainty" may make these individuals more inclined to seek out the company of those who reaffirm, rather than challenge, their views. Since survey studies of political homophily face well-documented methodological challenges, we instead test this proposition on a large sample of politically engaged users of the social-networking platform *Twitter*, whose ideologies we infer from the politicians and policy nonprofits they follow. **As predicted, we find that both more extreme and more conservative individuals tend to be more homophilous than more liberal and more moderate ones.**

**2.1.11** Boutyline, & Willer (2017). The social structure of political echo chambers: Variation in ideological homophily in online networks: Political echo chambers. *Political Psychology*.

ABSTRACT: We predict that people with different political orientations will exhibit systematically different levels of political homophily, the tendency to associate with others similar to oneself in political ideology. Research on personality differences across the political spectrum finds that both more conservative and more politically extreme individuals tend to exhibit greater orientations towards cognitive stability, clarity, and familiarity. We reason that such a "preference for certainty" may make these individuals more inclined to seek out the company of those who reaffirm, rather than challenge, their views. Since survey studies of political homophily face well-documented methodological challenges, we instead test this proposition on a large sample of politically engaged users of the social-networking platform Twitter, whose ideologies we infer from the politicians and policy nonprofits they follow. As predicted, **we find that both more extreme and more conservative individuals tend to be more homophilous than more liberal and more moderate ones.**

**2.1.12** Cookson, Engelberg, & Mullins (2020). Echo Chambers. *Soc ArXiV.*

ABSTRACT: We find evidence of selective exposure to confirmatory information among 400,000 users on the investor social network StockTwits. Self-described bulls are 5 times more likely to follow a user with a bullish view of the same stock than self-described bears. Consequently, **bulls see 62 more bullish messages and 24 fewer bearish messages than bears over the same 50-day period. These "echo chambers" exist even among professional investors and are strongest for investors who trade on their beliefs**. Finally, beliefs formed in echo chambers are associated with lower ex-post returns, more siloing of information and more trading volume.

**2.1.13** González-Bailón… & Tucker (2023). Asymmetric ideological segregation in exposure to political news on Facebook. *Science*.

ABSTRACT: Does Facebook enable ideological segregation in political news consumption? We analyzed exposure to news during the US 2020 election using aggregated data for 208 million US Facebook users. We compared the inventory of all political news that users could have seen in their feeds with the information that they

saw (after algorithmic curation) and the information with which they engaged. We show that **(i) ideological segregation is high and increases as we shift from potential exposure to actual exposure to engagement; (ii) there is an asymmetry between conservative and liberal audiences, with a substantial corner of the news ecosystem consumed exclusively by conservatives; and (iii) most misinformation, as identified by Meta's Third-Party Fact-Checking Program, exists within this homogeneously conservative corner, which has no equivalent on the liberal side.** Sources favored by conservative audiences were more prevalent on Facebook's news ecosystem than those favored by liberals.

[Other studies? What have we missed?]

## 2.2 STUDIES INDICATING NO

**2.2.1** Eady, Nagler, Guess, Zilinsky, & Tucker (2019). How many people live in political bubbles on social media? Evidence from linked survey and Twitter data. *SAGE Open*.

ABSTRACT: A major point of debate in the study of the Internet and politics is the extent to which social media platforms encourage citizens to inhabit online "bubbles" or "echo chambers," exposed primarily to ideologically congenial political information. To investigate this question, we link a representative survey of Americans with data from respondents' public Twitter accounts ($N$ = 1,496). We then quantify the ideological distributions of users' online political and media environments by merging validated estimates of user ideology with the full set of accounts followed by our survey respondents ($N$ = 642,345) and the available tweets posted by those accounts ($N$ ~ 1.2 billion). We study the extent to which liberals and conservatives encounter counter-attitudinal messages in two distinct ways: (a) by the accounts they follow and (b) by the tweets they receive from those accounts, either directly or indirectly (via retweets). More than a third of respondents do not follow any media sources, but among those who do, **we find a substantial amount of overlap (51%) in the ideological distributions of accounts followed by users on opposite ends of the political spectrum. At the same time, however, we find asymmetries in individuals' willingness to venture into cross-cutting spaces, with conservatives more likely to follow media and political accounts classified as left-leaning than the reverse.**

Finally, we argue that such choices are likely tempered by online news watching behavior.

ADDITIONAL EXCERPT: Our results provide a nuanced portrait of the information environments of Americans on Twitter. Most critically, **we do not find evidence supporting a strong characterization of "echo chambers" in which the majority of people's sources of news are mutually exclusive and from opposite poles: There is generally more overlap than divergence in the ideological distributions of media accounts followed by the most liberal and most conservative quintiles in our sample. However, we also show that fully 61% of members of the most conservative quintile in our sample follow very few media accounts even as far "left" as the *New York Times*, suggesting their online media diet is quite ideologically constrained.**

**2.2.2** Fletcher, Kalogeropolous, & Nielson (2021). More diverse, more politically varied: How social media, search engines, and aggregators shape news repertoire in the United Kingdom. *New Media & Society.*

ABSTRACT: There is still much to learn about how the rise of new, 'distributed', forms of news access through search engines, social media and aggregators are shaping people's news use. We analyse passive web tracking data from the United Kingdom to make a comparison between direct access (primarily determined by self-selection) and distributed access (determined by a combination of self-selection and algorithmic selection). **We find that (1) people who use search engines, social media and aggregators for news have more diverse news repertoires. However, (2) social media, search engine and aggregator news use is also associated with repertoires where more partisan outlets feature more prominently.** The findings add to the growing evidence challenging the existence of filter bubbles, and highlight alternative ways of characterizing people's online news use.

**2.2.3** Beam, Hutchens, & Hmielowski (2018). Facebook news and (de)polarization: Reinforcing spirals in the 2016 US election. *Information, Communication & Society*.

ABSTRACT: The rise of social media, and specifically Facebook, as a dominant force in the flow of news in the United States has led to concern that people incur greater isolation from diverse perspectives through filter bubbles (from algorithmic filtering) and

echo chambers (from an information environment populated by social recommendations coming from overwhelmingly like-minded others). This evolution in news diffusion comes at a time when Americans report increased affective partisan polarization. In particular, evidence shows increasingly negative attitudes about out-party members. Based on selective exposure and reinforcing spirals model perspectives, we examined the reciprocal relationship between Facebook news use and polarization using national 3-wave panel data collected during the 2016 US Presidential Election. **Over the course of the campaign, we found media use and attitudes remained relatively stable. Our results also showed that Facebook news use was related to a modest over-time spiral of depolarization. Furthermore, we found that people who use Facebook for news were more likely to view both pro- and counter-attitudinal news in each wave. Our results indicated that counter-attitudinal news exposure increased over time, which resulted in depolarization.** We found no evidence of a parallel model, where pro-attitudinal exposure stemming from Facebook news use resulted in greater affective polarization.

**2.2.4** Shore, Baek & Dellarocas (2018). Network structure and patterns of information diversity on Twitter. *MIS Quarterly.*

ABSTRACT: Social media have great potential to support diverse information sharing, but there is widespread concern that platforms like Twitter do not result in communication between those who hold contradictory viewpoints. Because users can choose whom to follow, prior research suggests that social media users exist in “echo chambers” or become polarized. We seek evidence of this in a complete cross section of hyperlinks posted on Twitter, using previously validated measures of the political slant of news sources to study information diversity. Contrary to prediction, **we find that the average account posts links to more politically moderate news sources than the ones they receive in their own feed.** However, members of **a tiny network core do exhibit cross-sectional evidence of polarization and are responsible for the majority of tweets received overall** due to their popularity and activity, which could explain the widespread perception of polarization on social media.

[NOTE from JH: this study connects to Michael Bang-Petersen’s work, on how the platforms don’t make people trollish; rather they empower trolls to reach many more people]

**2.2.5** Fletcher, Robertson, & Nielsen (2021). How many people live in politically partisan online news echo chambers in different countries? *Journal of Quantitative Description: Digital Media*.

ABSTRACT: Concern over online news echo chambers has been a consistent theme in recent debates on how people get news and information. Yet, we lack a basic descriptive understanding of how many people occupy bounded online news spaces in different countries. Using online survey data from seven countries we find that (i) **politically partisan left-right online news echo chambers are real, but only a minority of approximately 5% of internet news users inhabit them, (ii) in every country covered, more people consume no online news at all than occupy partisan online echo chambers, and (iii) except for the US, decisions over the inclusion or exclusion of particular news outlets make little difference to echo chamber estimates.** Differences within and between media systems mean we should be very cautious about direct comparisons between different echo chambers, but underlying patterns of audience overlap, and the continued popularity of mainstream outlets, often preclude the formation of large partisan echo chambers.

**2.2.6** Boulianne, Koc-Michalska, & Bimber (2020). Right-wing populism, social media and echo chambers in Western democracies. *New Media & Society*.

ABSTRACT: Many observers are concerned that echo chamber effects in digital media are contributing to the polarization of publics and in some places to the rise of right-wing populism. This study employs survey data collected in France, the United Kingdom, and the United States (1500 respondents in each country) from April to May 2017. **Overall, we do not find evidence that online/social media explain support for right-wing populist candidates and parties. Instead, in the USA, use of online media decreases support for right-wing populism.** Looking specifically at echo chambers measures, we find offline discussion with those who are similar in race, ethnicity, and class positively correlates with support for populist candidates and parties in the UK and France. The findings challenge claims about the role of social media and the rise of populism.

**2.2.7** Johnson, Kaye, & Lee (2017). Blinded by the spite? Path model of political attitudes, selectivity, and social media. *Atlantic Journal of Communication*.

ABSTRACT: Despite fears that selective exposure and selective avoidance could deepen polarization and negatively affect the democratic process, few studies have directly studied this phenomenon. This study explores whether selective exposure and avoidance to blogs, social network sites, and Twitter directly influence confidence in Congress and the president or more indirectly through polarization. **This study suggests that fears of selective exposure, selective avoidance, and polarization infecting the democratic process appear overstated. First, polarization was positively related to confidence in Congress and the president. Second, selective exposure to social media sites strengthens confidence in the president and in Congress. Twitter boosts confidence in Congress. T**hird, selective avoidance had a negative influence on other measures, which suggests people seek both information that challenges their views as well as ones that supports them. **Finally, selective exposure and avoidance proved weak indicators of polarization. I**nstead, strength of partisanship is the stronger predictor of confidence in Congress and the president.

**2.2.8** Scharkow, Mangold, Stier, & Breuer (2020). How social network sites and other online intermediaries increase exposure to news. *Proceedings of the National Academy of Sciences*.

ABSTRACT: Research has prominently assumed that social media and web portals that aggregate news restrict the diversity of content that users are exposed to by tailoring news diets toward the users' preferences. In our empirical test of this argument, we apply a random-effects within–between model to two large representative datasets of individual web browsing histories. This approach allows us to better encapsulate the effects of social media and other intermediaries on news exposure. **We find strong evidence that intermediaries foster more varied online news diets. The results call into question fears about the vanishing potential for incidental news exposure in digital media environments.**

**2.2.9** Stier, Mangold, Scharkow, & Breuer (2021). Post post-broadcast democracy? News exposure in the age of online intermediaries. *American Political Science Review*.

ABSTRACT: Online intermediaries such as social network sites or search engines are playing an increasingly central role in democracy by acting as mediators between information producers and citizens. Academic and public commentators have raised persistent concerns that algorithmic recommender systems would negatively affect the

provision of political information by tailoring content to the predispositions and entertainment preferences of users. At the same time, recent research indicates that intermediaries foster exposure to news that people would not use as part of their regular media diets. This study investigates these unresolved questions by combining the web browsing histories and survey responses of more than 7,000 participants from six major democracies. **The analysis shows that despite generally low levels of news use, using online intermediaries fosters exposure to nonpolitical and political news across countries and personal characteristics.** The findings have implications for scholarly and public debates on the challenges that high-choice digital media environments pose to democracy

**2.2.10** Dubois, & Blank (2018). The echo chamber is overstated: The moderating effect of political interest and diverse media. *Information, Communication & Society*.

ABSTRACT: In a high-choice media environment, there are fears that individuals will select media and content that reinforce their existing beliefs and lead to segregation based on interest and/or partisanship. This could lead to partisan echo chambers among those who are politically interested and could contribute to a growing gap in knowledge between those who are politically interested and those who are not. However, the high-choice environment also allows individuals, including those who are politically interested, to consume a wide variety of media, which could lead them to more diverse content and perspectives. This study examines the relationship between political interest as well as media diversity and being caught in an echo chamber (measured by five different variables). Using a nationally representative survey of adult internet users in the United Kingdom (N = 2000), **we find that those who are interested in politics and those with diverse media diets tend to avoid echo chambers. This work challenges the impact of echo chambers and tempers fears of partisan segregation since only a small segment of the population are likely to find themselves in an echo chamber.** We argue that single media studies and studies which use narrow definitions and measurements of being in an echo chamber are flawed because they do not test the theory in the realistic context of a multiple media environment.

**2.2.11** Nelson, & Webster (2017). The myth of partisan selective exposure: A portrait of the online political news audience: *Social Media + Society*.

ABSTRACT: Many assume that in a digital environment with a wide range of ideologically tinged news outlets, partisan selective exposure to like-minded speech is pervasive and a primary cause of political polarization. Yet, partisan selective exposure research tends to stem from experimental or self-reported data, which limits the applicability of their findings in a high-choice media environment. We explore observed online audience behavior data to present a portrait of the actual online political news audience. **We find that this audience frequently navigates to news sites from Facebook, and that it congregates among a few popular, well-known political news sites. We also find that political news sites comprise ideologically diverse audiences, and that they share audiences with nearly all smaller, more ideologically extreme outlets.** Our results call into question the strength of the so-called red/blue divide in actual web use.

**2.2.12** Liang, Hai (2018). Broadcast versus viral spreading: The structure of diffusion cascades and selective sharing on social media. *Journal of Communication.* [h/t Mike Burnham]

Sharing cross-ideological messages on social media exposes people to political diversity and generates other benefits for society. This study argues that the diffusion patterns of political messages can influence the degree of selective sharing. **Using a large-scale diffusion dataset from Twitter, this study found that messages that spread through multiple steps are more likely to involve cross-ideological sharing. Furthermore, the study found that this positive relationship is mediated by the distance between the sharers and originators of the messages and suppressed by the number of connections among the sharers. Overall, the study found that the viral diffusion model, in contrast to the broadcast model, increases the likelihood of cross-ideological sharing and thus increases political diversity on social media.**

**2.2.13** Muise, … & Watts (2022). Quantifying partisan news diets in Web and TV audiences. *Science Advances.*

ABSTRACT: Partisan segregation within the news audience buffers many Americans from countervailing political views, posing a risk to democracy. Empirical studies of the online media ecosystem suggest that only a small minority of Americans, driven by a mix of demand and algorithms, are siloed according to their political ideology. However, such research omits the comparatively larger television audience and often ignores

temporal dynamics underlying news consumption. By analyzing billions of browsing and viewing events between 2016 and 2019, with a novel framework for measuring partisan audiences, we first estimate that 17% of Americans are partisan-segregated through television versus roughly 4% online. Second, television news consumers are several times more likely to maintain their partisan news diets month-over-month. Third, TV viewers' news diets are far more concentrated on preferred sources. Last, partisan news channels' audiences are growing even as the TV news audience is shrinking. **Our results suggest that television is the top driver of partisan audience segregation among Americans.**

**2.2.14** Tornberg (2022). How digital media drive affective polarization through partisan sorting.

ABSTRACT: Politics has in recent decades entered an era of intense polarization. Explanations have implicated digital media, with the so-called echo chamber remaining a dominant causal hypothesis despite growing challenge by empirical evidence. **This paper suggests that this mounting evidence provides not only reason to reject the echo chamber hypothesis but also the foundation for an alternative causal mechanism.** To propose such a mechanism, the paper draws on the literatures on affective polarization, digital media, and opinion dynamics. From the affective polarization literature, we follow the move from seeing polarization as diverging issue positions to rooted in sorting: an alignment of differences which is effectively dividing the electorate into two increasingly homogeneous megaparties. To explain the rise in sorting, the paper draws on opinion dynamics and digital media research to present a model which essentially turns the echo chamber on its head: it is not isolation from opposing views that drives polarization but precisely the fact that digital media bring us to interact outside our local bubble. When individuals interact locally, the outcome is a stable plural patchwork of cross-cutting conflicts. **By encouraging nonlocal interaction, digital media drive an alignment of conflicts along partisan lines, thus effacing the counterbalancing effects of local heterogeneity. The result is polarization, even if individual interaction leads to convergence.** The model thus suggests that digital media polarize through partisan sorting, creating a maelstrom in which more and more identities, beliefs, and cultural preferences become drawn into an all-encompassing societal division.

See relevant twitter thread by lead author, Petter Tornberg.

[Other studies? What have we missed?]

## 2.3 MIXED RESULTS OR UNCLASSIFIED

**2.3.1** Chen, Pacheco, Yang, & Menczer (2021). Neutral bots probe political bias on social media. *Nature Communications*.

ABSTRACT: Social media platforms attempting to curb abuse and misinformation have been accused of political bias. We deploy neutral social bots who start following different news sources on Twitter, and track them to probe distinct biases emerging from platform mechanisms versus user interactions. We find no strong or consistent evidence of political bias in the news feed. Despite this, **the news and information to which U.S. Twitter users are exposed depend strongly on the political leaning of their early connections. The interactions of conservative accounts are skewed toward the right, whereas liberal accounts are exposed to moderate content shifting their experience toward the political center.** Partisan accounts, especially conservative ones, tend to receive more followers and follow more automated accounts. Conservative accounts also find themselves in denser communities and are exposed to more low-credibility content.

[Note from JH: conservatives shift toward extreme; liberals toward the center. Mixed results]

**2.3.2** Barberá, Jost, Nagler, Tucker, & Bonneau (2015). Tweeting from left to right. *Psychological Science.*

ABSTRACT: We estimated ideological preferences of 3.8 million Twitter users and, using a data set of nearly 150 million tweets concerning 12 political and nonpolitical issues, explored whether online communication resembles an “echo chamber” (as a result of selective exposure and ideological segregation) or a “national conversation.” We observed that **information was exchanged primarily among individuals with similar ideological preferences in the case of political issues (e.g., 2012 presidential election, 2013 government shutdown) but not many other current events** (e.g., 2013 Boston Marathon bombing, 2014 Super Bowl). Discussion of the Newtown shootings in 2012 reflected a dynamic process, beginning as a national conversation before transforming into a polarized exchange. **With respect to both political and nonpolitical issues, liberals were more likely than conservatives to engage in cross-ideological dissemination; this is an important asymmetry with**

**respect to the structure of communication that is consistent with psychological theory and research bearing on ideological differences in epistemic, existential, and relational motivation. Overall, we conclude that previous work may have overestimated the degree of ideological segregation in social-media usage.**

**2.3.3** Bakshy, Messing, & Adamic (2015). Exposure to ideologically diverse news and opinion on Facebook. *Science*.

ABSTRACT: Exposure to news, opinion, and civic information increasingly occurs through social media. How do these online networks influence exposure to perspectives that cut across ideological lines? Using deidentified data, we examined how 10.1 million U.S. Facebook users interact with socially shared news. We directly measured ideological homophily in friend networks and examined the extent to which heterogeneous friends could potentially expose individuals to cross-cutting content. We then quantified the extent to which individuals encounter comparatively more or less diverse content while interacting via Facebook's algorithmically ranked News Feed and further studied users' choices to click through to ideologically discordant content. **Compared with algorithmic ranking, individuals' choices played a stronger role in limiting exposure to cross-cutting content.**

**EXCERPT: "Although partisans tend to maintain relationships with like-minded contacts, on average more than 20% of an individual's Facebook friends who report an ideological affiliation are from the opposing party, leaving substantial room for exposure to opposing viewpoints.**..Perhaps unsurprisingly, we show that the composition of our friend networks is the most important factor limiting the mix of content encountered in social media. **The way that sharing occurs within these networks is not symmetric: Liberals tend to be connected to fewer friends who share conservative content than are conservatives (who tend to be linked to more friends who share liberal content).**

Within the population under study here, individual choices more than algorithms limit exposure to attitude-challenging content in the context of Facebook."

**2.3.4** Brown, Bisbee, Lai, Bonneau, Nagler, & Tucker (2022). Echo chambers, rabbit holes, and algorithmic bias: How YouTube recommends content to real users. *Social Science Research Network.*

ABSTRACT: To what extent does the YouTube recommendation algorithm push users into echo chambers, ideologically biased content, or rabbit holes? Despite growing popular concern, recent work suggests that the recommendation algorithm is not pushing users into these echo chambers. However, existing research relies heavily on the use of anonymous data collection that does not account for the personalized nature of the recommendation algorithm. We asked a sample of real users to install a browser extension that downloaded the list of videos they were recommended. We instructed these users to start on an assigned video and then click through 20 sets of recommendations, capturing what they were being shown in real time as they used the platform logged into their real accounts. Using a novel method to estimate the ideology of a YouTube video, **we demonstrate that the YouTube recommendation algorithm does, in fact, push real users into mild ideological echo chambers where, by the end of the data collection task, liberals and conservatives received different distributions of recommendations from each other, though this difference is small.** While we find evidence that this difference increases the longer the user followed the recommendation algorithm, **we do not find evidence that many go down `rabbit holes' that lead them to ideologically extreme content. Finally, we find that YouTube pushes all users, regardless of ideology, towards moderately conservative and an increasingly narrow range of ideological content the longer they follow YouTube's recommendations.**

**2.3.5** Heatherly, Lu, & Lee (2017). Filtering out the other side? Cross-cutting and like-minded discussions on social networking sites. *New Media & Society*.

ABSTRACT: Disagreement persists as to whether social networking sites (SNSs) are used more frequently to facilitate cross-cutting or like-minded discussions. We examine the relationship between the use of SNSs and involvement in discussions with politically similar and dissimilar others among a sample of US Democrats and Republicans. **Affective polarization is negatively related to involvement in cross-cutting discussions, suggesting that individuals extend their dislike of the opposing political party to out-party members within their online social networks.** Moreover, political discussion with one's friends on SNSs plays a mediating role in involvement in both cross-cutting and like-minded discussions. Finally, **party identification moderates the relationship between SNS use and involvement in cross-cutting discussions, indicating that Republicans participate more frequently than Democrats in cross-cutting exchanges on SNSs.** In the light of these findings, we discuss the contribution of SNSs to the ideals of deliberative democracy.

FIGURE:

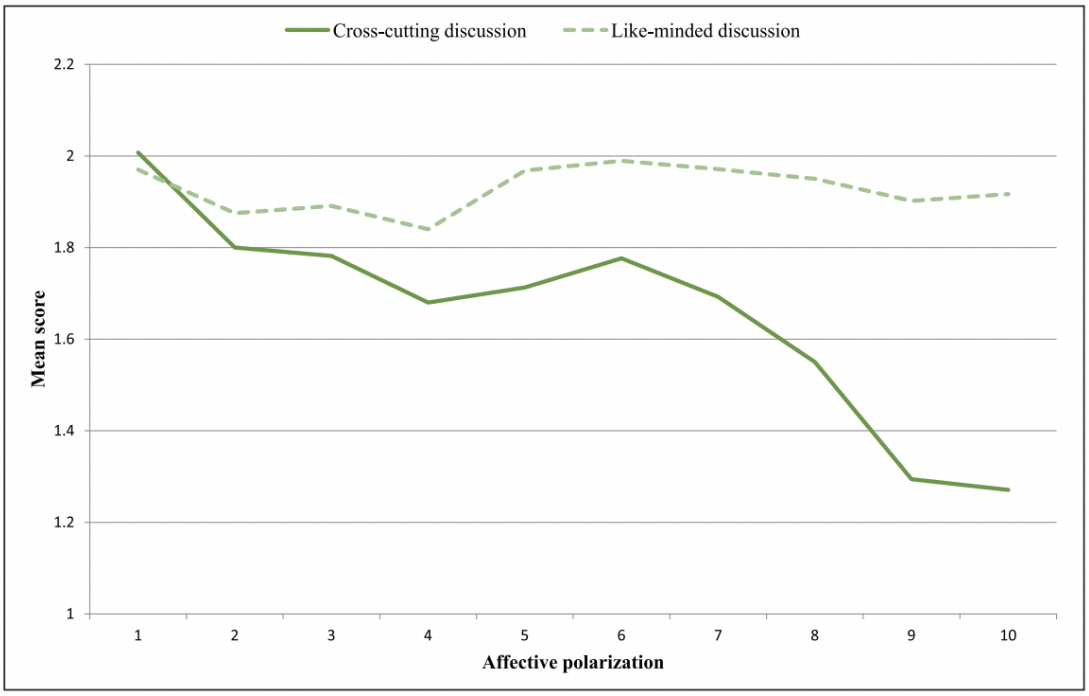



Figure 2. Affective polarization and involvement in cross-cutting and like-minded discussions on SNSs. In Figure 2, affective polarization was transformed to a 10-point scale to aid in interpretation. Y-Axis: . Mean scores of responses to “On social network sites, how often do you talk to people listed below?” (0 - 3).

**2.3.6** Kitchens, Johnson, & Gray (2020). Understanding echo chambers and filter bubbles: The impact of social media on diversification and partisan shifts in news consumption. *MIS Quarterly*.

ABSTRACT: Echo chambers and filter bubbles are potent metaphors that encapsulate widespread public fear that the use of social media may limit the information that users encounter or consume online. Specifically, the concern is that social media algorithms combined with tendencies to interact with like-minded others both limits users’ exposure to diverse viewpoints and encourages the adoption of more extreme ideological positions. Yet empirical evidence about how social media shapes information consumption is inconclusive. We articulate how characteristics of platform algorithms and users’ online social networks may combine to shape user behavior. We bring greater conceptual clarity to this phenomenon by expanding beyond discussion of a binary presence or absence of echo chambers and filter bubbles to a richer set of outcomes incorporating changes in both diversity and slant of users’ information

sources. Using a data set with over four years of web browsing history for a representative panel of nearly 200,000 U.S. adults, we analyzed how individuals' social media usage was associated with changes in the information sources they chose to consume. **We find differentiated impacts on news consumption by platform. Increased use of Facebook was associated with increased information source diversity and a shift toward more partisan sites in news consumption; increased use of Reddit with increased diversity and a shift toward more moderate sites; and increased use of Twitter with little to no change in either. Our results demonstrate the value of adopting a nuanced multidimensional view of how social media use may shape information consumption**

[NOTE from JH: Important point in this and several papers -- that different platforms yield different answers to the questions we ask in this review]

**2.3.7** Jürgens, & Stark (2022). Mapping exposure diversity: The divergent effects of algorithmic curation on news consumption. *Journal of Communication*. (h/t Richard Fletcher)

ABSTRACT: Diversity is a crucial precondition for a democratic public discourse. In today's high-choice media environments, exposure to diverse news is largely determined by individuals' personal selection. Yet these decisions are increasingly shaped by online platforms, whose curation mechanisms may serve to expand or contract the diversity of encountered content. In a major extension of existing research, we show that positive short-term effects of platforms mask detrimental long-term effects. Drawing on a four-month tracking dataset and a comprehensive content analysis covering the online news consumption of over 10,000 German citizens, we demonstrate that even though short-term usage of platforms uniformly increases exposure diversity, long-term reliance can lead to decreases. In addition, **platforms vary in their influences: News aggregators are beneficial to exposure diversity, while Twitter and search engines have a limiting effect; Facebook offers no significant influence.**

**2.3.8** Chen, Nyhan, Reifler, Robertson, & Wilson (2022). Subscriptions and external links help drive resentful users to alternative and extremist YouTube videos. *ArXiv*.

ABSTRACT: Do online platforms facilitate the consumption of potentially harmful content? Despite widespread concerns that YouTube's algorithms send people down "rabbit holes" with recommendations to extremist videos, little systematic evidence exists to support this conjecture. Using paired behavioral and survey data provided by participants recruited from a representative sample (n=1,181), **we show that exposure to alternative and extremist channel videos on YouTube is heavily concentrated among a small group of people with high prior levels of gender and racial resentment.** These viewers typically subscribe to these channels (causing YouTube to recommend their videos more often) and often follow external links to them. Contrary to the "rabbit holes" narrative, **non-subscribers are rarely recommended videos from alternative and extremist channels and seldom follow such recommendations when offered.**

**2.3.9** Tokita, Guess, & Tarnita (2021). Polarized information ecosystems can reorganize social networks via information cascades. *Proceedings of the National Academy of Sciences*.

ABSTRACT: The precise mechanisms by which the information ecosystem polarizes society remain elusive. Focusing on political sorting in networks, **we develop a computational model that examines how social network structure changes when individuals participate in information cascades, evaluate their behavior, and potentially rewire their connections to others as a result.** Individuals follow proattitudinal information sources but are more likely to first hear and react to news shared by their social ties and only later evaluate these reactions by direct reference to the coverage of their preferred source. **Reactions to news spread through the network via a complex contagion.** Following a cascade, individuals who determine that their participation was driven by a subjectively “unimportant” story adjust their social ties to avoid being misled in the future. In our model, this dynamic leads social networks to politically sort when news outlets differentially report on the same topic, even when individuals do not know others’ political identities. **Observational follow network data collected on Twitter support this prediction: We find that individuals in more polarized information ecosystems lose cross-ideology social ties at a rate that is higher than predicted by chance. Importantly, our model reveals that these emergent polarized networks are less efficient at diffusing information: Individuals avoid what they believe to be “unimportant” news at the expense of missing out on subjectively “important” news far more frequently. This suggests that “echo chambers”—to the extent that they exist—may not echo so much as silence.**

[NOTE from Chris Tokita: Our paper studies echo chamber formation on social media; however, we show/suggest that polarized media coverage is what is ultimately creating echo chambers online, as reactions to news coverage spread through social networks and cause people to adjust their social ties. We show that people in more polarized information ecosystems—that is, consuming more partisan news that is out of sync with other sources—lose social ties to people of the opposite ideology, even when they don't know each other's politics. This happens because people compare the behavior of their friends against what their preferred news outlet is reporting and break social ties with friends—some of whom might be consuming other news sources aligned with their personal politics—who appear to be acting "out of sync" with the reality presented by their news source. Therefore, we suggest that ultimately it is the information ecosystem (news coverage) that is reshaping our social networks, without us realizing it, although clearly we focus on how this is playing out on social media.]

**2.3.10** Williams, McMurray, Kurz, & Hugo Lambert (2015). Network analysis reveals open forums and echo chambers in social media discussions of climate change. *Global Environmental Change*. (h/t Olivia Fischer)

ABSTRACT: Action to tackle the complex and divisive issue of climate change will be strongly influenced by public perception. Online social media and associated social networks are an increasingly important forum for public debate and are known to influence individual attitudes and behaviours – yet online discussions and social networks related to climate change are not well understood. Here we construct several forms of social network for users communicating about climate change on the popular microblogging platform Twitter. We classify user attitudes to climate change based on message content and find that social networks are characterised by strong attitude-based homophily and segregation into polarised “sceptic” and “activist” groups. Most users interact only with like-minded others, in communities dominated by a single view. However, we also find mixed-attitude communities in which sceptics and activists frequently interact. Messages between like-minded users typically carry positive sentiment, while messages between sceptics and activists carry negative sentiment. We identify a number of general patterns in user behaviours relating to engagement with alternative views. Users who express negative sentiment are themselves the target of negativity. Users in mixed-attitude communities are less likely to hold a strongly polarised view, but more likely to express negative sentiment towards other users with differing views. Overall, **social media discussions of climate change often occur within polarising “echo chambers”, but also within “open forums”, mixed-attitude**

**communities that reduce polarisation and stimulate debate.** Our results have implications for public engagement with this important global challenge.

**2.3.11** Lai, Brown, Bisbee, Bonneau, Tucker, & Nagler (2022). Estimating the ideology of political YouTube videos. *SSRN.*

ABSTRACT: We present a method for estimating the ideology of political YouTube videos. As online media increasingly influences how people engage with politics, so does the importance of quantifying the ideology of such media for research. The subfield of estimating ideology as a latent variable has often focused on traditional actors such as legislators, while more recent work has used social media data to estimate the ideology of ordinary users, political elites, and media sources. We build on this work by developing a method to estimate the ideologies of YouTube videos, an important subset of media, based on their accompanying text metadata. First, we take Reddit posts linking to YouTube videos and use correspondence analysis to place those videos in an ideological space. We then train a text-based model with those estimated ideologies as training labels, enabling us to estimate the ideologies of videos not posted on Reddit. These predicted ideologies are then validated against human labels. Finally, we demonstrate the utility of this method by applying it to the watch histories of survey respondents with self-identified ideologies to evaluate the prevalence of echo chambers on YouTube. Our approach gives video-level scores based only on supplied text metadata, is scalable, and can be easily adjusted to account for changes in the ideological climate. This method could also be generalized to estimate the ideology of other items referenced or posted on Reddit.

ADDITIONAL EXCERPT: The resulting network (see figure 1 below) shows **clustering based on channel ideology, consistent with the idea that most online users prefer ideologically congruent information** in what can be referred to as ideological echo chambers... [However, when looking at video-level analysis], **we find substantial areas of overlap in the ideological content consumed by Democrats and Republicans** (see figure 2 below)**. This finding underscores an important benefit of our method: namely that video or channel-level analyses of echo chambers risk overstating their prevalence. Even though Democrats and Republicans may be watching different videos, there remains substantial overlap in the ideological content of what they watch.** By estimating ideology as a latent measure, and by applying this to the video level, we can paint a more nuanced picture of the extent of ideological echo chambers on YouTube.

FIGURE 1:

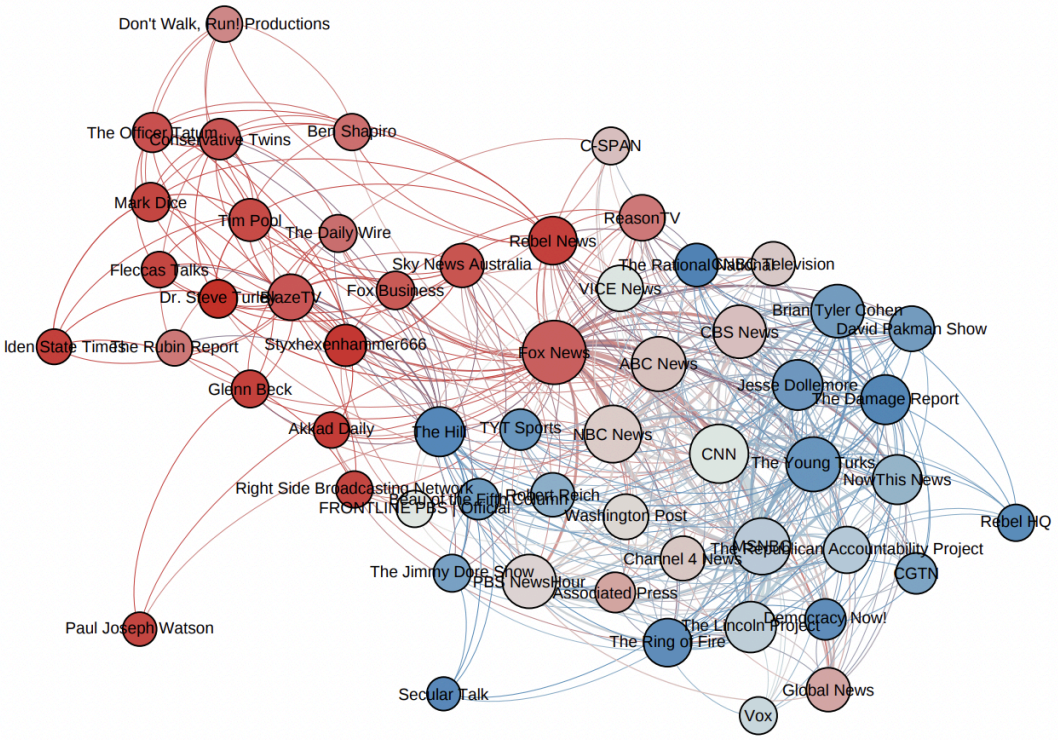



Figure 7: We draw a link between two channels if the same respondent has watched more than one video from both channels, and we plot the subset of channels which have at least 25 videos in our data set used in Section 3. The edge weight is the number of respondents who watched more than one video from both channels. Channels are colored according to the mean of their videos' ideology scores and nodes are sized based on degree.

FIGURE 2:

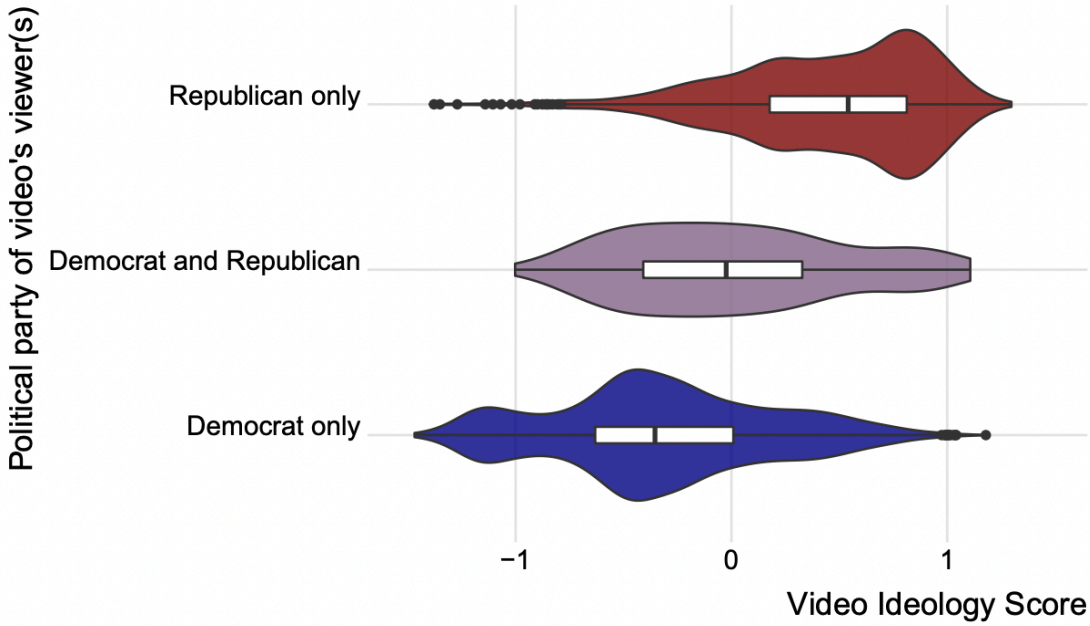



Figure 8: Distribution of video ideologies grouped by the self-reported party identification of the viewers. From top to bottom, we plot the ideology distribution of videos viewed only by Republicans, by both Democrats and Republicans, and by Democrats only.

**2.3.12** Flaxman, Goel, & Rao (2016). Filter bubbles, echo chambers, and online news consumption. *Public Opinion Quarterly*. (h/t Richard Fletcher)

ABSTRACT: Online publishing, social networks, and web search have dramatically lowered the costs of producing, distributing, and discovering news articles. Some scholars argue that such technological changes increase exposure to diverse perspectives, while others worry that they increase ideological segregation. We address the issue by examining web-browsing histories for 50,000 US-located users who regularly read online news. We find that social networks and search engines are associated with an increase in the mean ideological distance between individuals. However, somewhat counterintuitively, these same channels also are associated with an increase in an individual's exposure to material from his or her less preferred side of the political spectrum. Finally, **the vast majority of online news consumption is accounted for by individuals simply visiting the home pages of their favorite, typically mainstream, news outlets, tempering the consequences—both positive and negative—of recent technological changes. We thus uncover evidence for both sides of the debate, while also finding that the magnitude of the effects is relatively modest.**

[Other studies? What have we missed?]

## 2.4 DISCUSSION OF QUESTION #2

### 2.4.1 Note from Jon Haidt:

A few tentative conclusions seem to be warranted:

1. If we focus on exposure to *news stories*, the answer seems to be no. The major platforms such as Facebook and Twitter expose the AVERAGE user to a range of views and news stories, probably a wider range of views than the average person would encounter if not using any social media. (e.g., 2.1.1 Cinelli, Morales et al.; 2.2.5 Fletcher, Robertson, & Nielsen). So if the question is operationalized as "do the major social media platforms put most users into a news or information bubble?" the answer appears to be no.
2. If we focus on *social networks*, we see something very different: many studies find evidence of "homophily" (like goes with like; birds of a feather flock together." (e.g., 2.1.1 Cinelli, Morales et al.; 2.1.2 Barberá, P. (2015); 2.1.3 Hong & Kim (2016); 2.1.4 Mosleh, Martel et al. 2021. People do seem to immerse themselves into somewhat homogeneous partisan networks.
3. Simply encountering views from outside one's political community is not necessarily going to have the beneficial effects supposed by theorists of deliberative democracy, as explained by Zeynep Tufekci in this 2018 essay:

> The fourth lesson has to do with the much-touted issue of filter bubbles or echo chambers—the claim that online, we encounter only views similar to our own. This isn't completely true. While algorithms will often feed people some of what they already want to hear, **research shows that we probably encounter a wider variety of opinions online than we do offline,** or than we did before the advent of digital tools. Rather, **the problem is that when we encounter opposing views in the age and context of social media, it's not like reading them in a newspaper while sitting alone. It's like hearing them from the opposing team while sitting with our fellow fans in a football stadium.** Online, we're connected with our communities, and we seek approval from our

> like-minded peers. We bond with our team by yelling at the fans of the other one…. Belonging is stronger than facts.

See this similar idea, from Josh Pasek, in Tom Edsall's 6/15/22 column:

> "There is a tendency among many to assume that social media are putting us in echo chambers, where we only hear the stuff that confirms our worldviews. Instead, **it seems likely that social media are spurring polarization by facilitating the transmission of extreme information across the political spectrum**. The effect of this, likely, is to make ordinary people think that those on the other side are extreme. And this, in turn, fuels a deeper entrenchment in one's own group's attitudes and a willingness to support illiberal policies to avoid letting political opponents gain power."

These interpretations fit with **1.3.1** Bail, Argyle, Brown, Bumpus, Chen, Hunzaker, … Volfovsky (2018). Exposure to opposing views on social media can increase political polarization.

**2.4.2 Note from Chris Bail:**

> In my view, drawing definitive conclusions about the prevalence and power of social media echo chambers is inherently difficult because of the varied definitions of the term within the literature. If we adopt the broadest possible definition of an echo chamber— that is, that people tend to form social connections with those who are similar to them (i.e. what sociologists term "homophily")-- then I think it is somewhat easier to conclude that social media creates echo chambers. Note, however, that this new article provides substantial evidence that the echo chamber phenomenon is difficult to observe even within lifestyle issues that stretch far beyond politics.
>
> Most of the studies we review in this piece examine the relationship between political attitudes or behaviors and social media echo chambers. A general challenge that researchers face in this area is that publicly available data provide a very limited view of this phenomenon. The lion's share of research focuses on Twitter (and to a lesser extent Reddit). Many of these studies rely upon tweets to identify echo-chambers on such platforms. The challenge is that a) the vast majority of people tweet much less frequently than most people realize— perhaps as little as 1-2 times per month, according to the latest data from Pew. But also b) only a small fraction of those who tweet discuss politics. This means

that when we survey the sum total of political discourse on Twitter, we are actually only examining a fairly small subset of Twitter users (perhaps less than 5-6%). If a study *only* examines such publicly available data in a descriptive manner, then it is very difficult to make far-ranging conclusions about the broad population of people who do not engage in political discussions on places like Twitter.

There are a growing number of studies that either link publicly available data to private survey data or conduct field experiments in order to examine the echo chamber phenomenon. In my assessment of the literature, these studies generally suggest there is less evidence of the echo chamber phenomenon than many people realize— at least within the realm of politics.

Even if we could reliably establish whether echo chambers exist on social media, we would still face the broader— and perhaps more important— question of whether social media platforms encourage people to self-select into echo chambers. There are even fewer studies that are able to examine these dynamics— I am only aware of one that was conducted by Facebook many years ago. However, I do not think this single study is the “end all, be all” statement on the subject— not only because it is now quite dated, but also because it was conducted entirely by researchers inside Facebook. This is one of many reasons why I believe that advancing the research literature in this area will require greater access to data among researchers— particularly to sites other than Twitter or Reddit (which make up a relatively small percentage of the overall number of social media users).

* * * * * * * * * * * * * * *

# QUESTION 3: DOES SOCIAL MEDIA AMPLIFY POSTS THAT ARE MORE EMOTIONAL, INFLAMMATORY, OR FALSE?

Social media sites are routinely accused of increasing the prevalence of emotional, and inflammatory posts or false information— either because the algorithms on such platforms are thought to uprank posts that receive a lot of engagement, or because they do not effectively

moderate hateful content. We therefore consider studies that examine whether posts that are more emotional, inflammatory, and/or false tend to get more amplification than higher quality posts, and we also consider the related question of whether this amplification matters; if the amplification ends up not exposing many new people to the posts, then it doesn't really matter.

## 3.1 STUDIES INDICATING YES

**3.1.1** Brady, Wills, Jost, Tucker, & Van Bavel (2017). Emotion shapes the diffusion of moralized content in social networks. *Proceedings of the National Academy of Sciences*.

ABSTRACT: Political debate concerning moralized issues is increasingly common in online social networks. However, moral psychology has yet to incorporate the study of social networks to investigate processes by which some moral ideas spread more rapidly or broadly than others. Here, we show that the expression of moral emotion is key for the spread of moral and political ideas in online social networks, a process we call "moral contagion." Using a large sample of social media communications about three polarizing moral/political issues (n = 563,312), we observed **that the presence of moral-emotional words in messages increased their diffusion by a factor of 20% for each additional word. Furthermore, we found that moral contagion was bounded by group membership; moral-emotional language increased diffusion more strongly within liberal and conservative networks, and less between them.** Our results highlight the importance of emotion in the social transmission of moral ideas and also demonstrate the utility of social network methods for studying morality. These findings offer insights into how people are exposed to moral and political ideas through social networks, thus expanding models of social influence and group polarization as people become increasingly immersed in social media networks.

**3.1.2** Rathje, Van Bavel, & van der Linden (2021). Out-group animosity drives engagement on social media. *PNAS.*

ABSTRACT: There has been growing concern about the role social media plays in political polarization. We investigated whether out-group animosity was particularly successful at generating engagement on two of the largest social media platforms: Facebook and Twitter. **Analyzing posts from news media accounts and US**

**congressional members (*n* = 2,730,215), we found that posts about the political out-group were shared or retweeted about twice as often as posts about the in-group. Each individual term referring to the political out-group increased the odds of a social media post being shared by 67%. Out-group language consistently emerged as the strongest predictor of shares and retweets: the average effect size of out-group language was about 4.8 times as strong as that of negative affect language and about 6.7 times as strong as that of moral-emotional language**—both established predictors of social media engagement. Language about the out-group was a very strong predictor of "angry" reactions (the most popular reactions across all datasets), and language about the in-group was a strong predictor of "love" reactions, reflecting in-group favoritism and out-group derogation. This out-group effect was not moderated by political orientation or social media platform, but stronger effects were found among political leaders than among news media accounts. In sum, out-group language is the strongest predictor of social media engagement across all relevant predictors measured, suggesting that social media may be creating perverse incentives for content expressing out-group animosity.

**3.1.3** Alfano, Fard, Carter, Clutton, & Klein (2020). Technologically scaffolded atypical cognition: The case of YouTube's recommender system. *Synthese*.

ABSTRACT: YouTube has been implicated in the transformation of users into extremists and conspiracy theorists. The alleged mechanism for this radicalizing process is YouTube's recommender system, which is optimized to amplify and promote clips that users are likely to watch through to the end. YouTube optimizes for watch-through for economic reasons: people who watch a video through to the end are likely to then watch the *next* recommended video as well, which means that more advertisements can be served to them. This is a seemingly innocuous design choice, but it has a troubling side-effect. Critics of YouTube have alleged that the recommender system tends to recommend extremist content and conspiracy theories, as such videos are especially likely to capture and keep users' attention. To date, the problem of radicalization via the YouTube recommender system has been a matter of speculation. The current study represents the first systematic, pre-registered attempt to establish whether and to what extent the recommender system tends to promote such content. We begin by contextualizing our study in the framework of *technological seduction*. Next, we explain our methodology. **After that, we present our results, which are consistent with the radicalization hypothesis**. Finally, we discuss our findings, as well as directions for future research and recommendations for users, industry, and policy-makers.

EXCERPT: “YouTube’s recommender system is itself a moving target—indeed, **it has attracted attention precisely because a change to the algorithm appears to have shifted the balance towards promoting longer, more conspiratorial content. Our research supports that claim. Y**et YouTube is constantly tweaking its algorithm (making replication and reproducibility of work like ours tricky), and content-providers constantly tweak their output in order to maximize views within the system. Our research thus represents a snapshot of a complex, evolving system.

**3.1.4** Ribeiro, Ottoni, West, Almeida, & Meira, (2020). Auditing radicalization pathways on YouTube. *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*.

ABSTRACT: Non-profits, as well as the media, have hypothesized the existence of a radicalization pipeline on YouTube, claiming that users systematically progress towards more extreme content on the platform. Yet, there is to date no substantial quantitative evidence of this alleged pipeline. To close this gap, we conduct a large-scale audit of user radicalization on YouTube. We analyze 330,925 videos posted on 349 channels, which we broadly classified into four types: Media, the Alt-lite, the Intellectual Dark Web (I.D.W.), and the Alt-right. According to the aforementioned radicalization hypothesis, channels in the I.D.W. and the Alt-lite serve as gateways to fringe far-right ideology, here represented by Alt-right channels. Processing 72M+ comments, we show that the three channel types indeed increasingly share the same user base; that **users consistently migrate from milder to more extreme content; and that a large percentage of users who consume Alt-right content now consumed Alt-lite and I.D.W. content in the past.** We also probe YouTube's recommendation algorithm, looking at more than 2M video and channel recommendations between May/July 2019. **We find that Alt-lite content is easily reachable from I.D.W. channels, while Alt-right videos are reachable only through channel recommendations.** Overall, we paint a comprehensive picture of user radicalization on YouTube.

[Note from CB: The actual prevalence of the radicalization phenomenon documented in this article may be much lower than the abstract would lead one to conclude. See Munger and Phillips for more information].
]

**3.1.5** Vosoughi, Roy, & Aral (2018). The spread of true and false news online. *Science*.

ABSTRACT: We investigated the differential diffusion of all of the verified true and false news stories distributed on Twitter from 2006 to 2017. The data comprise ~126,000 stories tweeted by ~3 million people more than 4.5 million times. We classified news as true or false using information from six independent fact-checking organizations that exhibited 95 to 98% agreement on the classifications. **Falsehood diffused significantly farther, faster, deeper, and more broadly than the truth in all categories of information, and the effects were more pronounced for false political news than for false news about terrorism, natural disasters, science, urban legends, or financial information. We found that false news was more novel than true news, which suggests that people were more likely to share novel information. Whereas false stories inspired fear, disgust, and surprise in replies, true stories inspired anticipation, sadness, joy, and trust. Contrary to conventional wisdom, robots accelerated the spread of true and false news at the same rate, implying that false news spreads more than the truth because humans, not robots, are more likely to spread it.**

[NOTE: See important and relevant qualifications on these findings from the authors, responding to some over-interpretations commonly made]

**3.1.6** Kim, Guess, Nyhan & Reifler (2021). The distorting prism of social media: How self-selection and exposure to incivility fuel online comment toxicity. *Journal of Communication.*

ABSTRACT: Though prior studies have analyzed the textual characteristics of online comments about politics, less is known about how selection into commenting behavior and exposure to other people's comments changes the tone and content of political discourse. This article makes three contributions. First, we show that frequent commenters on Facebook are more likely to be interested in politics, to have more polarized opinions, and to use toxic language in comments in an elicitation task. Second, we find that **people who comment on articles in the real world use more toxic language on average than the public as a whole; levels of toxicity in comments scraped from media outlet Facebook pages greatly exceed what is observed in comments we elicit on the same articles from a nationally representative sample.** Finally, we demonstrate **experimentally that exposure to toxic language in comments increases the toxicity of subsequent comments.**

**3.1.7** Stella, Ferrara, & Domenico (2018). Bots increase exposure to negative and inflammatory content in online social systems. *Proceedings of the National Academy of Sciences*.

ABSTRACT: Social media can deeply influence reality perception, affecting millions of people's voting behavior. Hence, maneuvering opinion dynamics by disseminating forged content over online ecosystems is an effective pathway for social hacking. We propose a framework for discovering such a potentially dangerous behavior promoted by automatic users, also called "bots," in online social networks. We provide evidence that social bots target mainly human influencers but generate semantic content depending on the polarized stance of their targets. During the 2017 Catalan referendum, used as a case study, **social bots generated and promoted violent content aimed at Independentists, ultimately exacerbating social conflict online.** Our results open challenges for detecting and controlling the influence of such content on society.

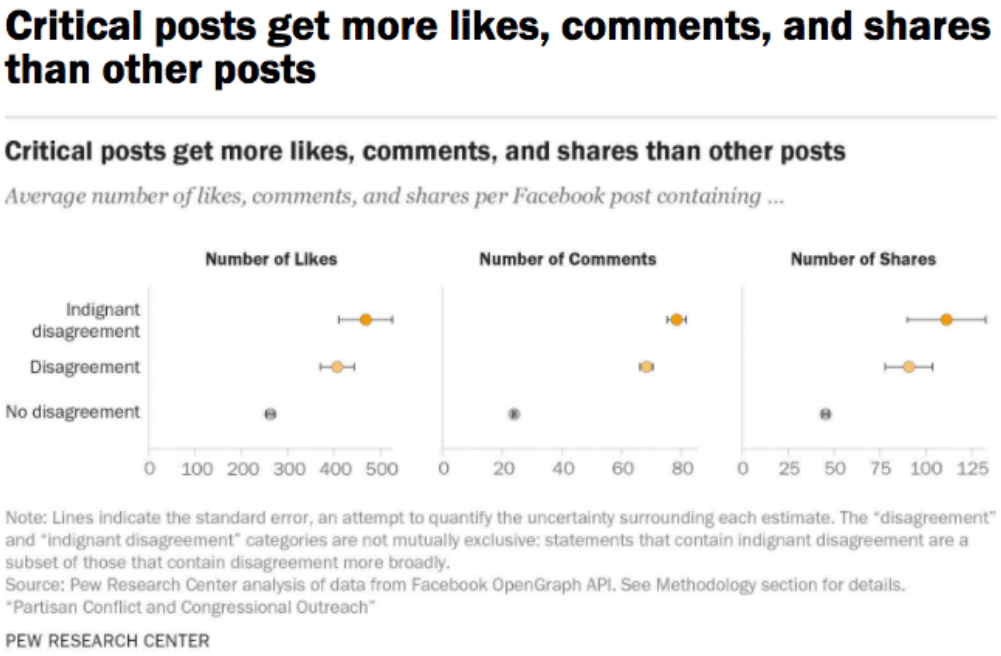
**3.1.8** Enders, Uscinski, Seelig, Klofstad, Wuchty, Funchion, Murthi, Premaratne, & Stoler (2021). The relationship between social media use and beliefs in conspiracy theories and misinformation. *Political Behavior*.

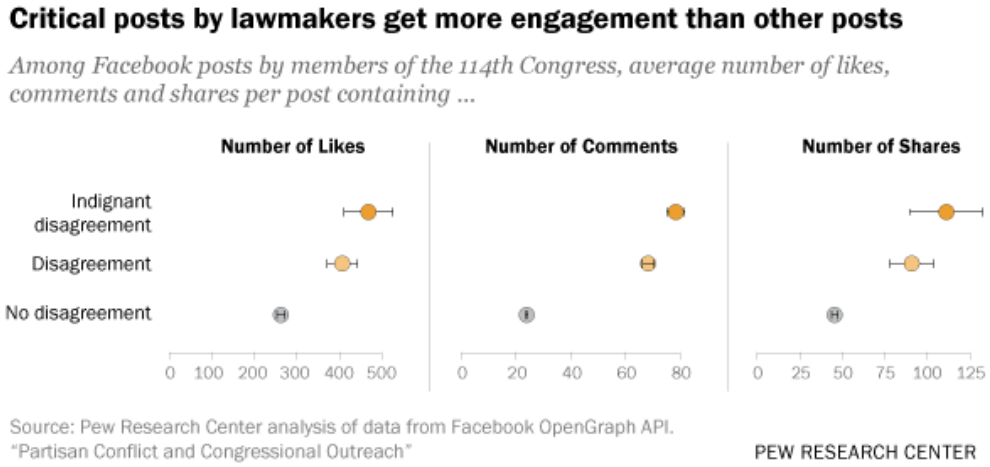
ABSTRACT: Numerous studies find associations between social media use and beliefs in conspiracy theories and misinformation. While such findings are often interpreted as evidence that social media causally promotes conspiracy beliefs, we theorize that this relationship is conditional on other individual-level predispositions. Across two studies, we examine the relationship between beliefs in conspiracy theories and media use, finding that **individuals who get their news from social media and use social media frequently express more beliefs in some types of conspiracy theories and misinformation. However, we also find that these relationships are conditional on conspiracy thinking—the predisposition to interpret salient events as products of conspiracies—such that social media use becomes more strongly associated with conspiracy beliefs as conspiracy thinking intensifies.** This pattern, which we observe across many beliefs from two studies, clarifies the relationship between social media use and beliefs in dubious ideas.

**3.1.9** Pröllochs, Bär, & Feuerriegel (2021). Emotions explain differences in the diffusion of true vs. False social media rumors. *Scientific Reports*.

ABSTRACT: False rumors (often termed “fake news”) on social media pose a significant threat to modern societies. However, potential reasons for the widespread diffusion of false rumors have been underexplored. In this work, we analyze whether sentiment words, as well as different emotional words, in social media content explain differences in the spread of true vs. false rumors. For this purpose, we collected $N$=126,301 rumor cascades from Twitter, comprising more than 4.5 million retweets that have been fact-checked for veracity. We then categorized the language in social media content to (1) sentiment (i.e., positive vs. negative) and (2) eight basic emotions (i. e., anger, anticipation, disgust, fear, joy, trust, sadness, and surprise). **We find that sentiment and basic emotions explain differences in the structural properties of true vs. false rumor cascades. False rumors (as compared to true rumors) are more likely to go viral if they convey a higher proportion of terms associated with a positive sentiment.** Further**, false rumors are viral when embedding emotional words classified as trust, anticipation, or anger.** All else being equal, **false rumors conveying one standard deviation more positive sentiment have a 37.58% longer lifetime and reach 61.44% more users.** Our findings offer insights into how true vs. false rumors spread and highlight the importance of managing emotions in social media content.
[Note from Haidt: This seems to be a rare finding that “good is stronger than bad,” although it still shows that emotions amplify false rumors.]

**3.1.10** Pew Research Center (2017). Critical posts get more likes, comments, and shares than other posts.




**3.1.11** de León & Trilling (2021). A sadness bias in political news sharing? The role of discrete emotions in the engagement and dissemination of political news on Facebook. *Social Media + Society.*

ABSTRACT. In this study, we address the role of emotions in political news sharing on Facebook to better understand the complex relationship between journalism, emotions, and politics. Categorizing Facebook Reactions (particularly, the Sad, Angry, Love, and

Wow Reactions) according to the discrete emotions model, we evaluate how positive versus negative political content relates to emotional responses, and how this consequentially influences the degree to which articles are shared across social media in the context of an election. We focus on the landmark 2018 Mexican elections to enable a nuanced conversation on how cues of user emotion predict the far-reaching dissemination of news articles on Facebook during a moment of heightened political attention. **Our findings demonstrate a negativity bias in news sharing and engagement, showing an outsized prevalence of anger in response to political news. In addition, we provide evidence of a novel sadness bias in the sharing of political coverage, suggesting that emotions considered as deactivating should be reevaluated in the context of social media.**

**3.1.12** Ciampaglia, Nematzadeh, Menczer, & Flammini, (2018). How algorithmic popularity bias hinders or promotes quality. *Scientific Reports*. [h/t Fil Menczer]

Algorithms that favor popular items are used to help us select among many choices, from top-ranked search engine results to highly-cited scientific papers. The goal of these algorithms is to identify high-quality items such as reliable news, credible information sources, and important discoveries–in short, high-quality content should rank at the top. Prior work has shown that choosing what is popular may amplify random fluctuations and lead to sub-optimal rankings. Nonetheless, it is often assumed that recommending what is popular will help high-quality content "bubble up" in practice. Here we identify the conditions in which popularity may be a viable proxy for quality content by studying a simple model of a cultural market endowed with an intrinsic notion of quality. A parameter representing the cognitive cost of exploration controls the trade-off between quality and popularity. **Below and above a critical exploration cost, popularity bias is more likely to hinder quality. But we find a narrow intermediate regime of user attention where an optimal balance exists: choosing what is popular can help promote high-quality items to the top.** These findings clarify the effects of algorithmic popularity bias on quality outcomes, and may inform the design of more principled mechanisms for techno-social cultural markets.

**3.1.13** Papakyriakopoulos, & Goodman (2022). The impact of Twitter labels on misinformation spread and user engagement: Lessons from Trump's election tweets. *Forthcoming in ACM WWW '22.* [h/t Orestis Papakyriakopoulos]

ABSTRACT: This study investigates the warning labels that Twitter placed on Donald Trump's false tweets about the 2020 US Presidential election. It specifically studies their relation to misinformation spread, and the magnitude and nature of user engagement. We categorize the warning labels by type – "veracity labels" calling out falsity and "contextual labels" providing more information. In addition, we categorize labels by their rebuttal strength and textual overlap (linguistic, topical) with the underlying tweet. We look at user interactions (liking, retweeting, quote tweeting, and replying), the content of user replies, and the type of user involved (partisanship and Twitter activity level) according to various standard metrics. **Using appropriate statistical tools, we find that, overall, label placement did not change the propensity of users to share and engage with labeled content, but the falsity of content did.**

**3.1.14** Corbu, Bârgăoanu, Buturoiu, & Ștefăniță (2020). Does fake news lead to more engaging effects on social media? Evidence from Romania. *Communications*. (h/t Olivia Fischer)

ABSTRACT: This study examines the potential of fake news to produce effects on social media engagement as well as the moderating role of education and government approval. We report on a 2x2x2 online experiment conducted in Romania (N=813), in which we manipulated the level of facticity of a news story, its valence, and intention to deceive. R**esults show that ideologically driven news with a negative valence (rather than fabricated news or other genres, such as satire and parody) have a greater virality potential.** However, neither the level of education nor government approval moderate this effect. Additionally, **both positive and negative ideologically driven news stories enhance the probability that people will sign a document to support the government** (i.e., potential for political engagement on social media). These latter effects are moderated by government approval: Lower levels of government approval lead to less support for the government on social media, as a consequence of fake news exposure.

**3.1.15** Frimer, Aujla, Feinberg, Skitka, Aquino, Eichstaedt, & Willer (2022). Incivility is rising among American politicians on Twitter. *Social Psychological and Personality Science*.

ABSTRACT: We provide the first systematic investigation of trends in the incivility of American politicians on Twitter, a dominant platform for political communication in the United States. Applying a validated artificial intelligence classifier to all 1.3 million tweets made by members of Congress since 2009, **we observe a 23% increase in incivility over a decade on Twitter.** Further analyses suggest that the **rise was partly driven by**

**reinforcement learning in which politicians engaged in greater incivility following positive feedback.** Uncivil tweets tended to receive more approval and attention, publicly indexed by large quantities of "likes" and "retweets" on the platform. **Mediational and longitudinal analyses show that the greater this feedback for uncivil tweets, the more uncivil tweets were thereafter.** We conclude by discussing how the structure of social media platforms might facilitate this incivility-reinforcing dynamic between politicians and their followers.

FIGURE 1:

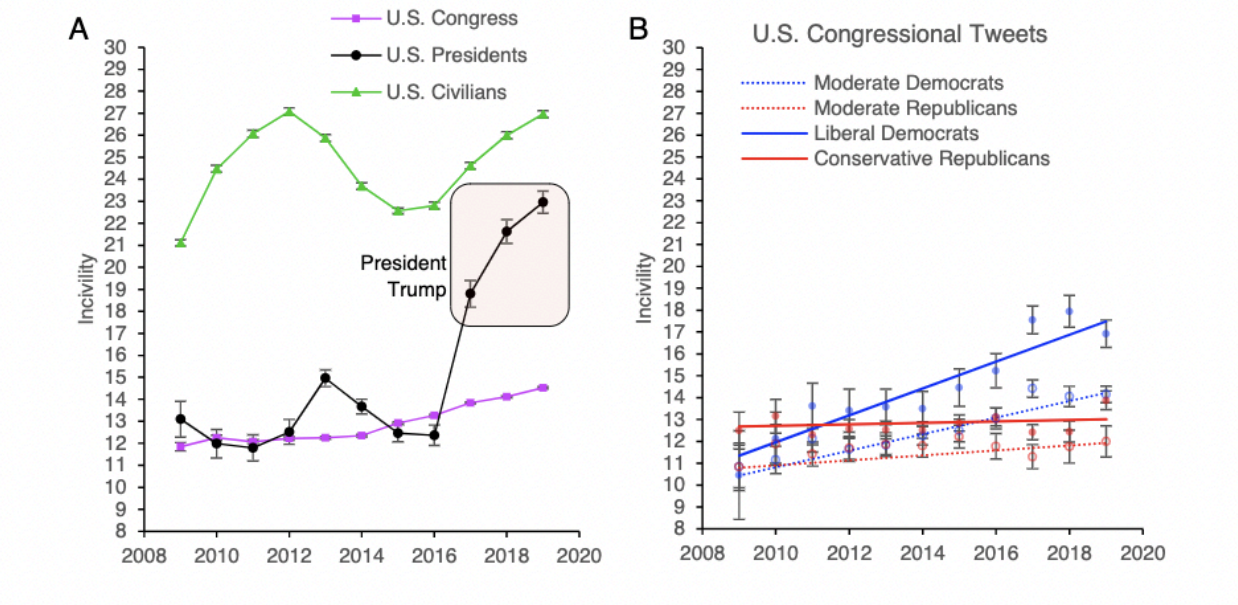



**Figure 1.** Changes in Levels of Incivility Among American Politicians Over Time on Twitter.
*Note.* Political incivility increased over time on Twitter both in the presidency and among members of the U.S. Congress (**A**), with the rise being the most pronounced among liberal Democrats (**B**). This partisan/ideological difference was mostly explained by reactions to President Trump (see the Supplement). Mean level estimates are from separate multilevel models for each year, with random intercepts for each Twitter user. Partisan levels (liberal, moderate, conservative) were derived from rollcall voting behavior. Trendlines were inferred from a party × partisanship × time multilevel model. Error bars represent 95% CIs.

FIGURE 2:

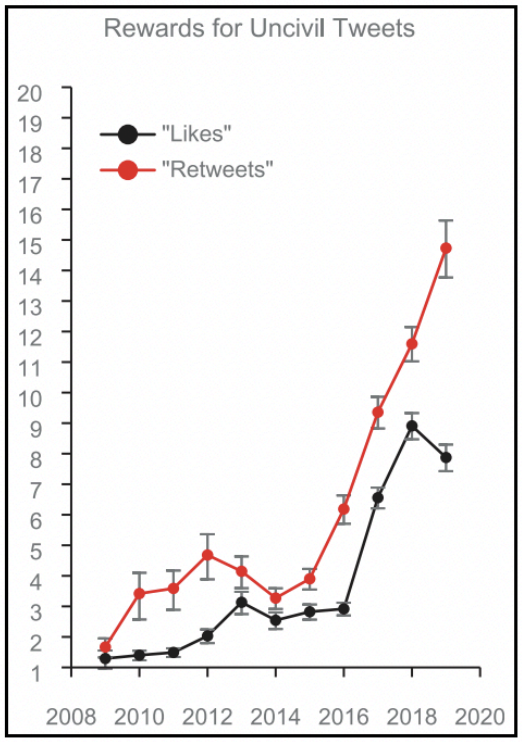



**Figure 4.** Rewards for Uncivil Tweeting Grew Stronger Over Time.
*Note.* The vertical axis representing the (model-implied) number of times more "retweets" and "likes" uncivil tweets (incivility = 100) received compared with civil tweets (incivility = 0). Error bars are 95% CIs.

**3.1.16** Wang, & Inbar (2022). Re-examining the spread of moralized rhetoric from political elites: Effects of valence and ideology. *Journal of Experimental Psychology.*

ABSTRACT: We examine the robustness of previous research finding increased diffusion of Twitter messages ("tweets") containing moral rhetoric. We use a distributed language model to examine the moral language used by U.S. political elites in two corpora of tweets: one from 2016 presidential candidates Hillary Clinton and Donald Trump, and one from U.S. Members of Congress. Consistent with previous research, we find greater diffusion for tweets containing moral rhetoric, but this is qualified by moral language valence and elite ideology. For both presidential candidates and Members of Congress, **negative moral language is associated with increased message diffusion. Positive moral language is not associated with diffusion for presidential candidates and is negatively associated with diffusion for Members of Congress.** In both data sets, the relationship between **negative moral language and message diffusion is stronger for liberals than conservatives.**

**3.1.17** Morris (2021). In Poland's politics, a 'social civil war' brewed as Facebook rewarded online anger. *Washington Post*.

EXCERPT: **An independent data analysis of major political parties in Poland that was conducted for The Post showed that after 2018, negative messages were more likely to receive a high number of shares. Previously, it appeared that more of a mix of positive and negative posts did well.**

Some Facebook employees recognized the need to act, according to the documents, but it was not just out of concern over the potentially damaging impact on society, internal documents show. Some employees also felt revisions to its algorithms were best for long-term growth, likening such outrage-centric content to junk food.

It was the presidential election in 2015 that woke Polish politics to the powers of Facebook, said Pawel Rybicki, who worked on the campaign for President Andrzej Duda. "We used social media full-scale," Rybicki said. Duda, an ally of Law and Justice, had been considered the underdog but won with 51.5 percent of the vote.

"It was like a war, and social media was the new gun for Polish political parties," recalled Rybicki, who met with the Facebook team when it was in Warsaw but said he largely raised concerns regarding moderation.

A consultant to the social media team for the Civic Platform party, who spoke on the condition of anonymity to discuss that party's social media strategy, described those days as the "wild West," with apparently little content-moderation on Facebook. He said that he, like others, noticed a shift in 2018 with more-extreme content breaking through.

**According to Facebook's internal report, the social media management team of one Polish political party (which was not named) described its shift: moving from a roughly even split of negative and positive messages, to 80 percent negative. The Civic Platform's social media team did not say whether the Facebook report mirrored feedback from someone at the party. Law and Justice officials declined interview requests.**

[Other studies? What have we missed?]

## 3.2 STUDIES INDICATING NO

**3.2.1** Grinberg, Joseph, Friedland, Swire-Thompson, & Lazer (2019). Fake news on Twitter during the 2016 U.S. presidential election. *Science*.

ABSTRACT: The spread of fake news on social media became a public concern in the United States after the 2016 presidential election. We examined exposure to and sharing of fake news by registered voters on Twitter and found that engagement with fake news sources was extremely concentrated. **Only 1% of individuals accounted for 80% of fake news source exposures, and 0.1% accounted for nearly 80% of fake news sources shared. Individuals most likely to engage with fake news sources were conservative leaning, older, and highly engaged with political news.** A cluster of fake news sources shared overlapping audiences on the extreme right, but for people across the political spectrum, most political news exposure still came from mainstream media outlets.

EXCERPT: This study estimated the extent to which people on Twitter shared and were exposed to content from fake news sources during the 2016 election season. **Although 6% of people who shared URLs with political content shared content from fake news sources, the vast majority of fake news shares and exposures were attributable to tiny fractions of the population.**

**3.2.2** Mosleh, Pennycook, & Rand (2020). Self-reported willingness to share political news articles in online surveys correlates with actual sharing on Twitter. *PLOS ONE*.

ABSTRACT: There is an increasing imperative for psychologists and other behavioral scientists to understand how people behave on social media. However, it is often very difficult to execute experimental research on actual social media platforms, or to link survey responses to online behavior in order to perform correlational analyses. Thus, there is a natural desire to use self-reported behavioral intentions in standard survey studies to gain insight into online behavior. But are such hypothetical responses hopelessly disconnected from actual sharing decisions? Or are online survey samples via sources such as Amazon Mechanical Turk (MTurk) so different from the average social media user that the survey responses of one group give little insight into the

on-platform behavior of the other? Here we investigate these issues by examining 67 pieces of political news content. We evaluate whether there is a meaningful relationship between (i) the level of sharing (tweets and retweets) of a given piece of content on Twitter, and (ii) the extent to which individuals (total N = 993) in online surveys on MTurk reported being willing to share that same piece of content. We found that the same news headlines that were more likely to be hypothetically shared on MTurk were also shared more frequently by Twitter users, r = .44. For example, across the observed range of MTurk sharing fractions, a 20 percentage point increase in the fraction of MTurk participants who reported being willing to share a news headline on social media was associated with 10x as many actual shares on Twitter. We also found that the correlation between sharing and various features of the headline was similar using both MTurk and Twitter data. These findings suggest that self-reported sharing intentions collected in online surveys are likely to provide some meaningful insight into what content would actually be shared on social media.

ADDITIONAL EXCERPT: In addition to examining headline veracity, we used the Linguistic Inquiry and Word Count (LIWC) dictionaries to determine the presence of Emotional, Moral, or Moral-Emotional language, as well as language related to Religion, Inhibition, and Insight; and we used the Dale-Chall formula to determine the complexity of the language used. When correlating these characteristics with sharing, we found identical patterns across the Mturk and Twitter sharing data. **False headlines were shared less on both MTurk ($r(66) = -0.536$, $p<0.001$) and Twitter ($r(66) = -.343$, $p = 0.004$) compared to true headlines; headlines containing Moral words were shared less on both on MTurk ($r(66) = -0.275$, $p = 0.023$) and Twitter ($r(66) = -0.317$, $p = 0.009$); and there was no significant correlation between any of the other headline characteristics and sharing on either MTurk or Twitter ($p>0.1$ for all).**

TABLE:

| | MTurk (z-scored) | Twitter (log-transformed) |
|---|---|---|
| Veracity | -0.536 ($p < .001$) | -0.343 ($p = 0.004$) |
| Emotional language | -0.118 ($p = 0.337$) | -0.076 ($p = 0.540$) |
| Moral language | -0.275 ($p = 0.023$) | -0.317 ($p = 0.009$) |
| Moral-Emotional language | 0.120 ($p = 0.331$) | 0.129 ($p = 0.295$) |
| Religious language | -0.073 ($p = 0.554$) | -0.025 ($p = 0.843$) |
| Insight language | -0.173 ($p = 0.157$) | -0.182 ($p = 0.137$) |
| Inhibition language | -0.051 ($p = 0.678$) | 0.098 ($p = 0.426$) |
| Language complexity | 0.160 ($p = 0.193$) | 0.026 ($p = 0.831$) |

https://doi.org/10.1371/journal.pone.0228882.t001

*Table 1. Correlation between various headline characteristics and the likelihood of being shared on MTurk and Twitter.*

### 3.2.3 Guess, Nagler, & Tucker (2019). Less than you think: Prevalence and predictors of fake news dissemination on Facebook. *Science Advances*.

ABSTRACT: So-called “fake news” has renewed concerns about the prevalence and effects of misinformation in political campaigns. Given the potential for widespread dissemination of this material, we examine the individual-level characteristics associated with sharing false articles during the 2016 U.S. presidential campaign. To do so, we uniquely link an original survey with respondents’ sharing activity as recorded in Facebook profile data. **First and foremost, we find that sharing this content was a relatively rare activity. Conservatives were more likely to share articles from fake news domains, which in 2016 were largely pro-Trump in orientation, than liberals or moderates. We also find a strong age effect, which persists after controlling for partisanship and ideology: On average, users over 65 shared nearly seven times as many articles from fake news domains as the youngest age group.**

EXCERPT: Holding constant ideology, party identification, or both, respondents in each age category were more likely to share fake news than respondents in the

next-youngest group, and the gap in the rate of fake news sharing between those in our oldest category (over 65) and youngest category is large and notable.

**3.2.4** Guess, Aslett, Tucker, Bonneau, & Nagler (2021). Cracking open the news feed: Exploring what U.S. Facebook users see and share with large-scale platform data. *Journal of Quantitative Description: Digital Media.*

ABSTRACT: In this study, we analyze for the first time newly available engagement data covering millions of web links shared on Facebook to describe how and by which categories of U.S. users different types of news are seen and shared on the platform. We focus on articles from low-credibility news publishers, credible news sources, purveyors of clickbait, and news specifically about politics, which we identify through a combination of curated lists and supervised classifiers. Our results support recent findings that more **fake news is shared by older users and conservatives and that both viewing and sharing patterns suggest a preference for ideologically congenial misinformation.** We also find that fake news articles related to politics are more popular among older Americans than other types, while the youngest users share relatively more articles with clickbait headlines. Across the platform, however, **articles from credible news sources are shared over 5 times more often and viewed over 7 times more often than articles from low-credibility sources.** These findings offer important context for researchers studying the spread and consumption of information — including misinformation — on social media.

**3.2.5** Hosseinmardi, Ghasemian, Clauset, Mobius, Rothschild, & Watts (2021). Examining the consumption of radical content on YouTube. *PNAS.*

ABSTRACT: Although it is under-studied relative to other social media platforms, YouTube is arguably the largest and most engaging online media consumption platform in the world. Recently, YouTube's scale has fueled concerns that YouTube users are being radicalized via a combination of biased recommendations and ostensibly apolitical "anti-woke" channels, both of which have been claimed to direct attention to radical political content. Here we test this hypothesis using a representative panel of more than 300,000 Americans and their individual-level browsing behavior, on and off YouTube, from January 2016 through December 2019. Using a labeled set of political news channels, we find that news consumption on YouTube is dominated by mainstream and largely centrist sources. Consumers of far-right content, while more engaged than average, represent a small and stable percentage of news consumers. However,

consumption of “anti-woke” content, defined in terms of its opposition to progressive intellectual and political agendas, grew steadily in popularity and is correlated with consumption of far-right content off-platform. **We find no evidence that engagement with far-right content is caused by YouTube recommendations systematically, nor do we find clear evidence that anti-woke channels serve as a gateway to the far right. Rather, consumption of political content on YouTube appears to reflect individual preferences that extend across the web as a whole.**

**3.2.6** Burton, Cruz, & Hahn (2021). Reconsidering evidence of moral contagion in online social networks. *Nature Human Behaviour*.

ABSTRACT: The ubiquity of social media use and the digital data traces it produces has triggered a potential methodological shift in the psychological sciences away from traditional, laboratory-based experimentation. The hope is that, by using computational social science methods to analyse large-scale observational data from social media, human behaviour can be studied with greater statistical power and ecological validity. However, current standards of null hypothesis significance testing and correlational statistics seem ill-suited to markedly noisy, high-dimensional social media datasets. We explore this point by probing the moral contagion phenomenon, whereby the use of moral-emotional language increases the probability of message spread. Through out-of-sample prediction, model comparisons and specification curve analyses, **we find that the moral contagion model performs no better than an implausible XYZ contagion model. This highlights the risks of using purely correlational evidence from large observational datasets and sounds a cautionary note for psychology’s merge with big data.**

**3.2.7** Valenzuela, Muñiz, & Santos (2022). Social media and belief in misinformation in Mexico: A case of maximal panic, minimal effects? *The International Journal of Press/Politics*. [h/t Sacha Yesilaltay]

ABSTRACT: Contrary to popular narratives, it is not clear whether using social media for news increases belief in political misinformation. Several of the most methodologically sound studies find small to nonexistent effects. However, extant research is limited by focusing on few platforms (usually Facebook, Twitter or YouTube) and is heavily U.S. centered. This leaves open the possibility that other platforms, such as those that rely on visual communication (e.g., Instagram) or are tailored to strong-tie network communication (e.g., WhatsApp), are more influential. Furthermore, the few studies conducted in other countries suggest that social media use increases political

misperceptions. Still, these works use cross-sectional designs, which are ill suited to dealing with omitted variable bias and temporal ordering of processes. Using a two-wave survey fielded in Mexico during the 2021 midterm elections (N = 596), we estimate the relationship between frequency of news exposure on Facebook, Twitter, YouTube, Instagram and WhatsApp, and belief in political misinformation, while controlling for both time-invariant and time-dependent individual differences. In contrast to political discussion, information literacy and digital skills, **none of the social platforms analyzed exhibits a significant association with misinformed beliefs. We also tested for possible indirect, moderated, and reciprocal relationships, but none of these analyses yielded a statistically significant result.** We conclude that the study is consistent with the "minimal media effects" paradigm, which suggests that efforts to address misinformation need to go beyond social platforms.

**3.2.8** Uscinski… & Murthi (2022). Have beliefs in conspiracy theories increased over time? *PLOS ONE*.

ABSTRACT: The public is convinced that beliefs in conspiracy theories are increasing, and many scholars, journalists, and policymakers agree. Given the associations between conspiracy theories and many non-normative tendencies, lawmakers have called for policies to address these increases. However, little evidence has been provided to demonstrate that beliefs in conspiracy theories have, in fact, increased over time. We address this evidentiary gap. Study 1 investigates change in the proportion of Americans believing 46 conspiracy theories; our observations in some instances span half a century. Study 2 examines change in the proportion of individuals across six European countries believing six conspiracy theories. Study 3 traces beliefs about which groups are conspiring against "us," while Study 4 tracks generalized conspiracy thinking in the U.S. from 2012 to 2021. **In no instance do we observe systematic evidence for an increase in conspiracism, however operationalized. We discuss the theoretical and policy implications of our findings.**

**3.2.9** Altay, Berriche, & Acerbi (2023). Misinformation on Misinformation: Conceptual and Methodological Challenges. *Social Media + Society*.

ABSTRACT: Alarmist narratives about online misinformation continue to gain traction despite evidence that its prevalence and impact are overstated. Drawing on research examining the use of big data in social science and reception studies, we identify six misconceptions about misinformation and highlight the conceptual and methodological

challenges they raise. The first set of misconceptions concerns the prevalence and circulation of misinformation. **First, scientists focus on social media because it is methodologically convenient, but misinformation is not just a social media problem. Second, the internet is not rife with misinformation or news, but with memes and entertaining content. Third, falsehoods do not spread faster than the truth; how we define (mis)information influences our results and their practical implications. The second set of misconceptions concerns the impact and the reception of misinformation. Fourth, people do not believe everything they see on the internet: the sheer volume of engagement should not be conflated with belief. Fifth, people are more likely to be uninformed than misinformed; surveys overestimate misperceptions and say little about the causal influence of misinformation. Sixth, the influence of misinformation on people's behavior is overblown as misinformation often "preaches to the choir."** To appropriately understand and fight misinformation, future research needs to address these challenges.

[Other studies? What have we missed?]

## 3.3 MIXED RESULTS OR UNCLASSIFIED

**3.3.1** Guess, Nyhan, & Reifler (2020). Exposure to untrustworthy websites in the 2016 US election. *Nature Human Behaviour*.

ABSTRACT: Although commentators frequently warn about echo chambers, little is known about the volume or slant of political misinformation that people consume online, the effects of social media and fact checking on exposure, or the effects of political misinformation on behaviour. Here, we evaluate these questions for websites that publish factually dubious content, which is often described as fake news. **Survey and web-traffic data from the 2016 US presidential campaign show that supporters of Donald Trump were most likely to visit these websites, which often spread through Facebook.** However, **these websites made up a small share of people's information diets on average and were largely consumed by a subset of Americans with strong preferences for pro-attitudinal information.** These results suggest that the widespread speculation about the prevalence of exposure to untrustworthy websites has been overstated.

ADDITIONAL EXCERPT: We made a more direct inference about the role of Facebook by examining the URLs visited by a respondent immediately before visiting an

untrustworthy website. **Facebook was among the three previous websites visited by respondents in the previous 30 s for 15.1% of the articles from untrustworthy news websites that we observed in our web data (see figure below). By contrast, Facebook appears in the comparable set of previous URLs for only 5.9% of articles on websites that were classified as hard news (excluding Amazon, Twitter and YouTube).**

We did not observe this pattern of differential visits immediately before visits to untrustworthy websites for Google (3.3% untrustworthy news versus 6.2% hard news) or Twitter (1.0% untrustworthy versus 1.5% hard news). It also exceeds what we observe for webmail providers such as Gmail (9.5% untrustworthy versus 5.4% hard news). **Our results demonstrate that Facebook was a key vector of distribution for untrustworthy websites.**

[CLARIFICATION from author, Andy Guess: We also show this exposure was already highly concentrated. **Taken together, this doesn't prove that FB was amplifying misinformation to people who weren't seeking it out**]

FIGURE 1:

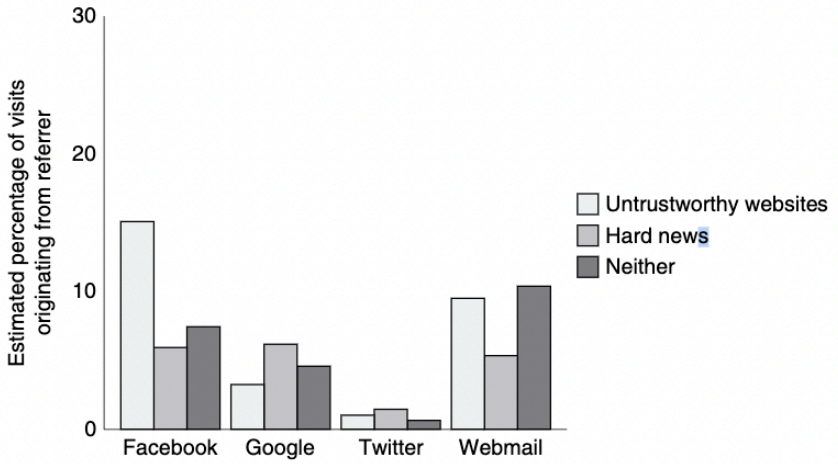



**Fig. 4 | Referrers to untrustworthy news websites and other sources.** Means and 95% CIs were calculated using survey weights among YouGov Pulse panel members; data from 7 October to 14 November 2016 ($n$ = 2,525). The denominator for information consumption includes total exposure to those websites as well as the number of pages visited on websites that were classified as focusing on hard-news topics (excluding Amazon, Twitter and YouTube). Respondents who did not visit any of these websites were excluded from the information-diet graph. Facebook, Google, Twitter or a webmail provider such as Gmail were identified as a referrer if they appeared within the last three URLs visited by the user during the 30 s before visiting the article.

**3.3.2** Brady, Crockett, & Van Bavel (2020). The MAD model of moral contagion: The role of motivation, attention, and design in the spread of moralized content online. *Perspectives on Psychological Science*.

ABSTRACT: With more than 3 billion users, online social networks represent an important venue for moral and political discourse and have been used to organize political revolutions, influence elections, and raise awareness of social issues. These examples rely on a common process to be effective: the ability to engage users and spread moralized content through online networks. Here, we review evidence that expressions of moral emotion play an important role in the spread of moralized content (a phenomenon we call *moral contagion*). Next, we propose a psychological model called the motivation, attention, and design (MAD) model to explain moral contagion. **The MAD model posits that people have group-identity-based *motivations* to share moral-emotional content, that such content is especially likely to capture our *attention*, and that the *design* of social-media platforms amplifies our natural**

**motivational and cognitive tendencies to spread such content.** We review each component of the model (as well as interactions between components) and raise several novel, testable hypotheses that can spark progress on the scientific investigation of civic engagement and activism, political polarization, propaganda and disinformation, and other moralized behaviors in the digital age.

**3.3.3** Vicario, Bessi, Zollo, Petroni, Scala, Caldarelli, Stanley, & Quattrociocchi (2016). The spreading of misinformation online. *Proceedings of the National Academy of Sciences*.

ABSTRACT: The wide availability of user-provided content in online social media facilitates the aggregation of people around common interests, worldviews, and narratives. However, the World Wide Web (WWW) also allows for the rapid dissemination of unsubstantiated rumors and conspiracy theories that often elicit rapid, large, but naive social responses such as the recent case of Jade Helm 15—where a simple military exercise turned out to be perceived as the beginning of a new civil war in the United States. In this work, we address the determinants governing misinformation spreading through a thorough quantitative analysis. In particular, we focus on how Facebook users consume information related to two distinct narratives: scientific and conspiracy news. **We find that, although consumers of scientific and conspiracy stories present similar consumption patterns with respect to content, cascade dynamics differ. Selective exposure to content is the primary driver of content diffusion and generates the formation of homogeneous clusters, i.e., "echo chambers." Indeed, homogeneity appears to be the primary driver for the diffusion of contents and each echo chamber has its own cascade dynamics.** Finally, we introduce a data-driven percolation model mimicking rumor spreading and we show that homogeneity and polarization are the main determinants for predicting cascades' size.

**3.3.4** Juul, & Ugander (2021). Comparing information diffusion mechanisms by matching on cascade size. *Proceedings of the National Academy of Sciences*.

ABSTRACT: Do some types of information spread faster, broader, or further than others? To understand how information diffusions differ, scholars compare structural properties of the paths taken by content as it spreads through a network, studying so-called cascades. Commonly studied cascade properties include the reach, depth, breadth, and speed of propagation. Drawing conclusions from statistical differences in

these properties can be challenging, as many properties are dependent. In this work, we demonstrate the essentiality of controlling for cascade sizes when studying structural differences between collections of cascades. We first revisit two datasets from notable recent studies of online diffusion that reported content-specific differences in cascade topology: an exhaustive corpus of Twitter cascades for verified true- or false-news content by Vosoughi et al. [S. Vosoughi, D. Roy, S. Aral. *Science* 359, 1146–1151 (2018)] and a comparison of Twitter cascades of videos, pictures, news, and petitions by Goel et al. [S. Goel, A. Anderson, J. Hofman, D. J. Watts. *Manage. Sci.* 62, 180–196 (2016)]. Using methods that control for joint cascade statistics, **we find that for false- and true-news cascades, the reported structural differences can almost entirely be explained by false-news cascades being larger. For videos, images, news, and petitions, structural differences persist when controlling for size.** Studying classical models of diffusion, we then give conditions under which differences in structural properties under different models do or do not reduce to differences in size. Our findings are consistent with the mechanisms underlying true- and false-news diffusion being quite similar, differing primarily in the basic infectiousness of their spreading process.

**3.3.5** Allcott, Gentzkow, & Yu, C. (2019). Trends in the diffusion of misinformation on social media. *Research & Politics*.

ABSTRACT; In recent years, there has been widespread concern that misinformation on social media is damaging societies and democratic institutions. In response, social media platforms have announced actions to limit the spread of false content. We measure trends in the diffusion of content from 569 fake news websites and 9540 fake news stories on Facebook and Twitter between January 2015 and July 2018. **User interactions with false content rose steadily on both Facebook and Twitter through the end of 2016. Since then, however, interactions with false content have fallen sharply on Facebook while continuing to rise on Twitter, with the ratio of Facebook engagements to Twitter shares decreasing by 60%.** In comparison, interactions with other news, business, or culture sites have followed similar trends on both platforms. Our results suggest that the relative magnitude of the misinformation problem on Facebook has declined since its peak.

**3.3.6** Garrett (2019). Social media's contribution to political misperceptions in U.S. Presidential elections. *PLOS ONE*.

ABSTRACT: There is considerable concern about the role that social media, such as Facebook and Twitter, play in promoting misperceptions during political campaigns. These technologies are widely used, and inaccurate information flowing across them has a high profile. This research uses three-wave panel surveys conducted with representative samples of Americans during both the 2012 and 2016 U.S. Presidential elections to assess whether use of social media for political information promoted endorsement of falsehoods about major party candidates or important campaign issues. Fixed effects regression helps ensure that observed effects are not due to individual differences. Results indicate that social media use had a small but significant influence on misperceptions about President Obama in the 2012 election, and that this effect was most pronounced among strong partisans. Social media had no effect on belief accuracy about the Republican candidate in that election. The 2016 survey focused on campaign issues. There is no evidence that social media use influenced belief accuracy about these topics in aggregate, but Facebook users were unique. Social media use by this group reduced issue misperceptions relative to those who only used other social media. These **results demonstrate that social media can alter citizens' willingness to endorse falsehoods during an election, but that the effects are often small.**

**3.3.7** Luca, Munger, Nagler, & Tucker (2021). You won't believe our results! But they might: Heterogeneity in beliefs about the accuracy of online media. *Journal of Experimental Political Science*.

ABSTRACT: "Clickbait" media has long been espoused as an unfortunate consequence of the rise of digital journalism. But little is known about *why* readers choose to read clickbait stories. Is it merely curiosity, or might voters think such stories are more likely to provide useful information? We conduct a survey experiment in Italy, where a major political party enthusiastically embraced the esthetics of new media and encouraged their supporters to distrust legacy outlets in favor of online news. We offer respondents a monetary incentive for correct answers to manipulate the relative salience of the motivation for accurate information. This incentive *increases* differences in the preference for clickbait; **older and less educated subjects become even more likely to opt to read a story with a clickbait headline when the incentive to produce a factually correct answer is higher. Our model suggests that a politically relevant subset of the population prefers Clickbait Media because they trust it more.**

**3.3.8** Rosenzweig, Bago, Berinsky, & Rand, (2021). Happiness and surprise are associated with worse truth discernment of COVID-19 headlines among social media users in Nigeria. *Harvard Kennedy School Misinformation Review*.

ESSAY SUMMARY:

- Using a survey of 1,341 Facebook users in Nigeria, we assess whether emotional reactions are associated with belief in COVID-19 headlines, information seeking, and sharing intentions. After viewing true and false COVID-19-related headlines, respondents reported what emotions, if any, they experienced. We assess how emotions correlate with our three outcomes of interest: i) *belief* about the accuracy of the headline, ii) interest in *clicking* to read, and iii) *sharing* intentions.
- **Respondents are more likely to believe, want to read and share headlines (regardless of veracity) when they feel any emotion. Emotional responses are associated with worse truth discernment and the ability to distinguish true from false headlines when assessing belief (but not reading or sharing). We find that happiness and surprise, in particular, are associated with believing and sharing false, relative to true, headlines.**
- Interventions to improve discernment of COVID-19 information should target youth, those who rely on intuition, and ruling party supporters in Nigeria.
- Understanding the role emotions play in reactions to misinformation has implications for technology platforms, governments, and citizens interested in combating the COVID-19 “infodemic.” Future research should test the causal relationship between emotions and belief in COVID-19 misinformation and interventions designed to regulate specific emotions in diverse settings.

**3.3.9** Bandy, & Diakopoulos (2021). Curating quality? How Twitter’s timeline algorithm treats different types of news. *Social Media + Society*.

ABSTRACT: This article explores how Twitter’s algorithmic timeline influences exposure to different types of external media. We use an agent-based testing method to compare chronological timelines and algorithmic timelines for a group of Twitter agents that emulated real-world archetypal users. We first find that algorithmic timelines exposed agents to external links at roughly half the rate of chronological timelines. Despite the reduced exposure, the proportional makeup of external links remained fairly stable in terms of source categories (major news brands, local news, new media, etc.). Notably, however, **algorithmic timelines slightly increased the proportion of “junk news”**

**websites in the external link exposures.** While our descriptive evidence does not fully exonerate Twitter's algorithm, it does **characterize the algorithm as playing a fairly minor, supporting role in shifting media exposure for end users, especially considering upstream factors that create the algorithm's input—factors such as human behavior, platform incentives, and content moderation.** We conclude by contextualizing the algorithm within a complex system consisting of many factors that deserve future research attention.

**3.3.10** Allcott, & Gentzkow (2017). Social Media and Fake News in the 2016 Election. *Journal of Economic Perspectives*.

ABSTRACT: Following the 2016 US presidential election, many have expressed concern about the effects of false stories ("fake news"), circulated largely through social media. We discuss the economics of fake news and present new data on its consumption prior to the election. Drawing on web browsing data, archives of fact-checking websites, and results from a new online survey, we find: 1) social media was an important but not dominant source of election news, with 14 percent of Americans calling social media their "most important" source; 2) **of the known false news stories that appeared in the three months before the election, those favoring Trump were shared a total of 30 million times on Facebook, while those favoring Clinton were shared 8 million times;** 3) the average American adult saw on the order of one or perhaps several fake news stories in the months around the election, with just over half of those who recalled seeing them believing them; and **4) people are much more likely to believe stories that favor their preferred candidate, especially if they have ideologically segregated social media networks.**

**3.3.11** Huszár, Ktena et al. (2021) Algorithmic amplification of politics on Twitter. *PNAS.*

ABSTRACT: Content on Twitter's home timeline is selected and ordered by personalization algorithms. By consistently ranking certain content higher, these algorithms may amplify some messages while reducing the visibility of others. There's been intense public and scholarly debate about the possibility that some political groups benefit more from algorithmic amplification than others. We provide quantitative evidence from a long-running, massive-scale randomized experiment on the Twitter platform that committed a randomized control group including nearly 2 million daily active accounts to a reverse-chronological content feed free of algorithmic personalization. We present two sets of findings. First, we studied tweets by elected

legislators from major political parties in seven countries. **Our results reveal a remarkably consistent trend: In six out of seven countries studied, the mainstream political right enjoys higher algorithmic amplification than the mainstream political left. Consistent with this overall trend, our second set of findings studying the US media landscape revealed that algorithmic amplification favors right-leaning news sources**. We further looked at **whether algorithms amplify far-left and far-right political groups more than moderate ones; contrary to prevailing public belief, we did not find evidence to support this hypothesis**. We hope our findings will contribute to an evidence-based debate on the role personalization algorithms play in shaping political content consumption.
[NOTE: this study does not exactly show amplification of content that is false; just of content that is right leaning; also interesting that it shows amplification of the right only, not of the left, at the expense of moderates]

**3.3.12** Osmundsen, Bor, Vahlstrup, Bechmann, & Petersen (2021). Partisan polarization is the primary psychological motivation behind political fake news sharing on Twitter. *American Political Science Review*.

ABSTRACT: The rise of “fake news” is a major concern in contemporary Western democracies. Yet, research on the psychological motivations behind the spread of political fake news on social media is surprisingly limited. Are citizens who share fake news *ignorant* and lazy? Are they fueled by sinister motives, seeking to *disrupt* the social status quo? Or do they seek to attack partisan opponents in an increasingly *polarized* political environment? This article is the first to test these competing hypotheses based on a careful mapping of psychological profiles of over 2,300 American Twitter users linked to behavioral sharing data and sentiment analyses of more than 500,000 news story headlines. The findings contradict the ignorance perspective but provide some support for the disruption perspective and strong support for the partisan polarization perspective. Thus, **individuals who report hating their political opponents are the most likely to share political fake news and selectively share content that is useful for derogating these opponents.** Overall, our findings show that **fake news sharing is fueled by the same psychological motivations that drive other forms of partisan behavior, including sharing partisan news from traditional and credible news sources.**

**3.3.13** Altay, Nielsen, & Fletcher (2022, Working Paper). The impact of news media and digital platform use on awareness of and belief in COVID-19 misinformation. *PsyArXiv.* [h/t Sacha Yesilaltay]

ABSTRACT: Does the news media exacerbate or reduce misinformation problems? Although some news media deliberately try to counter misinformation, it has been suggested that they might also inadvertently, and sometimes purposefully, amplify it. We conducted a two-wave panel survey in Brazil, India, and the UK (N = 4732) to investigate the effect of news and digital platform use, on awareness of and belief in COVID-19 misinformation over time. We found little support for the idea that the news exacerbates misinformation problems. News use broadened people's awareness of false claims, but did not increase the likelihood that people would believe them—and in some cases, news use actually weakened false belief acquisition, depending on access mode (online or offline) and outlet type. In line with previous research, we also find that news use strengthens political knowledge gain over time, again depending on outlets used. **The effect of digital platforms was inconsistent across countries, and in most cases not significant—though some, like Twitter, were associated with positive outcomes while a few others were associated with negative outcomes.** Overall, our findings challenge the notion that news media, by reporting on false and misleading claims, ultimately leave the public more misinformed, and support the idea that news helps people become more informed and, in some cases, more resilient to misinformation.

ADDITIONAL EXCERPT: 'In the UK, more frequent YouTube use broadened awareness of misinformation (b = .05 [.03, .08]). And more frequent Twitter and Facebook use weakened false belief acquisition (b = .17 [.26, .08]; b = -.10 [.18, -.01]) while more frequent FB Messenger and Pinterest use strengthened it (b = .14 [.04, .24]; b = .23 [.07, .39]). In India, more frequent Facebook use strengthened false belief acquisition (b = .22 [.01, .43]). In Brazil, more frequent Telegram use broadened awareness of false claims (b = .09 [.04, .13]), while higher FB Messenger and LinkedIn use strengthened false belief acquisition (b = .25 [.004, .49]; b = .29 [.02, .57]).'

**3.3.14** Shao, Ciampaglia, Varol, Yang, Flammini, & Menczer (2018). The spread of low-credibility content by social bots. *Nature Communications*. [h/t Fil Menczer]

ABSTRACT: The massive spread of digital misinformation has been identified as a major threat to democracies. Communication, cognitive, social, and computer scientists

are studying the complex causes for the viral diffusion of misinformation, while online platforms are beginning to deploy countermeasures. Little systematic, data-based evidence has been published to guide these efforts. Here we analyze 14 million messages spreading 400 thousand articles on Twitter during ten months in 2016 and 2017. **We find evidence that social bots played a disproportionate role in spreading articles from low-credibility sources. Bots amplify such content in the early spreading moments, before an article goes viral.** They also target users with many followers through replies and mentions. **Humans are vulnerable to this manipulation, resharing content posted by bots. Successful low-credibility sources are heavily supported by social bots.** These results suggest that curbing social bots may be an effective strategy for mitigating the spread of online misinformation.

**3.3.15** Majó-Vázquez, Congosto, Nicholls, & Nielsen (2021). The role of suspended accounts in political discussion on social media: Analysis of the 2017 French, UK and German Elections. *Social Media + Society.*

Content moderation on social media is at the center of public and academic debate. In this study, we advance our understanding on which type of election-related content gets suspended by social media platforms. For this, we assess the behavior and content shared by suspended accounts during the most important elections in Europe in 2017 (in France, the United Kingdom, and Germany). We identify significant differences when we compare the behavior and content shared by Twitter suspended accounts with all other active accounts, including a focus on amplifying divisive issues like immigration and religion and systematic activities increasing the visibility of specific political figures (often but not always on the right). Our analysis suggests that suspended accounts were overwhelmingly human operated and no more likely than other accounts to share "fake news." This study sheds light on the moderation policies of social media platforms, which have increasingly raised contentious debates, and equally importantly on the integrity and dynamics of political discussion on social media during major political events.

**3.3.16** Theocharis… & Štětka (2021). Does the platform matter? Social media and COVID-19 conspiracy theory beliefs in 17 countries. *New Media & Society*. [h/t Sacha Yesilaltay]

ABSTRACT: While the role of social media in the spread of conspiracy theories has received much attention, a key deficit in previous research is the lack of distinction

between different types of platforms. This study places the role of social media affordances in facilitating the spread of conspiracy beliefs at the center of its enquiry. We examine the relationship between platform use and conspiracy theory beliefs related to the COVID-19 pandemic. Relying on the concept of technological affordances, we theorize that variation across key features make some platforms more fertile places for conspiracy beliefs than others. Using data from a crossnational dataset based on a two-wave online survey conducted in 17 countries before and after the onset of the COVID-19 pandemic, **we show that Twitter has a negative effect on conspiracy beliefs—as opposed to all other platforms under examination which are found to have a positive effect.**

**3.3.17** Majó-Vázquez, Nielsen, Verdú, Rao, Domenico, Papaspiliopoulos (2020). Volume and patterns of toxicity in social media conversations during the Covid-19 pandemic. *Reuters Institute for the Study of Journalism.* [h/t Silvia Majo-Vazquez]

INTRODUCTION: In this RISJ Factsheet, we assess the volume and patterns of toxic conversations on social media during the Covid-19 pandemic. We specifically analyse worldwide conversations on Twitter targeting the World Health Organization (WHO), a central actor during the pandemic. We found that toxic messages amount to 21% of the overall conversation touching on the Covid-19 pandemic and the role of the WHO in the crisis. In other words, 21 out of 100 tweets in our sample are expected to convey a rude, disrespectful, or unreasonable comment.

The percentage of toxic tweets increases after 26 March (25%), when many countries were facing the growing adverse effects of the pandemic and passing measures to confine their populations. Peaks in toxicity can be divided into two different phases. At the beginning of the pandemic the highest percentage of toxic messages correlate with the WHO's statements or events, whereas at the end of the period studied, top-down messages from political leaders or specific media coverage coincide in time with the surge in toxicity. Our analysis contributes to the current research on the health of online debates amid the increasing role of social media as a critical entrance to information and mediator of public opinion building. Our analyses are based on a filtered dataset of about 303 million tweets including Covid-19 related terms, from which we obtained a final sub-subset of 222,774 tweets mentioning the WHO. The time window for this study spans 20 January to 23 April 2020. At that time, countries were at different stages of the pandemic. Some of them – mainly European, but also others such as China – were facing the most severe consequences of the peak of the outbreak, including strict lockdown measures, whereas others were just through the first stages of the crisis.

**3.3.18** Benkler, Tilton, Etling, Roberts, Clark, Faris, Kaiser, & Schmitt (2020). *Mail-In voter fraud: Anatomy of a disinformation campaign*. *Social Science Research Network.*

ABSTRACT: The claim that election fraud is a major concern with mail-in ballots has become the central threat to election participation during the COVID-19 pandemic and to the legitimacy of the outcome of the election across the political spectrum. President Trump has repeatedly cited his concerns over voter fraud associated with mail-in ballots as a reason that he may not abide by an adverse electoral outcome. Polling conducted in September 2020 suggests that nearly half of Republicans agree with the president that election fraud is a major concern associated with expanded mail-in voting during the pandemic. Few Democrats share that belief. Despite the consensus among independent academic and journalistic investigations that voter fraud is rare and extremely unlikely to determine a national election, tens of millions of Americans believe the opposite. This is a study of the disinformation campaign that led to widespread acceptance of this apparently false belief and to its partisan distribution pattern. Contrary to the focus of most contemporary work on disinformation, our findings suggest that this highly effective disinformation campaign, with potentially profound effects for both participation in and the legitimacy of the 2020 election, was an elite-driven, mass-media led process. Social media played only a secondary and supportive role.

Our results are based on analyzing over fifty-five thousand online media stories, five million tweets, and seventy-five thousand posts on public Facebook pages garnering millions of engagements. They are consistent with our findings about the American political media ecosystem from 2015-2018, published in Network Propaganda , in which we found that Fox News and Donald Trump's own campaign were far more influential in spreading false beliefs than Russian trolls or Facebook clickbait artists. This dynamic appears to be even more pronounced in this election cycle, likely because Donald Trump's position as president and his leadership of the Republican Party allow him to operate directly through political and media elites, rather than relying on online media as he did when he sought to advance his then-still-insurgent positions in 2015 and the first half of 2016.

Our findings here suggest that Donald Trump has perfected the art of harnessing mass media to disseminate and at times reinforce his disinformation campaign by using three core standard practices of professional journalism. These three are: elite institutional

focus (if the President says it, it's news); headline seeking (if it bleeds, it leads); and balance , neutrality, or the avoidance of the appearance of taking a side. He uses the first two in combination to summon coverage at will, and has used them continuously to set the agenda surrounding mail-in voting through a combination of tweets, press conferences, and television interviews on Fox News. He relies on the latter professional practice to keep audiences that are not politically pre-committed and have relatively low political knowledge confused, because it limits the degree to which professional journalists in mass media organizations are willing or able to directly call the voter fraud frame disinformation. The president is, however, not acting alone. Throughout the first six months of the disinformation campaign, the Republican National Committee (RNC) and staff from the Trump campaign appear repeatedly and consistently on message at the same moments, suggesting an institutionalized rather than individual disinformation campaign. The efforts of the president and the Republican Party are supported by the right-wing media ecosystem, primarily Fox News and talk radio functioning in effect as a party press. These reinforce the message, provide the president a platform, and marginalize or attack those Republican leaders or any conservative media personalities who insist that there is no evidence of widespread voter fraud associated with mail-in voting.

**The primary cure for the elite-driven, mass media communicated information disorder we observe here is unlikely to be more fact checking on Facebook. Instead, it is likely to require more aggressive policing by traditional professional media, the Associated Press, the television networks, and local TV news editors of whether and how they cover Trump's propaganda efforts, and how they educate their audiences about the disinformation campaign the president and the Republican Party have waged.**

**3.3.19** Bandy, & Diakopoulos (2021). More accounts, fewer links: How algorithmic curation impacts media exposure in Twitter timelines. *Proceedings of the ACM on Human-Computer Interaction*. (h/t Jack Bandy)

ABSTRACT: Algorithmic timeline curation is now an integral part of Twitter's platform, affecting information exposure for more than 150 million daily active users. Despite its large-scale and high-stakes impact, especially during a public health emergency such as the COVID-19 pandemic, the exact effects of Twitter's curation algorithm generally remain unknown. In this work, we present a sock-puppet audit that aims to characterize the effects of algorithmic curation on source diversity and topic diversity in Twitter timelines. We created eight sock puppet accounts to emulate representative real-world

users, selected through a large-scale network analysis. Then, for one month during early 2020, we collected the puppets' timelines twice per day. Broadly, **our results show that algorithmic curation increases source diversity in terms of both Twitter accounts and external domains**, even though it drastically decreases the number of external links in the timeline. In terms of topic diversity, **algorithmic curation had a mixed effect, slightly amplifying a cluster of politically-focused tweets while squelching clusters of tweets focused on COVID-19 fatalities and health information**. Finally, we present some evidence that the **timeline algorithm may exacerbate partisan differences in exposure to different sources and topics.** The paper concludes by discussing broader implications in the context of algorithmic gatekeeping.

**3.3.20** Martel, Pennycook, & Rand (2020). Reliance on emotion promotes belief in fake news. *Cognitive Research: Principles and Implications*.

ABSTRACT: What is the role of emotion in susceptibility to believing fake news? Prior work on the psychology of misinformation has focused primarily on the extent to which reason and deliberation hinder versus help the formation of accurate beliefs. Several studies have suggested that people who engage in more reasoning are less likely to fall for fake news. However, the role of reliance on emotion in belief in fake news remains unclear. To shed light on this issue, we explored the relationship between experiencing specific emotions and believing fake news (Study 1; N = 409). We found that across a wide range of specific emotions, heightened emotionality at the outset of the study was predictive of greater belief in fake (but not real) news posts. Then, in Study 2, we measured and manipulated reliance on emotion versus reason across four experiments (total N = 3884). **We found both correlational and causal evidence that reliance on emotion increases belief in fake news: self-reported use of emotion was positively associated with belief in fake (but not real) news, and inducing reliance on emotion resulted in greater belief in fake (but not real) news stories compared to a control or to inducing reliance on reason.** These results shed light on the unique role that emotional processing may play in susceptibility to fake news.

**3.3.21** Pennycook, Cannon, & Rand (2018). Prior Exposure Increases Perceived Accuracy of Fake News. *Journal of Experimental Psychology: General.*

ABSTRACT: The 2016 U.S. presidential election brought considerable attention to the phenomenon of “fake news”: entirely fabricated and often partisan content that is

presented as factual. Here we demonstrate one mechanism that contributes to the believability of fake news: fluency via prior exposure. Using actual fake-news headlines presented as they were seen on Facebook, **we show that even a single exposure increases subsequent perceptions of accuracy, both within the same session and after a week. Moreover, this "illusory truth effect" for fake-news headlines occurs despite a low level of overall believability and even when the stories are labeled as contested by fact checkers or are inconsistent with the reader's political ideology. These results suggest that social media platforms help to incubate belief in blatantly false news stories and that tagging such stories as disputed is not an effective solution to this problem.** It is interesting, however, that we also found that prior exposure does not impact entirely implausible statements (e.g., "The earth is a perfect square"). These observations indicate that although extreme implausibility is a boundary condition of the illusory truth effect, only a small degree of potential plausibility is sufficient for repetition to increase perceived accuracy. As a consequence, the scope and impact of repetition on beliefs is greater than has been previously assumed.

## 3.4. DISCUSSION OF QUESTION 3

[To come: We will add a discussion section at the end of each of our 7 questions, where Jon, Chris, and other researchers will weigh in on what can be concluded from the preponderance of the evidence about this question. If you are a researcher and want to offer your thoughts in brief form, please request edit access]

[Other studies? What have we missed?]

* * * * * * * * * * * * * * *

# QUESTION 4: DOES SOCIAL MEDIA INCREASE THE PROBABILITY OF VIOLENCE?

There are accusations that social media platforms have caused or amplified violence, including inter-communal violence. We include research on how groups that use violence recruit new members and then radicalize them.

The paradigm case, often cited, is that Facebook played a determining role in inciting offline violence in the Rohingya genocide in Myanmar. A Facebook executive even

admitted that the company “failed to prevent its platform from being used to ‘foment division and incite offline violence’ in the country.” Here is Facebook's official statement, along with an independent report conducted by BSR (but commissioned by Facebook). And here is one key book on the topic.

## 4.1 STUDIES INDICATING YES

**4.1.1** Atari, Davani, Kogon, Kennedy, Saxena, Anderson, & Dehghani (2021). Morally homogeneous networks and radicalism. *Social Psychological and Personality Science.*

ABSTRACT: Online radicalization is among the most vexing challenges the world faces today. Here, we demonstrate that homogeneity in moral concerns results in increased levels of radical intentions. In Study 1, we find that in **Gab – a right-wing extremist network – the degree of moral convergence within a cluster, predicts the number of hate-speech messages members post. In Study 2, we replicate this effect in another extremist network; Incels.** In Study 3 (N = 333), we demonstrate that **experimentally leading people to believe that others in their group share their moral views increases their radical intentions.** Study 4 (N = 510) replicates this effect in a stratified representative sample, and finds that this causal link may be explained by the degree to which individuals’ identities are fused with their ingroup. Our findings highlight the role of moral convergence and identity fusion in radicalization, emphasizing the need for diversity of moral worldviews within social networks.

**4.1.2** Mooijman, Hoover, Lin, Ji, & Dehghani (2018). Moralization in social networks and the emergence of violence during protests. *Nature Human Behaviour*.

ABSTRACT: In recent years, protesters in the United States have clashed violently with police and counter-protesters on numerous occasions. Despite widespread media attention, little scientific research has been devoted to understanding this rise in the number of violent protests. We propose that this phenomenon can be understood as a function of an individual’s moralization of a cause and the degree to which they believe others in their social network moralize that cause. Using data from the 2015 Baltimore protests, we show that not only did the degree of moral rhetoric used on social media increase on days with violent protests but also that the hourly frequency of morally relevant tweets predicted the future counts of arrest during protests, suggesting an

association between moralization and protest violence. To better understand the structure of this association, we ran a series of controlled behavioural experiments demonstrating that **people are more likely to endorse a violent protest for a given issue when they moralize the issue; however, this effect is moderated by the degree to which people believe others share their values. We discuss how online social networks may contribute to inflations of protest violence**.

**4.1.3** Müller & Schwarz (2021). Fanning the flames of hate: Social media and hate crime. Journal of the European Economic Association.

ABSTRACT: This paper investigates the link between social media and hate crime. We show that **anti-refugee sentiment on Facebook predicts crimes against refugees in otherwise similar municipalities with higher social media usage.** To establish causality, we exploit exogenous variation in the timing of major Facebook and internet outages. Consistent with a role for "echo chambers," we find that right-wing social media posts contain narrower and more loaded content than news reports. **Our results suggest that social media can act as a propagation mechanism for violent crimes by enabling the spread of extreme viewpoints.**

**4.1.4** Phadke, & Mitra (2020). Many faced hate: A cross platform study of content framing and information sharing by online hate groups. *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. (h/t Tanu Mitra)

ABSTRACT: Hate groups are increasingly using multiple social media platforms to promote extremist ideologies. Yet we know little about their communication practices across platforms. How do hate groups (or "in-groups"), frame their hateful agenda against the targeted group or the "out-group?" How do they share information? Utilizing "framing" theory from social movement research and analyzing domains in the shared links, we juxtapose the Facebook and Twitter communication of 72 Southern Poverty Law Center (SPLC) designated hate groups spanning five hate ideologies. **Our findings show that hate groups use Twitter for educating the audience about problems with the out-group, maintaining positive self-image by emphasizing in-group's high social status, and for demanding policy changes to negatively affect the out-group. On Facebook, they use fear appeals, call for active participation in group events (membership requests), all while portraying themselves as being oppressed by the out-group and failed by the system**. Our

study unravels the ecosystem of cross-platform communication by hate groups, suggesting that they use Facebook for group radicalization and recruitment, while Twitter for reaching a diverse follower base.

**4.1.5** Müller & Scwarz (2020). From hashtag to Hate Crime: Twitter and Anti-Minority Sentiment. *SSRN.* (h/t Naman Garg)

ABSTRACT: We study whether social media can contribute to hatred against minorities with a focus on Donald Trump's political rise. To establish causality, we construct an instrument for Twitter usage based on the platform's early adopters at the South by Southwest (SXSW) festival in 2007, who were crucial for Twitter's diffusion across US counties. Instrumenting with the home counties of SXSW followers who joined in March 2007, while controlling for the counties of SXSW followers who joined before the festival, **we find that a one standard deviation increase in Twitter usage is associated with a 32% larger increase in anti-Muslim hate crimes since the 2016 presidential primaries**. Further, Trump's tweets about Islam-related topics predict increases in xenophobic tweets by his followers, cable news attention paid to Muslims, and hate crimes on the following days. These correlations persist in an instrumental variable framework exploiting that Trump is more likely to tweet about Muslims on days he plays golf.

**4.1.6** Karell, Linke, Holland, & Hendrickson (2023). “Born for a Storm”: Hard-Right Social Media and Civil Unrest. *American Sociological Review*.

ABSTRACT: Does activity on hard-right social media lead to hard-right civil unrest? If so, why? We created a spatial panel dataset comprising hard-right social media use and incidents of unrest across the United States from January 2020 through January 2021. Using spatial regression analyses with core-based statistical area (CBSA) and month fixed effects, we find that greater CBSA-level hard-right social media activity in a given month is associated with an increase in subsequent unrest. The results of robustness checks, placebo tests, alternative analytical approaches, and sensitivity analyses support this finding. To examine why hard-right social media activity predicts unrest, we draw on an original dataset of users’ shared content and status in the online community. **Analyses of these data suggest that hard-right social media shift users’ perceptions of norms, increasing the likelihood they will participate in contentious events they once considered taboo.** Our study sheds new light on social

media's offline effects, as well as the consequences of increasingly common hard-right platforms.

[Other studies? What have we missed?]

## 4.2 STUDIES INDICATING NO

**4.2.1** Asimovic, Nagler, Bonneau, & Tucker (2021). Testing the effects of Facebook usage in an ethnically polarized setting. *Proceedings of the National Academy of Sciences.*

ABSTRACT: Despite the belief that social media is altering intergroup dynamics—bringing people closer or further alienating them from one another—the impact of social media on interethnic attitudes has yet to be rigorously evaluated, especially within areas with tenuous interethnic relations. We report results from a randomized controlled trial in Bosnia and Herzegovina (BiH), exploring the effects of exposure to social media during 1 wk around genocide remembrance in July 2019 on a set of interethnic attitudes of Facebook users. We find evidence that, counter to preregistered expectations, **people who deactivated their Facebook profiles report lower regard for ethnic outgroups than those who remained active. Moreover, we present additional evidence suggesting that this effect is likely conditional on the level of ethnic heterogeneity of respondents' residence.** We also extend the analysis to include measures of subjective well-being and knowledge of news. Here, we find that **Facebook deactivation leads to suggestive improvements in subjective wellbeing and a decrease in knowledge of current events**, replicating results from recent research in the United States in a very different context, thus increasing our confidence in the generalizability of these effects.

[Other studies? What have we missed?]

## 4.3 MIXED RESULTS OR UNCLASSIFIED

**4.3.1** Chang, & Park (2021). Social media use and participation in dueling protests: The case of the 2016–2017 presidential corruption scandal in South Korea. *The International Journal of Press/Politics*.

ABSTRACT: This study examines how citizens' social media use may have influenced their participation in highly polarizing protests during the 2016–2017 corruption scandal in South Korea. As social media users mobilize politically by acquiring varied political information from other users, **social media use created more incentives for citizens to participate in both pro- and anti-impeachment protests during the scandal.** Given that social media is an important arena for political activism, participation in rival protests also influences many motivated protesters to strengthen their side's voices online. Thus, protests may increase citizens' political use of social media. Our empirical analysis suggests that social network service use does not influence citizens' political activities in a unidirectional manner. **We have found that social media use and participation in rival protests reciprocally influence each other.**

[Other studies? What have we missed?]

## 4.4 DISCUSSION OF QUESTION 4

[TO COME]

* * * * * * * * * * * * * * *

# QUESTION 5: DOES SOCIAL MEDIA ENABLE FOREIGN GOVERNMENTS TO INCREASE POLITICAL DYSFUNCTION IN THE UNITED STATES AND OTHER DEMOCRACIES?

This section contains research and reports on the long running Russian disinformation and manipulation campaign against American democracy. Is there evidence that it has been effective? We also include research on what other nations are doing. We focus on interventions that use social media to foment anger, division, conflict and distrust. We hope to give readers a sense of the size of the problem and the specific ways that social media is being used by foreign governments to weaken American democracy and society.

# 5.1 STUDIES AND REPORTS INDICATING YES

**5.1.1** DiResta (Nov 28, 2018). The digital maginot line. *Ribbonfarm.*

*[Note: this is not an empirical report but it is included here because DiResta is a research director at the Stanford Internet Observatory and is among the most knowledgeable people about how bad actors are using social media, and this essay gives a helpful overview of what is going on]*

EXCERPT: There are state-sponsored trolls, destabilizing societies in some countries, and rendering all information channels except state media useless in others. They operate at the behest of rulers, often through military or intelligence divisions. Sometimes, as in the case of Duterte in the Philippines, these digital armies focus on interference in their own elections, using paid botnets and teams of sockpuppet personas to troll and harass opponents, or to amplify their owner's candidacy. Other times, the trolls reach beyond their borders to manipulate politics elsewhere, as was the case with Brexit and the U.S. presidential election of 2016. Sometimes, as in Myanmar, elections aren't the goal at all: there, military-run digital teams incited a genocide…

Influence operations exploit divisions in our society using vulnerabilities in our information ecosystem. We have to move away from treating this as a problem of giving people better facts, or stopping some Russian bots, and move towards thinking about it as an ongoing battle for the integrity of our information infrastructure – easily as critical as the integrity of our financial markets. When it's all done and over with, we'll look back on this era as being as consequential in reshaping the future of the United States and the world as World War II.

**5.1.2** Howard, Ganesh, Liotsiou, Kelly, & François (2019). The IRA, social media and political polarization in the United States, 2012-2018. *U.S. Senate Documents*.

ABSTRACT: Russia's Internet Research Agency (IRA) launched an extended attack on the United States by using computational propaganda to misinform and polarize US voters. This report provides the first major analysis of this attack based on data provided by social media firms to the Senate Select Committee on Intelligence (SSCI). This analysis answers several key questions about the activities of the known IRA accounts. In this analysis, **we investigate how the IRA exploited the tools and platforms of Facebook, Instagram, Twitter, and YouTube to impact US users**. We identify which

aspects of the IRA's campaign strategy got the most traction on social media and the means of microtargeting US voters with particular messages.

- **Between 2013 and 2018, the IRA's Facebook, Instagram, and Twitter campaigns reached tens of millions of users in the United States.**
- **Over 30 million users, between 2015 and 2017, shared the IRA's Facebook and Instagram posts with their friends and family, liking, reacting to, and commenting on them along the way.**
- Peaks in advertising and organic activity often correspond to important dates in the US political calendar, crises, and international events.
- IRA activities focused on the US began on Twitter in 2013 but quickly evolved into a multi-platform strategy involving Facebook, Instagram, and YouTube amongst other platforms.
- The most far reaching IRA activity is in organic posting, not advertisements.
- **Russia's IRA activities were designed to polarize the US public and interfere in elections by:**
    - **campaigning for African American voters to boycott elections or follow the wrong voting procedures in 2016, and more recently for Mexican American and Hispanic voters to distrust US institutions;**
    - **encouraging extreme right-wing voters to be more confrontational; and**
    - **spreading sensationalist, conspiratorial, and other forms of junk political news and misinformation to voters across the political spectrum.**
- Surprisingly, these campaigns did not stop once Russia's IRA was caught interfering in the 2016 election. Engagement rates increased and covered a widening range of public policy issues, national security issues, and issues pertinent to younger voters.
- The highest peak of IRA ad volume on Facebook is in April 2017—the month of the Syrian missile strike, the use of the Mother of All Bombs on ISIS tunnels in eastern Afghanistan, and the release of the tax reform plan.
- IRA posts on Instagram and Facebook increased substantially after the election, with Instagram seeing the greatest increase in IRA activity.
- The IRA accounts actively engaged with disinformation and practices common to Russian "trolling". Some posts referred to Russian troll factories that flooded online conversations with posts, others denied being Russian trolls, and some even complained about the platforms' alleged political biases when they faced account suspension.

**5.1.3** DiResta, Shaffer…Johnson (2019). The tactics & tropes of the Internet Research Agency. *U.S. Senate Documents*.

ABSTRACT: Upon request by the United States Senate Select Committee on Intelligence (SSCI), New Knowledge reviewed an expansive data set of social media posts and metadata provided to SSCI by Facebook, Twitter, and Alphabet, plus a set of related data from additional platforms. The data sets were provided by the three primary platforms to serve as evidence for an investigation into the Internet Research Agency (IRA) influence operations. The organic post content in this data set has never previously been seen by the public. Our report quantifies and contextualizes Internet Research Agency (IRA) influence operations targeting American citizens from 2014 through 2017, and articulates the significance of this long-running and broad influence operation. It includes an overview of Russian influence operations, a collection of summary statistics, and a set of key takeaways that are then discussed in detail later in the document. The document includes links to full data visualizations, hosted online, that permit the reader to explore facets of the IRA-created manipulation ecosystem. Finally, we share our concluding notes and recommendations. We also provide a comprehensive slide deck accommodating a wide array of selected images directly from the data set illustrating our observations, and, as an appendix, a comprehensive summary of relevant statistics related to the data set.

**Broadly, Russian interference in the U.S. Presidential Election of 2016 took three distinct forms, one of which is within the scope of our analysis: ... 3. A sweeping and sustained social influence operation consisting of various coordinated disinformation tactics aimed directly at US citizens, designed to exert political influence and exacerbate social divisions in US culture. This last form of interference, a multi-year coordinated disinformation effort conducted by the Russian state-supported Internet Research Agency (IRA)**, is the topic of this analysis.

**5.1.4** Farkas, & Bastos (2018). IRA propaganda on Twitter: Stoking antagonism and tweeting local news. *Proceedings of the 9th International Conference on Social Media and Society*.

ABSTRACT: This paper presents preliminary findings of a content analysis of tweets posted by false accounts operated by the Internet Research Agency (IRA) in St Petersburg. We relied on a historical database of tweets to retrieve 4,539 tweets posted

by IRA-linked accounts between 2012 and 2017 and coded 2,501 tweets manually. The messages cover newsworthy events in the United States, the Charlie Hebdo terrorist attack in 2015, and the Brexit referendum in 2016. Tweets were annotated using 19 control variables to investigate whether IRA operations on social media are consistent with classic propaganda models. **The results show that the IRA operates a composite of user accounts tailored to perform specific tasks, with the lion's share of their work focusing on US daily news activity and the diffusion of polarized news across different national contexts.**

**5.1.5** Bradshaw & Howard (2019). The global disinformation order: 2019 global inventory of organized social media manipulation.

EXECUTIVE SUMMARY: Over the past three years, we have monitored the global organization of social media manipulation by governments and political parties. Our 2019 report analyses the trends of computational propaganda and the evolving tools, capacities, strategies, and resources.

1. Evidence of organized social media manipulation campaigns which have taken place **in 70 countries, up from 48 countries in 2018 and 28 countries in 2017**. In each country, there is at least one political party or government agency using social media to shape public attitudes domestically.
2. **Social media has become co-opted by many authoritarian regimes**. In 26 countries, **computational propaganda is being used as a tool of information control in three distinct ways: to suppress fundamental human rights, discredit political opponents, and drown out dissenting opinions.**
3. **A handful of sophisticated state actors use computational propaganda for foreign influence operations. Facebook and Twitter attributed foreign influence operations to seven countries (China, India, Iran, Pakistan, Russia, Saudi Arabia, and Venezuela) who have used these platforms to influence global audiences.**
4. **China has become a major player** in the global disinformation order. Until the 2019 protests in Hong Kong, most evidence of Chinese computational propaganda occurred on domestic platforms such as Weibo, WeChat, and QQ. But China’s new-found interest in aggressively using Facebook, Twitter, and YouTube should raise concerns for democracies
5. Despite there being more social networking platforms than ever, **Facebook remains the platform of choice for social media manipulation**. In 56

countries, we found evidence of formally organized computational propaganda campaigns on Facebook.

**5.1.6** Freelon & Lokot (2020). Russian Twitter disinformation campaigns reach across the American political spectrum. *Harvard Kennedy School Misinformation Review*.

ABSTRACT: The IRA is a private company sponsored by the Russian government, which distributes Kremlin-friendly disinformation on social media under false identities (see DiResta et al., 2018; Howard, Ganesh, Liotsiou, Kelly, & Francois, 2018).

- **The IRA engaged with several distinct communities of authentic users—primarily conservatives, progressives, and Black people—which exhibited only minimal overlap on Twitter.**
- Authentic users primarily engaged with IRA accounts that shared their own ideological and/or racial identities.
- **Racist stereotyping, racial grievances, the scapegoating of political opponents, and outright false statements were four of the most common appeals found among the most replied-to IRA tweets.**
- We conducted a network analysis of 2,057,747 authentic replies to IRA tweets over nine years, generated ideology ratings for a random sample of authentic users, and qualitatively analyzed some of the most replied-to IRA tweets.
- **State-sponsored disinformation agents have demonstrated success in infiltrating distinct online communities. Political content attracts far more engagement than non-political content and appears crafted to exploit intergroup distrust and enmity.**
- Collaboration between different political groups and communities might be successful in detecting IRA campaigns more effectively.

**5.1.7** China-linked influence operation on Twitter detected engaging with the U.S. Presidential Election (2021). *Crime and Security Research Institute.*

EXCERPT: **A network of China-linked accounts operating on Twitter was detected in the run up to the 2020 US Presidential election. These accounts were originally identified as being of interest because they repeatedly posted negative messages about President Trump and Joe Biden, made allegations of election fraud, and engaged with negative narratives about the US response to the coronavirus**

**pandemic.** More recently, they have been detected amplifying reactions to the Capitol Building riot in Washington D.C. on 6th January, drawing comparisons to the West's responses to political protests in Hong Kong and quickly disseminating tailored propaganda videos. The network possesses signatures of a co-ordinated information-influence operation.

**5.1.8** Dubow, Lucas, & Morris (2021). Jabbed in the back: Mapping Russian and Chinese information operations during COVID-19. *CEPA*.

EXECUTIVE SUMMARY: During the COVID-19 pandemic, the Chinese Communist Party (CCP) spread disinformation about the efficacy of vaccines and the virus's origins, a shift from Beijing's previous disinformation campaigns, which had a narrower focus on China-specific issues such as Tibet, Hong Kong, and Taiwan.

- Most of Beijing's COVID-19 narratives aimed at shaping perceptions of China's response to the pandemic and only rarely targeted other countries specifically.
- Russia recycled previous narratives and exacerbated tensions in Western society while attempting some propaganda about Russian scientific prowess.
- The Kremlin and the CCP learned from each other. While limited evidence exists of explicit cooperation, instances of narrative overlap and circular amplification of disinformation show that China is following a Russian playbook with Chinese characteristics. Russia is simultaneously learning from the Chinese approach.
- The largest difference between China's and Russia's information warfare tactics remains China's insistence on narrative consistency, compared with Russia's firehose of falsehoods strategy.1 Even with substantially greater resources, this largely prevents Chinese narratives from swaying public opinion or polarizing societies.
- The two authoritarian countries' information operations have evolved over the last 18 months and will continue to do so with the spread of variants, vaccines, and inquiries into the virus's origins.

[Other studies or reports? What have we missed?]

## 5.2 STUDIES AND REPORTS INDICATING NO, OR MINIMAL EFFECTS

**5.2.1** Bail, Guay, Maloney, Combs, Hillygus, Merhout, Freelon, & Volfovsky (2020). Assessing the Russian Internet Research Agency's impact on the political attitudes and behaviors of American Twitter users in late 2017. *Proceedings of the National Academy of Sciences*.

ABSTRACT: There is widespread concern that Russia and other countries have launched social-media campaigns designed to increase political divisions in the United States. Though a growing number of studies analyze the strategy of such campaigns, it is not yet known how these efforts shaped the political attitudes and behaviors of Americans. We study this question using longitudinal data that describe the attitudes and online behaviors of 1,239 Republican and Democratic Twitter users from late 2017 merged with nonpublic data about the Russian Internet Research Agency (IRA) from Twitter. Using Bayesian regression tree models, **we find no evidence that interaction with IRA accounts substantially impacted 6 distinctive measures of political attitudes and behaviors over a 1-mo period.** We also find that interaction with IRA accounts were most common among respondents with strong ideological homophily within their Twitter network, high interest in politics, and high frequency of Twitter usage. **Together, these findings suggest that Russian trolls might have failed to sow discord because they mostly interacted with those who were already highly polarized.** We conclude by discussing several important limitations of our study—especially our inability to determine whether IRA accounts influenced the 2016 presidential election—as well as its implications for future research on social media influence campaigns, political polarization, and computational social science.

**5.2.2** Eady, Paskhalis, Zilinsky, Bonneau, Nagler, & Tucker (2023). 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. *Nature Communications*.

ABSTRACT: There is widespread concern that foreign actors are using social media to interfere in elections worldwide. Yet data have been unavailable to investigate links between exposure to foreign influence campaigns and political behavior. Using longitudinal survey data from US respondents linked to their Twitter feeds, we quantify the relationship between exposure to the Russian foreign influence campaign and attitudes and voting behavior in the 2016 US election. We demonstrate, first, that exposure to Russian disinformation accounts was heavily concentrated: only 1% of users accounted for 70% of exposures. Second, exposure was concentrated among users who strongly identified as Republicans. Third, exposure to the Russian influence

campaign was eclipsed by content from domestic news media and politicians. **Finally, we find no evidence of a meaningful relationship between exposure to the Russian foreign influence campaign and changes in attitudes, polarization, or voting behavior. The results have implications for understanding the limits of election interference campaigns on social media.**

[Other studies? What have we missed?]

## 5.3 UNCLASSIFIED

**5.3.1** McKay, & Tenove (2021). Disinformation as a threat to deliberative democracy. *Political Research Quarterly*.

ABSTRACT: It is frequently claimed that online disinformation threatens democracy, and that disinformation is more prevalent or harmful because social media platforms have disrupted our communication systems. These intuitions have not been fully developed in democratic theory. **This article builds on systemic approaches to deliberative democracy to characterize key vulnerabilities of social media platforms that disinformation actors exploit, and to clarify potential anti-deliberative effects of disinformation.** The disinformation campaigns mounted by Russian agents around the United States' 2016 election illustrate the use of anti-deliberative tactics, including ***corrosive falsehoods, moral denigration*, and *unjustified inclusion***. We further propose that these tactics might contribute to the system-level **anti-deliberative properties of *epistemic cynicism, techno-affective polarization*, and *pervasive inauthenticity***. These harms **undermine a polity's capacity to engage in communication characterized by the use of facts and logic, moral respect, and democratic inclusion.** Clarifying which democratic goods are at risk from disinformation, and how they are put at risk, can help identify policies that go beyond targeting the architects of disinformation campaigns to address structural vulnerabilities in deliberative systems.

[Explains *why* disinformation campaigns by foreign powers are harmful, but does not show that they are happening and if they are harmful]

**5.3.2** McKay & Tenove (2021). Disinformation as a threat to deliberative democracy. *Political Research Quarterly*.

ABSTRACT: The Chinese military's focus on information warfare is expanding to include information operations on social media. Given the possibility of U.S.-China conflict over Taiwan or another regional contingency, understanding how the People's Liberation Army (PLA) thinks about the use of disinformation campaigns on social media has emerged as an important question for U.S. national security policymakers and defense planners. This report describes how the PLA might direct social media disinformation campaigns against the United States and its armed forces, especially the U.S. Air Force. The authors conducted interviews with regional experts during three trips to Asia and reviewed Chinese-language writings and analyses of publicly attributed, or at least reasonably suspected, examples of Chinese disinformation and other malign social media activity on both Chinese and foreign platforms. The authors identify key Chinese practices and the supporting infrastructure and conditions needed to engage in successful social media disinformation campaigns and conclude that China is using Taiwan as a test bed for developing attack vectors. The authors recommend being competitive in shaping and countering messages on social media, working to engage and protect Chinese-American service members (China's most likely targets), and incorporating adversary social media disinformation into future wargames.

KEY FINDINGS:

- China is treating Taiwan as a test bed for developing attack vectors using disinformation on social media.
- To date, in the case of Taiwan, China's use of disinformation has achieved mixed and somewhat limited results that are primarily in the political, not operational, domain.
- **China has not carried out substantial disinformation attacks on other U.S. allies or partners (such as Singapore, the Philippines, or Japan).**
- Nonetheless, as Chinese disinformation during the COVID-19 crisis has shown, **Chinese disinformation campaigns will likely be used to target the United States in the event of a crisis or conflict.** As China moves to incorporate social media further into its military operations, it will increasingly engage in some level of shaping operations during what Western observers would consider the preconflict stage.

[Note: this study is in the “unclassified” section because it does not say that China is already fomenting dysfunction in the United States, although it makes the case that this could happen soon]

## 5.4 DISCUSSION OF QUESTION 5

[TO COME]

* * * * * * * * * * * * * * *

# QUESTION 6: DOES SOCIAL MEDIA DECREASE TRUST?

We include democratic institutions including government and elections. Because of the importance of epistemic institutions for successful liberal democracies, we also include research on journalism, science, and universities.

## 6.1 STUDIES INDICATING YES

**6.1.1** Park, Fisher, Flew, & Dulleck (2020). Global mistrust in news: The impact of social media on trust. *International Journal on Media Management*.

ABSTRACT: Digital platforms such as search engines and social media have become major gateways to news. Algorithms are used to deliver news that is consistent with consumers' preferences and individuals share news through their online social networks. This networked environment has resulted in growing uncertainty about online information which has had an impact on news industries globally. While it is well established that perceptions of trust in news found on social media or via search engines are lower than traditional news media, there has been less discussion about the impact of social media use on perceptions of trust in the news media more broadly. This study fills that gap by examining the influence of social media as news sources and pathways to news on perceptions of the level of news trust at a country level. A secondary data analysis of a 26-country survey in 2016 and 2019 was conducted. The analysis revealed **an increase in social media use for accessing news resulted in a decline in trust in news media generally across the globe. Higher levels of general mistrust in news were related to an increased use of sharing of news.** This paper argues the use of social media for news is closely linked to the increase in news

mistrust, which is likely to continue to rise as the number of people using social media to access news continues to grow.

[NOTE: This article makes causal claims ("impact") using cross-sectional survey data. Make sure to interpret the results of this paper with that in mind]

**6.1.2** Klein, & Robison (2019). Like, post, and distrust? How social media use affects trust in government. *Political Communication*.

ABSTRACT: There is much discussion about the potential negative effects of social media use on people's political attitudes. But, does social media use shape trust in government? We use evidence from the 2012 and 2016 ANES as well as the 2018 American Institutional Confidence Poll to test competing expectations regarding this **question: that social media polarizes versus de-polarizes trust judgments across partisan lines. Our analyses provide greater support for the expectation of polarization.** We then unpack the potential mechanisms behind these findings. We use the number of "stealth" issue campaigns targeted to the respondent's state in 2016 as a proxy for the amount of political conflict the respondent was likely to have experienced when using social media during the 2016 Presidential election. Notably, we find that polarization is substantially impacted by the nature of the voter's broader political environment. These findings are consequential for our understanding of how social media influences public opinion and draws attention to the role of the broader political context for this relationship.

[NOTE: This article makes causal claims using cross-sectional survey data. Make sure to interpret the results of this paper with that in mind]

**6.1.3** Sabatini, & Sarracino (2019). Online social networks and trust. *Social Indicators Research*.

ABSTRACT: We use Italian data from the Multipurpose Household Survey to explore how participation in social networking sites (SNS) such as Facebook and Twitter affects the most economically relevant aspect of social capital, trust. We account for measures of trust in strangers (often referred to as social trust), trust in neighbours (particularized trust) and trust in the police (institutional trust). We address endogeneity in the use of SNS by exploiting the variation in the availability of broadband for high-speed Internet, which relates to technological characteristics of the pre-existing voice

telecommunication infrastructures. **We find that all the forms of trust significantly decrease with participation in online networks.** We discuss several interpretations of the results in light of the specific features of Internet-mediated social interaction.

**6.1.4** Bekmagambetov et al. (2018). Critical social media information flows: political trust and protest behaviour among Kazakhstani college students. *Central Asian Survey.*

ABSTRACT: In political regimes where traditional mass media are under state control, social networking sites may be the only place where citizens are exposed to and exchange dissident information. Despite all the attempts, complete control of social media seems to be implausible. **We argue that the critical information that people see, read and share online undermines their trust in political institutions.** This diminishing trust may threaten the legitimacy of the ruling regime and stimulate protest behaviour. **We rely on original survey data of Kazakhstani college students to confirm these expectations.** The data are unique in that they directly measure exposure to critical/dissident information, as opposed to simply assuming it. The analysis leverages Coarsened Exact Matching to simulate experimental conditions. This allows us to better identify the consequential mechanism and the attitudinal precursor by which social media influence protest in an authoritarian context.

**6.1.5** Praprotnik, Perlot, Ingruber, & Filzmaier (2019). Social media as information channel. *Österreichische Zeitschrift für Politikwissenschaft*.

ABSTRACT: A vivid political discourse is a crucial component of a functioning democracy. Since user numbers of social networks are increasing, the political debate via these channels becomes more important. Therefore, the present study analyses the consumers of political information through social networks, using the case of Austria as an example. The models are based upon a secondary data analysis of the Digitalmonitor (N=1.200). Our results show that **social media consumers of political information are, among other things, highly interested in politics, hold rather extreme values on a political left-right scale and have little trust in traditional media channels.** We conclude that social media does not guarantee equal access to information. However, for people dissatisfied with the traditional media, it provides an alternative.

**6.1.6** Choli, & Kuss (2021). Perceptions of blame on social media during the coronavirus pandemic. *Computers in Human Behavior*.

ABSTRACT: The outbreak of the coronavirus (COVID-19) disease is overwhelming resources, economies and countries around the world. Millions of people have been infected and hundreds of thousands have succumbed to the virus. Research regarding the coronavirus pandemic is published every day. However, there is limited discourse regarding societal perception. Thus, this paper examines blame attribution concerning the origin and propagation of the coronavirus crisis according to public perception. Specifically, data were extracted from the social media platform Twitter concerning the coronavirus during the early stages of the outbreak and further investigated using thematic analysis. The findings revealed **the public predominantly blames national governments for the coronavirus pandemic. In addition, the results documented the explosion of conspiracy theories among social media users regarding the virus' origin.** In the early stages of the pandemic, the blame tendency was most frequent to conspiracy theories and restriction of information from the government, whilst in the later months, responsibility had shifted to political leaders and the media. The findings indicate an emerging government mistrust that may result in disregard of preventive health behaviours and the amplification of conspiracy theories, and an evolving dynamic of blame. This study argues for a transparent, continuing dialogue between governments and the public to stop the spread of the coronavirus.

**6.1.7** Enders, Uscinski, Seelig…Stoler (2021). The relationship between social media use and beliefs in conspiracy theories and misinformation. *Political Behavior*.

ABSTRACT: Numerous studies find associations between social media use and beliefs in conspiracy theories and misinformation. While such findings are often interpreted as evidence that social media causally promotes conspiracy beliefs, we theorize that this relationship is conditional on other individual-level predispositions. Across two studies, we examine the relationship between beliefs in conspiracy theories and media use, finding that **individuals who get their news from social media and use social media frequently express more beliefs in some types of conspiracy theories and misinformation. However, we also find that these relationships are conditional on conspiracy thinking—the predisposition to interpret salient events as products of conspiracies**—such that social media use becomes more strongly associated with conspiracy beliefs as conspiracy thinking intensifies. This pattern, which we observe across many beliefs from two studies, clarifies the relationship between social media use and beliefs in dubious ideas.

**6.1.8** Aruguete, Calvo, Scartascini, & Ventura (2022, Working Paper). Trustful voters, trustworthy politicians: A survey experiment on the influence of social media on trust. *OSF working paper.* (h/t Tiago Ventura)

ABSTRACT: Recent increases in polarization in social media raise questions about the relationship between social media and the decline in political trust around the world. To evaluate this claim causally, we implement a variant of the well-known trust game in a survey experiment with 4,800 respondents in Brazil and Mexico and test for the effect of social media exposure on trust and trustworthiness. We measure the extent to which voters place their trust in others and are themselves trustworthy after being treated with social media messages from in-group or out-group politicians, and with a polarizing partisan or non-partisan message. **Results provide robust support for a negative effect of polarizing partisan discourse on trust behavior and null results on trustworthiness. The negative effect on trust is considerably greater among randomly treated respondents who decided to engage with social media messages.** Findings showing that engagement is an important mediator in reducing trust provide several theoretical implications for studies on behavioral effects of social media incidental exposure.

**6.1.9** Kiratli (2023). Social Media Effects on Public Trust in the European Union. *Public Opinion Quarterly*.

ABSTRACT: This paper scrutinizes the effect of social media use on institutional trust in the European Union (EU) among European citizens. Fixed-effects regression models on data from the Eurobarometer survey conducted in 2019, the year of the most recent European Parliament (EP) elections, demonstrate that higher social media use is associated with lower trust in the EU**. More importantly, social media usage habits exert particularly detrimental effects in regions with wider and faster internet connections. In such high-information environments, those who more frequently use online social networks, tend to trust those networks, and receive information on EU affairs from these networks have less faith in the EU compared to those in regions with lower-quality internet access. In contrast, in regions with lower broadband access, receiving EU information from social media fosters political trust.**

[Other studies? What have we missed?]

## 6.2 STUDIES INDICATING NO, OR MINIMAL EFFECTS

[It is interesting to note that none of the 3 studies in this section are about the USA]

**6.2.1** Valenzuela, Halpern & Araneda (2021). A downward spiral? A panel study of misinformation and media trust in Chile. *The International Journal of Press/Politics*.

ABSTRACT: Despite widespread concern, research on the consequences of misinformation on people's attitudes is surprisingly scant. To fill in this gap, the current study examines the long-term relationship between misinformation and trust in the news media. Based on the reinforcing spirals model, we analyzed data from a three-wave panel survey collected in Chile between 2017 and 2019. We found a weak, over-time relationship between misinformation and media skepticism. Specifically, **initial beliefs on factually dubious information were negatively correlated with subsequent levels of trust in the news media. Lower trust in the media, in turn, was related over time to higher levels of misinformation.** However, **we found no evidence of a reverse, parallel process where media trust shielded users against misinformation, further reinforcing trust in the news media. The lack of evidence of a downward spiral suggests that the corrosive effects of misinformation on attitudes toward the news media are less serious than originally suggested**. We close with a discussion of directions for future research.

**6.2.2** Huber, Barnidge, Gil de Zuniga, & Liu (2019). Fostering public trust in science: The role of social media. *Public Understanding of Science.*

ABSTRACT: The growing importance of social media for getting science news has raised questions about whether these online platforms foster or hinder public trust in science. Employing multilevel modeling, this study leverages a **20-country survey to examine the relationship between social media news use and trust in science. Results show a positive relationship between these variables across countries.** Moreover, the between-country variation in this relationship is related to two cultural characteristics of a country, individualism/collectivism and power distance.

**6.2.3** Placek (2017). #Democracy: Social media use and democratic legitimacy in Central and Eastern Europe. *Democratization*.

ABSTRACT: Since 1989, many of the former communist countries in Central and Eastern Europe (CEE) have made the dramatic change from communist regimes to democratic nations that are integrated in the European sphere. While these sweeping changes have given rise to a successful transition to democracy unlike any the world has ever seen, there remain issues with governance as well as citizen support for the regime. While other studies have shown that mass media can influence a person's attitudes and opinions in the region, none has explored what effect social media can have on orientations toward democracy in the region. In the following paper, I build several hypotheses based on previous studies of media effects and democratic survival. I then employ survey data to empirically test whether social media increases support for democracy. **The study finds that not only does using social media increase support for democracy, but also simple usage rather than information seeking provides more consistent effects on a person's support for democracy in CEE.**

[Other studies? What have we missed?]

## 6.3 MIXED RESULTS OR UNCLASSIFIED

[what have we missed?]

## 6.4 DISCUSSION OF QUESTION 6

[to come]

* * * * * * * * * * * * * * * *

# QUESTION 7: DOES SOCIAL MEDIA STRENGTHEN POPULIST MOVEMENTS?

Although we look at populism across the political spectrum, the majority of the research we have found examines right-wing movements.

## 7.1 STUDIES INDICATING YES

**7.1.1** Schumann, Thomas, Ehrke, Bertlich, & Dupont (2021). Maintenance or change? Examining the reinforcing spiral between social media news use and populist attitudes. *Information, Communication & Society*.

ABSTRACT: Citizens around the world increasingly express support for populism. Here, we apply the reinforcing spirals model to examine whether, and how, social media news use shapes populist attitudes over time. Specifically, we assess if using social media as a news source serves to maintain existing populist attitudes or facilitates a shift in attitudes to a more extreme position. **A cross-sectional survey (*N*1 = 195)** highlighted a **positive correlation between social media news use and populist attitudes.** A **four-wave longitudinal survey (*N*2 = 386)** further showed that this relationship reflects media and selection effects. Over a period of three months, **more frequent social media news use predicted stronger populist attitudes at subsequent measuring points.** In addition, higher levels of populist attitudes were related to more frequent social media news consumption in the following waves. However, **the frequency of social media news use did not change over time and populist attitudes did not become stronger during the study period.** Taken together, the findings indicate that **social media news use contributed to the maintenance of populist attitudes at a stable level. There is no evidence to suggest social media news use predicted more extreme populist attitudes.** We discuss these results with respect to the (potentially continued) rise of populism; we also critically reflect on the phenomenon of attitude polarization online.

**7.1.2** Müller, & Bach (2021). Populist alternative news use and its role for elections: Web-tracking and survey evidence from two campaign periods. *New Media & Society*.

ABSTRACT: This study explores voters' populist alternative news use during (different types of) democratic elections and investigates starting points for preventing potentially harmful effects. We draw from two combined data sets of web-tracking and survey data which were collected during the 2017 German *Bundestag* campaign (1523 participants) and the 2019 European Parliamentary election campaign in Germany (1009 participants). Results indicate that while populist alternative news outlets drew more interest during the first-order election campaign, they reached only 16.5% of users even then. Moreover, most users visited their websites rather seldom. Nonetheless, our **data suggest that alternative news exposure is strongly linked to voting for (right-wing) populist parties. Regarding the origins of exposure, our analyses punctuate the role of platforms in referring users to populist alternative news. About 40% of website visits originated from Facebook alone in both data sets and another third of visits from search engines.** This raises questions about algorithmic accountability.

**7.1.3** Heiss, & Matthes (2020). Stuck in a nativist spiral: Content, selection, and effects of right-wing populists' communication on Facebook. *Political Communication*.

ABSTRACT: Although social media have become important venues for right-wing populist (RWP) campaigns, the content, selection, and effects of RWP messages on social media remain largely unknown. Using content and panel analysis in two studies, we investigated the potential reciprocal relationship between RWP communication on social media and citizens' anti-immigrant attitudes, anti-elitist attitudes, and feelings of anger and anxiety. In Study 1, we analyzed 13,358 Facebook posts from German and Austrian political parties and their leading candidates. Among our results, RWP actors conveyed anti-immigrant and anti-elitist messages more often than non-RWP actors, and anti-immigrant messages especially induced negative emotional responses among followers of RWP actors. In Study 2, our analysis of data from a two-wave panel study with 559 respondents revealed that **anti-immigrant attitudes drove selective exposure to RWP content on Facebook, which consequently fueled anti-immigrant attitudes, and that selective exposure to such content increased individuals' anti-elitist attitudes and anxiety.**

**7.1.4** Mosca & Quaranta (2021). Are digital platforms potential drivers of the populist vote? A comparative analysis of France, Germany and Italy. *Information, Communication & Society*.

ABSTRACT: Populist parties are often argued to be very skilled in using digital media to attract supporters and strengthen linkages with their followers. However, only rarely has research shown this linkage empirically. This study explores whether arguments about the relation between digital platforms and populist voting can be substantiated using comparative survey data in France, Germany and Italy. Digital media include a variety of online platforms that can affect populist vote in different ways. This article addresses the relation between the *political use* of digital platforms and the populist vote. First, it looks at how the use of Social Networking Sites (SNS) and Mobile Instant Messaging Services (MIMS) is related to voting for populist parties. Second, it assesses whether the role of digital platforms is different for supporting digital 'immigrant' and digital 'native' populist parties. Third, it explores country differences in the relation between SNS and MIMS' use and the populist vote. Using original online surveys, the article shows that **political activities on SNS and MIMS platforms (sending messages or posting, discussing or convincing others to vote for a candidate) increase the probability of voting for populist parties. However, it also finds that the political use of digital media is associated with the populist vote under certain (and limited) circumstances, that is only for a subset of populist parties.** Finally, it identifies important differences in how SNS and MIMS are linked to the populist vote in countries presenting diverse institutional features, web regulations and constellations of media systems.

**7.1.5** Schumann, Boer, Hanke, & Liu (2021). Social media use and support for populist radical right parties: Assessing exposure and selection effects in a two-wave panel study. *Information, Communication & Society*.

ABSTRACT: Vote shares for populist radical right parties (PRRPs) have increased considerably in recent years, and this advancement of PRRPs has been attributed in part to social media. We assess the affinity between social media and populist radical right parties by examining a) whether more frequent social media use for news enhances the willingness to vote for a PRRP (exposure effect) as well as b) whether individuals who have voted for a PRRP in the past use social media more frequently to access news (selection effect). To address these research questions, we analysed data of a two-wave survey study that was conducted in Germany, focusing on the party Alternative for Germany (AfD). **Binary logistic regression highlighted that social media use increased the likelihood of supporting the AfD. Pre-registered multinominal analyses, however, showed that this effect was driven by specific party comparisons.** That is, using the AfD as a reference category, social media use reduced intentions to vote for parties that expressed similar positions as the AfD on the

issue of immigration and with which the PRRP competes over votes. Social media selection effects were not supported.

**7.1.6** Schulze (2020). Who uses right-wing alternative online media? An exploration of audience characteristics. *Politics and Governance*.

ABSTRACT: Accompanying the success of the radical right and right-wing populist movements, right-wing alternative online media have recently gained prominence and, to some extent, influence on public discourse and elections. The existing scholarship so far focuses primarily on the role of content and social media distribution and pays little attention to the audiences of right-wing alternative media, especially at a cross-national level and in the European context. The present paper addresses this gap by exploring the characteristics of the audiences of right-wing alternative online media. Based on a secondary data analysis of the 2019 Reuters Digital News Survey, this article presents a cross-national analysis of right-wing alternative media use in Northern and Central Europe. The results indicate a comparatively high prevalence of right-wing alternative online media in Sweden, whereas in Germany, Austria, and Finland, these news websites seem to be far less popular. With regard to audience characteristics, the strongest predictors of right-wing alternative online media use are political interest and a critical stance towards immigration, accompanied by a skeptical assessment of news quality, in general, and distrust, especially in public service broadcasting media. Additionally, **the use of social media as a primary news source increases the likelihood of right-wing alternative news consumption.** This corroborates the high relevance of social media platforms as distributors and multipliers of right-wing alternative news content. The findings suggest that right-wing alternative online media should not be underestimated as a peripheral phenomenon, but rather have to be considered influential factors for center-right to radical right-leaning politics and audiences in public discourse, with a high mobilizing and polarizing potential.

**7.1.7** Santini, Salles, & Tucci (2021). Comparative approaches to mis/disinformation | When machine behavior targets future voters: The use of social bots to test narratives for political campaigns in Brazil. *International Journal of Communication*.

ABSTRACT: In 2018, the election of Jair Bolsonaro for the Brazilian presidency was associated with dubious propaganda strategies implemented through social media. The purpose of this article is to understand the early development of key communication

strategies of his presidential campaign since 2016. We used a combination of observational, discourse, and content analysis based on digital trace data to investigate how Bolsonaro had been testing his campaign targets and segmentation, as well as cultivating bot accounts and botnets on Twitter during the 2016 Rio de Janeiro municipal election. Our research suggests that the automation of different supporter profiles to target potential voter identities and the experimental dissemination of divisive narratives ensured the effectiveness of his communication persuasion. This finding contributes to the growing body of knowledge regarding his controversial online efforts, adding to the urgent research agenda on Brazil's democratic setback.

**7.1.8** Serrano, Shahrezaye, Papakyriakopoulos, & Hegelich (2019). The rise of Germany's AfD: A social media analysis. *Proceedings of the 10th International Conference on Social Media and Society*.

ABSTRACT: In 2017, a far-right party entered the German parliament for the first time in over half a century. The Alternative für Deutschland (AfD) became the third largest party in the government. Its campaign focused on Euroscepticism and a nativist stance against immigration. The AfD used all available social media channels to spread this message. This paper seeks to understand the AfD's social media strategy over the last years on the full gamut of social media platforms and to verify the effectiveness of the party's online messaging strategy. For this purpose, we collected data related to Germany's main political parties from Facebook, Twitter, YouTube, and Instagram. This data was subjected to a unified multi-platform analysis, which relies on four measures: party engagement, user engagement, message spread, and acceptance. This analysis proves the **AfD's superior online popularity relative to the rest of Germany's political parties. The evidence also indicates that automated accounts contributed to this online superiority.** Finally, we demonstrate that as part of its social media strategy, the AfD avoided discussion of its economic proposals and instead focused on pushing its anti-immigration agenda to gain popularity.

**7.1.9** Bobba, Cremonesi, Mancosu, & Seddone (2018). Populism and the gender gap: Comparing digital engagement with populist and non-populist Facebook pages in France, Italy, and Spain. *The International Journal of Press/Politics*.

ABSTRACT: This paper clarifies whether and to what extent populist communication could drive different gender-oriented reactions. We adopted an original research design intending Facebook as a natural environment where investigating the interaction

between social media users and populist and non-populist parties. Our case selection considers three countries falling into the pluralist polarized media system: France, Italy, and Spain. A human content analysis was carried out on a sample of 2,235 Facebook posts published during thirty days in 2016 by the four main parties/leaders in each country. An original algorithm allowed to identify the gender of users liking each message. We tested whether men tend more to provide likes to messages posted by populist parties, messages published by radical populists, messages containing populist contents, and different components of populist messages. Findings confirm the existence of a gender-oriented reaction to populism: **Men tend to support populist actors and parties on Facebook more than women do, by providing likes to their content. Yet the difference in gender gap between radical and moderate parties is not significant. We also found that the antielite component of populist discourse obtains more likes by male Facebook users.** This pattern is common for both populist and non-populist parties.

**7.1.10** Bliuc, Betts, Faulkner, Vergani, Chow, Iqbal, & Best (2020). The effects of local socio-political events on group cohesion in online far-right communities. *PLoS ONE.*

ABSTRACT: In recent years, the reach and influence of far-right ideologies have been extended through online communities with devastating effects in the real world. In this research, we **examine how far-right online communities can be empowered by socio-political events that are significant to them.** Using over 14 years of data extracted from an Australian national sub-forum of a global online white supremacist community, we investigate whether the group cohesion of the community is affected by local race riots. Our analysis shows that **the online community, not only became more cohesive after the riots, but was also reinvigorated by highly active new members who joined during the week of the riots or soon after.** These changes were maintained over the longer-term, highlighting pervasive ramifications of the local socio-political context for this white supremacist community. Pre-registered analyses of data extracted from other white supremacist online communities (in South Africa and the United Kingdom) show similar effects on some of the indicators of group cohesion, but of reduced magnitude, and not as enduring as the effects found in the context of the Australian far-right online community.

**7.1.11** Wilkerson, Riedl, & Whipple (2021). Affective affordances: Exploring Facebook reactions as emotional responses to hyperpartisan political news. *Digital Journalism*.

ABSTRACT: This research examines the key characteristics of hyperpartisan news pages on Facebook and how audiences interact with politically polarized content through the visual-emotional shorthand of Facebook Reactions. Through a quantitative content analysis of 4,236 posts shared by the most popular hyperpartisan U.S. Facebook pages before, during, and after the 2016 U.S. Presidential Election, the researchers introduce the concept of affective affordances to analyse emotional reactions elicited through Facebook Reactions in response to right- and left-leaning Facebook news posts, as well as the political topics, rhetorical devices, stylistic devices and emotionally charged content that are most likely to elicit emotional responses and inspire shares and comments from audiences in reaction to liberal and conservative content. The results are interpreted in light of the theory of affective intelligence.

**7.1.12** Reuning,Whitesell, & Hannah (2022). Facebook algorithm changes may have amplified local republican parties. *Research & Politics*.

ABSTRACT: In this research note we document changes to the rate of comments, shares, and reactions on local Republican Facebook pages. Near the end of 2018, local Republican parties started to see a much higher degree of interactions on their posts compared to local Democratic parties. We show how this increase in engagement was unique to Facebook and happened across a range of over a thousand local parties. In addition, we use a changepoint model to identify when the change happened and find it lines up with reported information about the change in Facebook's algorithm in 2018. We conclude that it seems possible that changes in how Facebook rated content led to a doubling of the total shares of local Republican party posts compared to local Democratic party posts in the first half of 2019 even though Democratic parties posted more often during this period. **Regardless of Facebook's motivations, their decision to change the algorithm might have given local Republican parties greater reach to connect with citizens and shape political realities for Americans.** The fact that private companies can so easily control the political information flow for millions of Americans raises clear questions for the state of democracy.**.**

[Other studies? What have we missed?]

## 7.2 STUDIES INDICATING NO, OR MINIMAL EFFECTS

**7.2.1** Carrella (2020). #Populism on Twitter: Statistical analysis of the correlation between tweet popularity and "populist" discursive features, *Brno Studies in English.*

ABSTRACT: Recent political events, such as the Brexit or Donald Trump's electoral success, have led to a proliferation of studies focusing on populism nature (Müller 2017; Mudde and Kaltwasser 2017). Part of the literature has also investigated communicative aspects of populism, highlighting how populists are benefitting from the use of social media (Bartlett 2014; Gerbaudo 2018). This research offers further insights on the subject by analyzing populist discourse on Twitter and exploring the correlation between the presence of linguistic features linked to populism, such as emotionalization, simplified rhetoric and intensified claims (Canovan 1999; Heinisch 2008), and tweet popularity. The use of linear mixed effects models revealed a **positive correlation between the linguistic elements of interest and tweet popularity, not only in the populist sample, but also in the control group composed by establishment politicians.** Surprisingly, **reference tweets received more popularity than populist messages when the discursive features analyzed were present.**

**7.2.2** Boulianne, Koc-Michalska, & Bimber (2020). Right-wing populism, social media and echo chambers in Western democracies. *New Media & Society*.

ABSTRACT: Many observers are concerned that echo chamber effects in digital media are contributing to the polarization of publics and in some places to the rise of right-wing populism. This study employs survey data collected in France, the United Kingdom, and the United States (1500 respondents in each country) from April to May 2017. **Overall, we do not find evidence that online/social media explain support for right-wing populist candidates and parties. Instead, in the USA, use of online media decreases support for right-wing populism.** Looking specifically at echo chambers measures, we find offline discussion with those who are similar in race, ethnicity, and class positively correlates with support for populist candidates and parties in the UK and France. The findings challenge claims about the role of social media and the rise of populism.

**7.2.3** Jeroense, Luimers, Jacobs, & Spierings (2021). Political social media use and its linkage to populist and postmaterialist attitudes and vote intention in the Netherlands. *European Political Science*.

ABSTRACT: This study focuses on social media use of citizens from two groups that are often associated with the rise of social media: populist and postmaterialist citizens. Considering their ideological underpinnings, we theorize that they will make more political use of social media and that this further reifies their political attitudes into voting for populist and postmaterialist parties, respectively. Using unique survey data including the relatively new populist attitudes and political use of social media, we test this theory on the Dutch case. We find that both groups do not read political news or connect to politicians more, but both are more likely to react to political content. Moreover, **social media use does not seem to lead to a retention in one's own ideological funnel signified by populist or postmaterialist voting. Among more postmaterialist citizens, passive social media use even makes it more likely to vote for other parties.**

[Other studies? What have we missed?]

## 7.3 MIXED RESULTS OR UNCLASSIFIED

[Other studies? What have we missed?]

## 7.4 DISCUSSION OF QUESTION 7

[TO COME]

* * * * * * * * * * * * * * *

# 8. OTHER STUDIES NOT YET CLASSIFIED

**8.1** Chang, Cheng, & Danescu-Niculescu-Mizil (2020). Don't let me be misunderstood: comparing intentions and perceptions in online discussions. *Proceedings of The Web Conference*.

ABSTRACT: Discourse involves two perspectives: a person's intention in making an utterance and others' perception of that utterance. The misalignment between these perspectives can lead to undesirable outcomes, such as misunderstandings, low productivity and even overt strife. In this work, we present a computational framework for exploring and comparing both perspectives in online public discussions.

We combine logged data about public comments on Facebook with a survey of over 16,000 people about their intentions in writing these comments or about their perceptions of comments that others had written. Unlike previous studies of online discussions that have largely relied on third-party labels to quantify properties such as sentiment and subjectivity, our approach also directly captures what the speakers actually intended when writing their comments. In particular, our analysis focuses on judgments of whether a comment is stating a fact or an opinion, since these concepts were shown to be often confused.

We show that intentions and perceptions diverge in consequential ways. **People are more likely to perceive opinions than to intend them, and linguistic cues that signal how an utterance is intended can differ from those that signal how it will be perceived.**

**Further, this misalignment between intentions and perceptions can be linked to the future health of a conversation: when a comment whose author intended to share a fact is misperceived as sharing an opinion, the subsequent conversation is more likely to derail into uncivil behavior than when the comment is perceived as intended.** Altogether, these findings may inform the design of discussion platforms that better promote positive interactions.

**8.2** Serrano, Papakyriakopoulos, & Hegelich (2020). Dancing to the partisan beat: A first analysis of political communication on TikTok. *12th ACM Conference on Web Science*.

ABSTRACT: TikTok is a video-sharing social networking service, whose popularity is increasing rapidly. It was the world's second-most downloaded app in 2019. Although the platform is known for having users posting videos of themselves dancing,

lip-syncing, or showcasing other talents, user-videos expressing political views have seen a recent spurt. This study aims to perform a primary evaluation of political communication on TikTok. We collect a set of US partisan Republican and Democratic videos to investigate how users communicated with each other about political issues. With the help of computer vision, natural language processing, and statistical tools, we illustrate that political communication on TikTok is much more interactive in comparison to other social media platforms, with users combining multiple information channels to spread their messages. We show that political communication takes place in the form of communication trees since users generate branches of responses to existing content. In terms of user demographics, we find that users belonging to both the US parties are young and behave similarly on the platform. However, Republican users generated more political content and their videos received more responses; on the other hand, Democratic users engaged significantly more in cross-partisan discussions.

**8.3** Munger, & Phillips (2020). A supply and demand framework for YouTube politics. *The International Journal of Press/Politics.*

ABSTRACT: YouTube is the most used social network in the United States and the only major platform that is more popular among right-leaning users. We propose the “Supply and Demand” framework for analyzing politics on YouTube, with an eye toward understanding dynamics among right-wing video producers and consumers. We discuss a number of novel technological affordances of YouTube as a platform and as a collection of videos, and how each might drive supply of or demand for extreme content. We then provide large-scale longitudinal descriptive information about the supply of and demand for conservative political content on YouTube. We demonstrate that viewership of far-right videos peaked in 2017.

**8.4** Engesser, Ernst, Esser, & Büchel (2017). Populism and social media: How politicians spread a fragmented ideology. *Information, Communication & Society*, *20*(8), 1109–1126.

ABSTRACT: Populism is a relevant but contested concept in political communication research. It has been well-researched in political manifestos and the mass media. The present study focuses on another part of the hybrid media system and explores how politicians in four countries (AT, CH, IT, UK) use Facebook and Twitter for populist purposes. Five key elements of populism are derived from the literature: emphasizing the sovereignty of the people, advocating for the people, attacking the elite, ostracizing

others, and invoking the ‘heartland’. A qualitative text analysis reveals that **populism manifests itself in a fragmented form on social media.** Populist statements can be found across countries, parties, and politicians’ status levels. While a broad range of politicians advocate for the people, attacks on the economic elite are preferred by left-wing populists. Attacks on the media elite and ostracism of others, however, are predominantly conducted by right-wing speakers. Overall, the paper provides an in-depth analysis of populism on social media. It shows that **social media gives the populist actors the freedom to articulate their ideology and spread their messages.** The paper also contributes to a refined conceptualization and measurement of populism in future studies.

ADDITIONAL EXCERPT: **“We conclude that social media are particularly well-suited to meet the communicative preferences of populist actors and that they provide them with a convenient instrument to spread their messages. We could even go so far and argue that populism thrives on the logic of connective action.”**

**8.5** Crockett (2017). Moral outrage in the digital age. *Nature Human Behaviour*.

EXCERPT: As digital media infiltrates our social lives, it is crucial that we understand how this technology might transform the expression of moral outrage and its social consequences. Here, I describe a simple psychological framework for tackling this question (Fig. 1). Moral outrage is triggered by stimuli that call attention to moral norm violations. These stimuli evoke a range of emotional and behavioural responses that vary in their costs and constraints. Finally, expressing outrage leads to a variety of personal and social outcomes. This framework reveals that digital media may exacerbate the expression of moral outrage by inflating its triggering stimuli, reducing some of its costs and amplifying many of its personal benefits.

Figure 1:

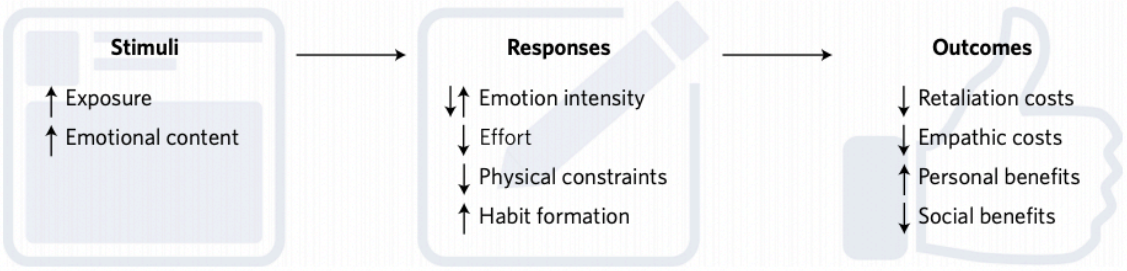



**Fig. 1 | How digital media might transform moral outrage.** Moral outrage is an emotion elicited by stimuli appraised as signifying moral norm violations. The subjective experience of outrage in reaction to such stimuli motivates the expression of behavioural responses such as gossip, shaming or punishment. Expressing outrage can lead to positive and negative outcomes for oneself and for society. Digital media may promote the expression of outrage by magnifying its triggers, reducing its personal costs and amplifying its personal benefits, while at the same time reducing its benefits for society.

**8.6** Angyal & Fellner (2020). How are online and offline political activities connected? A comparison of studies. *Intersections. East European Journal of Society and Politics*.

ABSTRACT: In general, political participation means all the action of citizens that has the aim or the effect of influencing government or politics. Studies argue that media consumption and political participation are correlated: offline and online political participation affect each other. Knowing the relationship between online and offline political activity can improve estimations of offline political events based on social media data.

By comparing these empirical results, in this study we investigate whether social media usage reinforces or weakens the willingness to become involved in a demonstration or other offline political activity. Numerous studies have already attempted to measure this effect, with contradictory findings related to the direction and volume of the latter.

We explore this connection by synthesizing recent empirical political science papers. For this purpose, we compare the results of the former using Bayesian updating – a tool for comparing studies regardless of their methodology or data collection method. This method of data analysis is also insensitive to the operationalization of either the dependent or the explanatory variables.

**Based on the aforementioned studies, our results prove that online political activity has a significant positive effect on offline political activity**, in spite of the fact that some research has found an insignificant connection.

**8.7** Wittenberg, Tappin, Berinsky, & Rand (2021). The (minimal) persuasive advantage of political video over text. *Proceedings of the National Academy of Sciences*.

ABSTRACT: Video is an increasingly common source of political information. Although conventional wisdom suggests that video is much more persuasive than other communication modalities such as text, this assumption has seldom been tested in the political domain. Across two large-scale randomized experiments, we find clear evidence that 'seeing is believing': **individuals are more likely to believe an event took place when shown information in video versus textual form.** When it comes to persuasion, however, the advantage of video over text is markedly less pronounced, with only small effects on attitudes and behavioral intentions. Together, these results challenge popular narratives about the unparalleled persuasiveness of political video versus text.

**8.8** Phadke, Samory, & Mitra (2020). What makes people join conspiracy communities?: Role of social factors in conspiracy engagement, *Proceedings of the ACM on Human-Computer Interaction.* (h/t Tanu Mitra)

ABSTRACT: Widespread conspiracy theories, like those motivating anti-vaccination attitudes or climate change denial, propel collective action and bear society-wide consequences. Yet, empirical research has largely studied conspiracy theory adoption as an individual pursuit, rather than as a socially mediated process. What makes users join communities endorsing and spreading conspiracy theories? We leverage longitudinal data from 56 conspiracy communities on Reddit to compare individual and social factors determining which users join the communities. Using a quasi-experimental approach, we first identify 30K future conspiracists—(FC) and 30K matched non-conspiracists—(NC). **We then provide empirical evidence of importance of social factors across six dimensions relative to the individual factors by analyzing 6 million Reddit comments and posts**. Specifically **in social factors, we find that dyadic interactions with members of the conspiracy communities and marginalization outside of the conspiracy communities, are the most important social precursors to conspiracy joining—even outperforming individual factor baselines**. Our results offer quantitative backing to understand social processes and echo chamber effects in conspiratorial engagement, with important implications for democratic institutions and online communities.

[NOTE: We may include this in section 3.1 if we expand the question to include whether social media drives individuals toward conspiracy communities]

**8.9** Feezell & Ortiz (2019). 'I saw it on Facebook': an experimental analysis of political learning through social media. *Information, Communication & Society.* (h/t Jessica Feezell)

ABSTRACT: The maldistribution of political knowledge in society has important consequences for individual-level political behavior and the representativeness of governmental policies. Increased media selectivity threatens to widen the gap between the politically well-informed and the less-informed by decreasing chance encounters with incidental political information. This study asks: Does exposure to incidental political information through social media promote political learning among users? **We conduct two longitudinal, controlled experiments administered through the Facebook platform, and find no statistical difference in the levels of factual political knowledge among participants exposed to political information compared to those who were not. However, those in the treatment group with low political interest may be more likely to venture an incorrect guess than those in the control group, suggesting that exposure to incidental political information through social media may lead to an increase in self-perceived knowledge among some.**

[Other studies? What have we missed?]

* * * * * * * * * * * * * * *

# 9. MAJOR REVIEW ARTICLES, REPORTS, AND DATABASES

**9.1.1*** Barrett, Hendrix, & Sims (2021). Fueling the fire: How social media intensifies political polarization - and what can be done about it. *NYU Stern Center for Business and Human Rights.* [NOTE: Not peer-reviewed]

EXCERPT: This report analyzes the evidence bearing on social media's role in polarization, assesses the effects of severe divisiveness, and recommends steps the government and the social media industry can take to ameliorate the problem. **We conclude that Facebook, Twitter, and YouTube are not the original or main cause of rising U.S. political polarization, a phenomenon that long predates the social media industry. But use of those platforms intensifies divisiveness and thus contributes to its corrosive consequences...**We focus on "affective polarization," a form of partisan hostility characterized by seeing one's opponents as not only wrong on important issues, but also abhorrent, unpatriotic, and a danger to the country's future. This kind of hatred now infects American politics, and social media has helped spread the disease. But as we illustrate, affective polarization and its consequences are not distributed evenly across the political spectrum.

**9.1.2** Fletcher & Jenkins (2019). Polarisation and the news media in Europe: A literature review of the effect of news use on polarisation across Europe. European Parliamentary Research Service. [NOTE: Not peer-reviewed]

ABSTRACT: Across Europe there is as yet little evidence to support the idea that increased exposure to news featuring like-minded or opposing views leads to the widespread polarisation of attitudes. Though some studies have found that both can strengthen the attitudes of a minority who already hold strong views. **Most studies of news use on social media have failed to find evidence of echo chambers and/or filter bubbles, where people are over-exposed to like-minded views. Some studies even find evidence that it increases the likelihood of exposure to opposing views.** The extent to which people self-select news sources in Europe based on their political preferences, as well as the extent to which news outlets produce partisan coverage, still varies greatly by country. In addition to differences between European countries, comparative research often tends to show that the US has much higher levels of partisan news consumption and polarisation, making it difficult to generalise from these findings. There are large gaps in our understanding of the relationship between the news media and polarisation, particularly outside of Western and Northern Europe, and particularly concerning our knowledge of new, more partisan digital-born news sources.

[Note from JH: This is about mere exposure to news, people do encounter other side]

**9.1.3** Deb, Donohue, & Glaisyer (2017). Omidyar white paper: Is social media a threat to democracy? The Omidyar Group. [NOTE: white paper, not peer reviewed]

EXCERPT: It is becoming increasingly apparent that fundamental principles underlying democracy—trust, informed dialogue, a shared sense of reality, mutual consent, and participation—are being put to the test by certain features and attributes of social media. As technology companies increasingly achieve financial success by monetizing public attention, it is worth examining some of the key issues and unintended consequences arising as a result…

**6 KEY ISSUES:**

1. **Echo chambers, polarization, and hyper-partisanship:** Social media platform design, combined with the proliferation of partisan media in traditional channels, has exacerbated political divisions and polarization. Additionally, some social media algorithms reinforce divisions and create echo chambers that perpetuate increasingly extreme or biased views over time.
2. **Spread of false and/or misleading information:** Today, social media acts as an accelerant, and an at-scale content platform and distribution channel, for both viral “dis”-information (the deliberate creation and sharing of information known to be false) and “mis”-information (the inadvertent sharing of false information). These two types of content—sometimes mistakenly conflated into the term “fake news”—are created and disseminated by both state and private actors, in many cases using bots. Each type poses distinct threats for public dialogue by flooding the public square with multiple, competing realities and exacerbating the lack of agreement about what constitutes truth, facts, and evidence.
3. **Conversion of popularity into legitimacy:** The algorithms behind social media platforms convert popularity into legitimacy, overwhelming the public square with multiple, conflicting assertions. In addition, some social media platforms assume user intentionality (e.g. in search queries) and conflate this with interest, via features such as auto-fill search terms. These design mechanisms impute or impose certain ways of thinking, while also further blurring the lines between specialists and laypeople, or between verified and unverified assertions, thus contributing to the already reduced trust in traditional gatekeepers.
4. **Manipulation by “populist” leaders, governments, and fringe actors:** “Populist” leaders use these platforms, often aided by trolls, “hackers for hire” and bots, on open networks such as Twitter and YouTube. Sometimes they are seeking to communicate directly with their electorate. In using such platforms, they subvert established protocol, shut down

dissent, marginalize minority voices, project soft power across borders, normalize hateful views, showcase false momentum for their views, or create the impression of tacit approval of their appeals to extremism. And they are not the only actors attempting to use these platforms to manipulate political opinion— such activity is now acknowledged by governments of democratic countries (like the UK), as well.

5. **Personal data capture and targeted messaging/advertising:** Social media platforms have become a preferred channel for advertising spend. Not only does this monetization model drive businesses reliant on the capture and manipulation of huge swathes of user data and attention, it also widens the gap between the interests of publishers and journalists and erodes traditional news organizations' revenues. The resulting financial strain has left news organizations financially depleted and has reduced their ability to produce quality news and hold the powerful to account. In addition, advanced methods for capturing personal data have led to sophisticated psychographic analysis, behavioral profiling, and micro-targeting of individuals to influence their actions via so-called "dark ads."
6. **Disruption of the public square:** Some social media platforms have user policies and technical features that enable unintended consequences, like hate speech, terrorist appeals, and racial and sexual harassment, thus encouraging uncivil debate. This can lead members of frequently targeted groups—such as women and minorities—to self-censor or opt out of participating in public discourse. Currently, there are few options for redress. At the same time, platforms are faced with complex legal and operational challenges with respect to determining how they will manage speech, a task made all the more difficult since norms vary widely by geographic and cultural context.

### 9.1.4 Finkel, Bail, Cikara, Ditto, Iyengar, Klar, Mason, McGrath...Druckman (2020). Political sectarianism in America. *Science*.

ABSTRACT: Political polarization, a concern in many countries, is especially acrimonious in the United States. For decades, scholars have studied polarization as an ideological matter—how strongly Democrats and Republicans diverge vis-à-vis political ideals and policy goals. Such competition among groups in the marketplace of ideas is a hallmark of a healthy democracy. But more recently, researchers have identified a second type of polarization, one focusing less on triumphs of ideas than on dominating

the abhorrent supporters of the opposing party. This literature has produced a proliferation of insights and constructs but few interdisciplinary efforts to integrate them. We offer such an integration, pinpointing the superordinate construct of political sectarianism and identifying its three core ingredients: othering, aversion, and moralization. We then consider the causes of political sectarianism and its consequences for U.S. society—especially the threat it poses to democracy. Finally, we propose interventions for minimizing its most corrosive aspects.

ADDITIONAL EXCERPT: **Social media technology employs popularity based algorithms that tailor content to maximize user engagement...Maximizing engagement increases affective polarization, they added, especially within "homogeneous networks," or groupings of like-thinking users. This is "in part because of the contagious power of content that elicits sectarian fear or indignation."**

**9.1.5** Tucker, Guess, Barberá, Vaccari, Siegel, Sanovich, Stukal, & Nyhan (2018). Social media, political polarization, and political disinformation: A review of the scientific literature. *Hewlett Foundation.* [NOTE: Not peer-reviewed]

EXECUTIVE SUMMARY: The following report is intended to provide an overview of the current state of the literature on the relationship between social media; political polarization; and political "disinformation," a term used to encompass a wide range of types of information about politics found online, including "fake news," rumors, deliberately factually incorrect information, inadvertently factually incorrect information, politically slanted information, and "hyperpartisan" news. The review of the literature is provided in six separate sections, each of which can be read individually but that cumulatively are intended to provide an overview of what is known—and unknown—about the relationship between social media, political polarization, and disinformation. The report concludes by identifying key gaps in our understanding of these phenomena and the data that are needed to address them.

**9.1.6** Kubin, & von Sikorski (2021). The role of (social) media in political polarization: A systematic review. *Annals of the International Communication Association*.

ABSTRACT: Rising political polarization is, in part, attributed to the fragmentation of news media and the spread of misinformation on social media. Previous reviews have yet to assess the full breadth of research on media and polarization. We systematically

examine 94 articles (121 studies) that assess the role of (social) media in shaping political polarization. Using quantitative and qualitative approaches, **we find an increase in research over the past 10 years and consistently find that pro-attitudinal media exacerbates polarization. We find a hyperfocus on analyses of Twitter and American samples and a lack of research exploring ways (social) media can depolarize. Additionally, we find ideological and affective polarization are not clearly defined, nor consistently measured.** Recommendations for future research are provided.

ADDITIONAL EXCERPT: "*Social Media Use and Polarization*. A majority of papers focused on the effects of selectively exposing oneself to social media content on political polarization. **These studies showed that social media use predicted both ideological and affective polarization** (Cho et al., 2018). However, some suggest the effect of social media use and polarization is **small** (Johnson et al., 2017), and that it is not about what we see on social media, but rather **what we choose to *share*** on social media that drives political polarization (Johnson et al., 2020). Others find real-world implications for social media use, showing that **social media use is linked to participation in polarizing political protests** (Chang & Park, 2020). Also, some research suggests a reciprocal relationship between media exposure and increased political polarization (Chang & Park, 2020).

However, not all research supports this link between social media use and increased political polarization. Two studies suggest there is no effect of social media on polarization (e.g. Valenzuela et al., 2019). However, neither examined Twitter or Facebook, the two primary social media sites where people see political information (e.g. Stier et al., 2018). One study found evidence of *depolarizing* effects on social media (i.e. Facebook), due to exposure to diverse information (Beam et al., 2018).

Given these divergent findings, **the true effect of social media exposure on political polarization remains unclear. It seems in some cases social media exposure may exacerbate polarization while in other contexts or on certain platforms the effects are unobservable or even lead to depolarization.** Future research should consider more clearly defining the conditions where selective exposure to social media exacerbates political polarization."

**9.1.7** Zhuravskaya, Petrova, & Enikolopov (2020). Political effects of the internet and social media. *Annual Review of Economics*.

ABSTRACT: How do the Internet and social media affect political outcomes? We review empirical evidence from the recent political economy literature, focusing primarily on work that considers traits that distinguish the Internet and social media from traditional off-line media, such as low barriers to entry and reliance on user-generated content. We discuss the main results about the effects of the Internet in general, and social media in particular, on voting, street protests, attitudes toward government, political polarization, xenophobia, and politicians' behavior. We also review evidence on the role of social media in the dissemination of fake news, and we summarize results about the strategies employed by autocratic regimes to censor the Internet and to use social media for surveillance and propaganda. We conclude by highlighting open questions about how the Internet and social media shape politics in democracies and autocracies.

ADDITIONAL EXCERPT: The literature has concluded that **in places where the main public grievances are related to corruption, subversion of power, and control of traditional media by autocrats, free Internet and social media do improve accountability by informing the public and facilitating the organization of protests.** This is exactly why autocrats increasingly resort to censoring the Internet, banning those social media that they cannot monitor and flooding with misinformation the social media networks that they cannot ban.

…Yet, the political roles of the Internet and social media are not yet fully understood. **There is some evidence that so far in democracies, populist parties—on both the extreme right and the extreme left of the political spectrum—benefit more than actors in the center from social media's and the Internet's amplification of existing grievances.** However, there are more open questions than answers. First, an important question is whether these results are temporary, namely, whether people will adapt to the new environment and learn to be more critical of what they see online and learn how to fact-check the information they get. **One piece of evidence that points in this direction is the fact that younger people (who are usually more experienced users) seem to be much less affected by false news than older people—or at least, the young share false news much less.**

KEY FINDINGS:

- "The literature shows that false news does spread through social media, and its spread is faster and wider than that of true news. Future research needs to document how persuasive false news is when exposure occurs on social media."

- “The evidence does suggest that extreme voices are propagated through social media and this has real implication for hate crimes.”
- “The available evidence about whether social media increase political polarization is not conclusive.”
- “There is convincing evidence that low entry barriers and the potential for horizontal flows of information make social media a vehicle to facilitate political protests.”
- “The evidence about the Internet, social media, and voting can be summarized as follows. The spread of the Internet and social media has contributed, at least in part, to the electoral success of populists in Europe and to reduced political support for the ruling parties in immature democracies and semi-autocratic regimes. There is also evidence that social media can be used to mobilize voters.”

### **9.1.8** Pew Reports on Polarization.

Pew Research Center has been conducting excellent research on political polarization in the USA since the 1990s. You can access many of their reports by searching for keywords, such as “media polarization,” or “political polarization.” Some of these reports address the role of social media in polarization and political dysfunction, e.g.,

- 64% of Americans say social media have a mostly negative effect on the way things are going in the U.S. today (2020)

### **9.1.9** Lewandowsky, Smillie, Garcia, Hertwig, Weatherall, … & Leiser (2020). Technology and Democracy: Understanding the influence of online technologies on political behaviour and decision-making. *Publications Office of the European Union.*

EXCERPT: Drawing from many disciplines, the report adopts a behavioural psychology perspective to argue that “social media changes people’s political behaviour”. Four pressure points are identified and analysed in detail: the attention economy; choice

architectures; algorithmic content curation; and mis/disinformation. Policy implications are outlined in detail.

**9.1.10** Terren, & Borge-Bravo (2021). Echo chambers on social media: A systematic review of the literature. *Review of Communication Research*.

ABSTRACT: There have been growing concerns regarding the potential impact of social media on democracy and public debate. While some theorists have claimed that ICTs and social media would bring about a new independent public sphere and increase exposure to political divergence, others have warned that they would lead to polarization through the formation of echo chambers. The issue of social media echo chambers is both crucial and widely debated. This article attempts to provide a comprehensive account of the scientific literature on this issue, shedding light on the different approaches, their similarities, differences, benefits, and drawbacks, and offering a consolidated and critical perspective that can hopefully support future research in this area. Con**cretely, it presents the results of a systematic review of 55 studies i**nvestigating the existence of echo chambers on social media, providing a first classification of the literature and identifying patterns across the studies' foci, methods and findings. **We found that conceptual and methodological choices influence the results of research on this issue. Most importantly, articles that found clear evidence of echo chambers on social media were all based on digital trace data. In contrast, those that found no evidence were all based on self-reported data.** Future studies should take into account the possible biases of the different approaches and the significant potential of combining self-reported data with digital trace data.

**9.1.11** Knight Foundation (2018). Avoiding the echo chamber about echo chambers. [NOTE: Not peer-reviewed]

ABSTRACT: Is the expansion of media choice good for democracy? Not according to critics who decry 'echo chambers,' 'filter bubbles,' and 'information cocoons' — the highly polarized, ideologically homogeneous forms of news and media consumption that are facilitated by technology. However, these claims overstate the prevalence and severity of these patterns, which at most capture the experience of a minority of the public.

In this review essay, we summarize the most important findings of the academic literature about where and how Americans get news and information. We focus

particular attention on how much consumers engage in selective exposure to media content that is consistent with their political beliefs and the extent to which this pattern is exacerbated by technology. **As we show, the data frequently contradict or at least complicate the 'echo chambers' narrative, which has ironically been amplified and distorted in a kind of echo chamber effect.**

We instead emphasize three fundamental features of preferences for news about politics. First, there is diversity in the sources and media outlets to which people pay attention. In particular, only a subset of Americans are devoted to a particular outlet or set of outlets; others have more diverse information diets. Second, though some people have high levels of motivation to follow the latest political news, many only pay attention to politics at critical moments, or hardly at all. Finally, the context in which we encounter information matters. Endorsements from friends on social media and algorithmic rankings can influence the information people consume, but these effects are more modest and contingent than many assume. Strikingly, our vulnerability to echo chambers may instead be greatest in offline social networks, where exposure to diverse views is often more rare.

ADDITIONAL EXCERPT: **The evidence for 'echo chambers' is more equivocal than the alarmist tone of popular discussion suggests.** It is true that people tend to prefer congenial political content in studies when given the choice, but these findings are more limited and contingent than people realize. For instance, these tendencies are asymmetric; **people tend to prefer pro-attitudinal information to a greater extent than they avoid counter-attitudinal information. Selective exposure can also be overridden by other factors such as social cues. In addition, behavioral data shows that tendencies toward selective exposure do not translate into real-world outcomes as often as public discussion would suggest.** Commentators often neglect how little political news most people consume — much of the public is not attentive to politics and thus unlikely to be in an echo chamber of any sort. Moreover, among those who do consume more than a negligible amount of political news, most do not get all or even most of it from congenial media outlets.

**9.1.12** Pennycook, & Rand (2021). The psychology of fake news. *Trends in Cognitive Sciences*.

ABSTRACT: We synthesize a burgeoning literature investigating why people believe and share false or highly misleading news online. Contrary to a common narrative whereby politics drives susceptibility to fake news, people are 'better' at discerning truth from falsehood (despite greater overall belief) when evaluating politically concordant

news. Instead, poor truth discernment is associated with lack of careful reasoning and relevant knowledge, and the use of heuristics such as familiarity. Furthermore, there is a substantial disconnect between what people believe and what they share on social media. This dissociation is largely driven by inattention, more so than by purposeful sharing of misinformation. Thus, interventions can successfully nudge social media users to focus more on accuracy. Crowdsourced veracity ratings can also be leveraged to improve social media ranking algorithms.

**9.1.13** Lorenz-Spreen, Oswald, Lewandowsky, & Hertwig (2022). A systematic review of worldwide causal and correlational evidence on digital media and democracy. Nature Human Behavior.

ABSTRACT: One of today's most controversial and consequential issues is whether the global uptake of digital media is causally related to a decline in democracy. We conducted a systematic review of causal and correlational evidence (N = 496 articles) on the link between digital media use and different political variables. Some associations, such as increasing political participation and information consumption, are likely to be beneficial for democracy and were often observed in autocracies and emerging democracies. **Other associations, such as declining political trust, increasing populism and growing polarization, are likely to be detrimental to democracy and were more pronounced in established democracies.** While the impact of digital media on political systems depends on the specific variable and system in question, several variables show clear directions of associations. The evidence calls for research efforts and vigilance by governments and civil societies to better understand, design and regulate the interplay of digital media and democracy.

[Note from Philipp Lorenz-Spreen: The excel list of all articles that we included in our review, with coded methods and outcome measures is publicly available here: https://osf.io/7ry4a/

[Additional quotes from the results section, summarizing findings on our key outcome variables:] [note that these quotes are from the preprint version, 2021]

**Trust.** Many articles in our sample found detrimental associations between digital media and various dimensions of trust (Fig. 2). For example, detrimental associations were found for trust in governments and politics [56, 57, 63, 75–79], trust in media [80], and social and institutional trust [81]. During the COVID-19 pandemic, digital media use was reported to be negatively associated with trust in vaccines [82, 83]. Yet the results about associations with trust are not entirely homogeneous. One multinational survey found beneficial associations with trust in science [84]; others found increasing trust in democracy with digital media use in Eastern and

Central European samples [85, 86]. Nevertheless, **the large majority of reported associations between digital media use and trust appear to be detrimental for democracy**.

**Polarization. Most articles found detrimental associations between digital media and different forms of political polarization** [110–114]. Our review found evidence for increasing out-group polarization on social media in a range of political contexts and on various platforms [115–118]. Increasing polarization was also linked to exposure to viewpoints opposed to one's own on social media feed [66, 119]. Articles comparing several political systems found associations that were country-dependent [120], again highlighting the importance of political context [121]. Nevertheless, increased digital use was for the most part linked to increased polarization overall, although there was some evidence for balanced online discourse without pronounced patterns of polarization [122–124], as well as evidence for potentially depolarizing association with social media [125]. **The body of causal articles largely supported the detrimental associations of digital media on polarization that we identified in correlational articles. Among established Western democracies, both social media use and overall internet use increased political polarization** [60, 67]. This was also the case in an experimental treatment that exposed users to opposing views on Twitter [66].

**Populism.** Articles on populism in our review examined either vote share and other popularity indicators for populist parties or the prevalence of populist messages and communication styles on digital media. Overall, articles using panel surveys, tracking data, and methods linking surveys to social media data **consistently found that increased digital media use was associated with increased populism. For example, digital platforms were observed to benefit populist parties more than they benefit established politicians** [127–130]. In a panel survey in Germany, a decline in trust that accompanied increasing digital media consumption was also linked to a turn towards the hard-right populist AfD party [77]. There is also evidence for an association between increased social media use and online right-wing radicalization in Austria, Sweden, and Australia [131–133].

FIGURES: [show that benefits were more frequent in less democratic countries; harms were more prevalent in the advanced democracies]

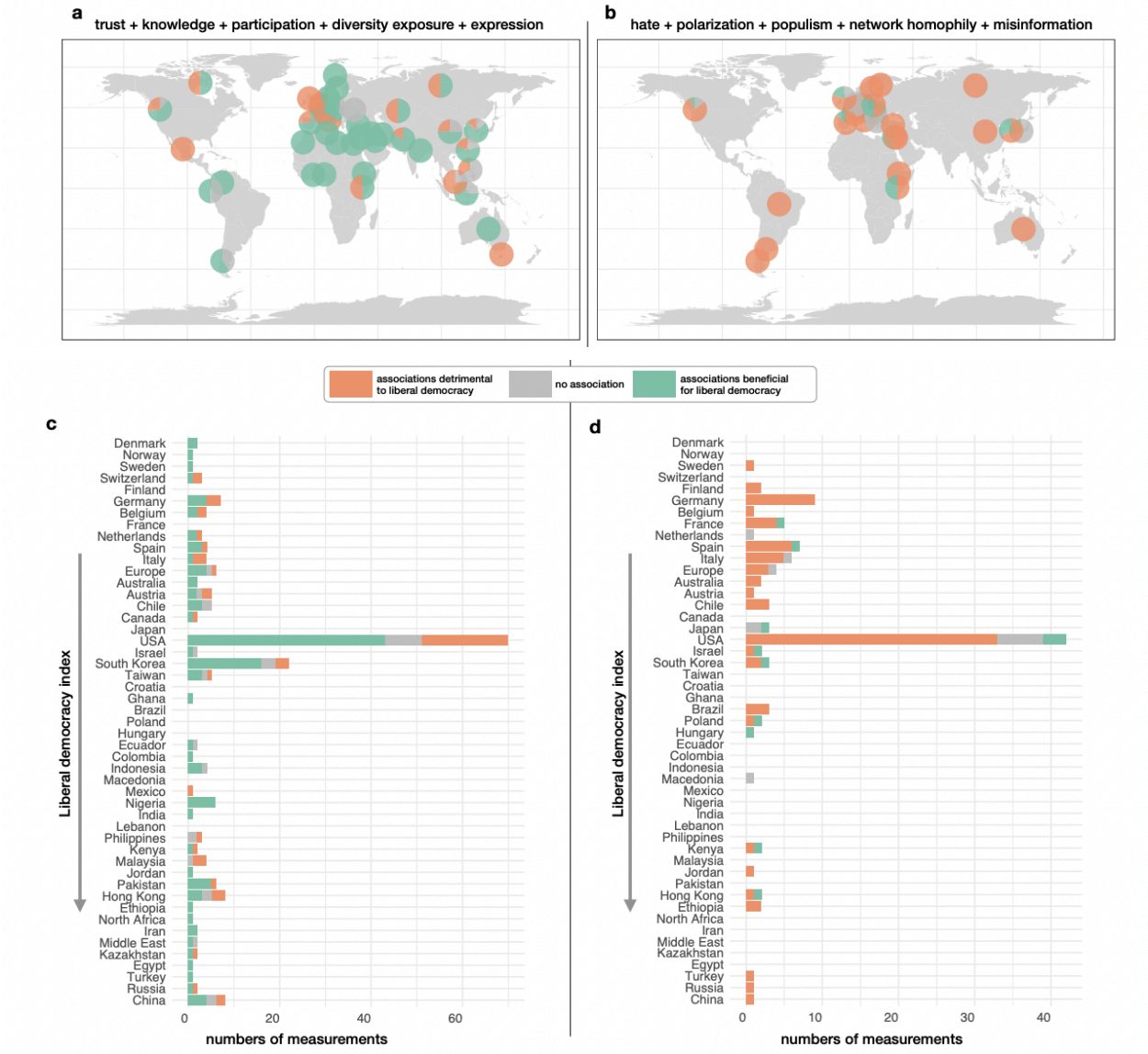



Figure 5: Geographical distribution of associations, split vertically between beneficial and detrimental outcomes. **a** Geographical distribution of reported associations for the variables trust, knowledge, participation, exposure, and expression. Pie charts show the composition of directions for each country studied. **b** Geographic representation of reported associations for the variables hate, polarization, populism, homophily, and misinformation. **c** Data and variables in **a**, in absolute numbers of reported associations and sorted along the Liberal Democracy Index [17]. **d** Data and variables in **b**, in absolute numbers of reported associations and sorted along the Liberal Democracy Index.

**9.1.14** Bavel, Rathje, Harris, Robertson, & Sternisko (2021). How social media shapes polarization. *Trends in Cognitive Sciences*.

ABSTRACT: This article reviews the empirical evidence on the relationship between social media and political polarization. We argue that social media shapes polarization through the following social, cognitive, and technological processes: partisan selection, message content, and platform design and algorithms.

FIGURE:

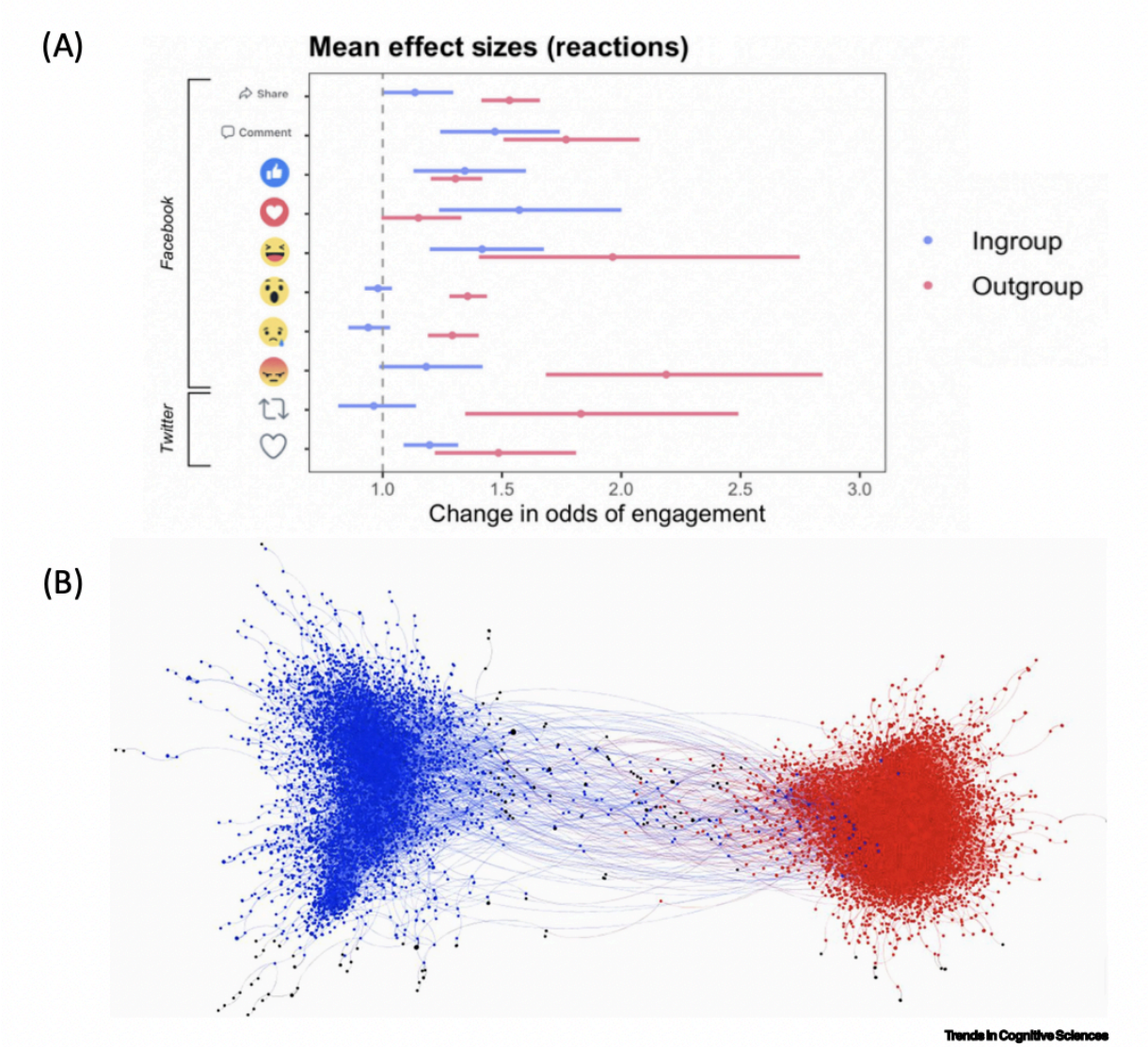



Figure 1. Partisan rhetoric on social media. (A) Shows the effects of ingroup and outgroup language on retweets, shares, and 'reactions' on Facebook (from [8]). (B) Shows the retweet networks of liberals (in blue) and conservatives (in red) when they include moral emotional language (from [12]).

*Figure 1. Partisan rhetoric on social media. (A) Shows the effects of ingroup and outgroup language on retweets, shares, and 'reactions' on Facebook. (B) Shows the*

*retweet networks of liberals (in blue) and conservatives (in red) when they include moral emotional language.*

**9.1.15** Arguedas, Robertson, & Nielsen (2022). Echo chambers, filter bubbles, and polarisation: A literature review. *Reuters Institute for the Study of Journalism.*

EXECUTIVE SUMMARY: Terms like echo chambers, filter bubbles, and polarisation are widely used in public and political debate but not in ways that are always aligned with, or based on, scientific work. And even among academic researchers, there is not always a clear consensus on exact definitions of these concepts.

In this literature review we examine, specifically, social science work presenting evidence concerning the existence, causes, and effect of online echo chambers and consider what related research can tell us about scientific discussions online and how they might shape public understanding of science and the role of science in society. Echo chambers, filter bubbles, and the relationship between news and media use and various forms of polarisation has to be understood in the context of increasingly digital, mobile, and platform-dominated media environments where most people spend a limited amount of time with news and many internet users do not regularly actively seek out online news, leading to significant inequalities in news use.

When defined as a bounded, enclosed media space that has the potential to both magnify the messages delivered within it and insulate them from rebuttal, studies in the UK estimate that between six and eight percent of the public inhabit politically partisan online news echo chambers.

**More generally, studies both in the UK and several other countries, including the highly polarised US, have found that most people have relatively diverse media diets, that those who rely on only one source typically converge on widely used sources with politically diverse audiences (such as commercial or public service broadcasters) and that only small minorities, often only a few percent, exclusively get news from partisan sources.**

Studies in the UK and several other countries show that the forms of **algorithmic selection offered by search engines, social media, and other digital platforms generally lead to slightly more diverse news use – the opposite of what the "filter bubble" hypothesis posits – but that self-selection, primarily among a small**

**minority of highly partisan individuals, can lead people to opt in to echo chambers, even as the vast majority do not.**

Research on polarisation offers a complex picture both in terms of overall developments and the main drivers and there is in many cases limited empirical work done outside the United States. Overall, ideological polarisation has, in the long run, declined in many countries but affective polarisation has in some, but not all, cases increased. News audience polarisation is much lower in most European countries, including the United Kingdom. Much depends on the specifics of individual countries and what point in time one measures change from and there are no universal patterns.

There is limited research outside the United States systematically examining the possible role of news and media use in contributing to various kinds of polarisation and the work done does not always find the same patterns as those identified in the US. In the specific context of the United States where there is more research, it seems that exposure to like-minded political content can potentially polarise people or strengthen the attitudes of people with existing partisan attitudes and that cross- cutting exposure can potentially do the same for political partisans.

Public discussions around science online may exhibit some of the same dynamics as those observed around politics and in news and media use broadly, but fundamentally there is at this stage limited empirical research on the possible existence, size, and drivers of echo chambers in public discussions around science. More broadly, existing research on science communication, mainly from the United States, documents the important role of self-selection, elite cues, and small, highly active communities with strong views in shaping these debates and highlights the role especially political elites play in shaping both news coverage and public opinion on these issues.

In summary, the work reviewed here suggests **echo chambers are much less widespread than is commonly assumed, finds no support for the filter bubble hypothesis and offers a very mixed picture on polarisation and the role of news and media use in contributing to polarisation.**

**9.1.16** Serrano, Carlos Medina, Hegelich, Shahrezaye, & Papakyriakopoulos (2018). Social media report: The 2017 German federal elections.

EXCERPT: **The first finding is that the AfD dominated in social media. On both Twitter and Facebook, the right-wing political party managed to spread their**

**message to more users. There is a possibility that part of their success in the 2017 elections relates to these results.** Already in 2016, Schelter et al. [ 20 ] formulated that "**the rise of the AfD can be associated with an amount of social media coverage and user engagement that is unprecedented in the German political landscape**".

The second finding is that online manipulation mechanisms existed that targeted the German election process on Twitter. Nevertheless, the observed amount was less than expected by experts. It is difficult to measure the effects that the detected social bots, fake news stories and foreign intervention techniques had on the German public. However, the results are consistent with Neudert et al. [ 16 ], which also found that the bots were working in favor of the AfD and with Saengerlaub et al. [19 ], who presented an analysis on fake news in Germany.

The third finding is that the German public is less prone to being affected by online misinformation than the US public. The closeness of right- and left-wing media in Germany to the mainstream media shows that citizens of different political parties are consuming information from validated sources. We further conclude that false news did not play a major role in the conversation regarding the election. The top shared news on Facebook and Twitter connected to political parties had only a few misleading stories and no completely fabricated news. The news items related to migration were those that had the most misleading facts.

**9.1.17** German National Academy of Sciences (2021). Digitalisation and democracy. [h/t Tobias Dienlen]

SUMMARY AND RECOMMENDATION [first 2 paras of many]: In the course of digitalisation, the democratic public sphere has already changed fundamentally. Alongside traditional media such as press and broadcast media, new digital forms of communication such as online media and social networks have emerged. With respect to their democratisation potential, these have given rise to great expectations, but they also facilitate critical developments.This development has enabled easier access to information for the general public as well as greater opportunities for political participation and to strengthen civil society. However, it has also resulted in an increase in misinformation, attempts to manipulate and hate speech.

In order to properly understand the relationship between digitalisation and democratic public spheres, four aspects need to be considered: (a) the digitalisation of

infrastructures of democratic public spheres, (b) changes in information and communication effectuated by digital media, (c) the increase in democratic participation due to new, digital formats and (d) the shift in political self determination.

**9.1.18** Iandoli, Primario, & Zollo (2021). The impact of group polarization on the quality of online debate in social media: A systematic literature review. *Technological Forecasting and Social Change*. [h/t Olivia Fischer]

ABSTRACT: Social media are often accused of worsening the quality of online debate. In this paper, we focus on group polarization in the context of social media-enabled interaction, a dysfunctional group dynamic by which participants become more extreme in their initial position on an issue. Through a systematic literature review, we identified a corpus of 121 research papers investigating polarization in social media and other online conversational platforms and reviewed the main empirical findings, as well as theoretical and methodological approaches. We use this knowledge base to assess some recurrent accusations against social media in terms of their supposed tendency to worsen online debate. Our analysis shows that, **while some concerns have been exaggerated, social media do contribute to increase polarization either by amplifying and escalating social processes that also occur offline or in specific ways enabled by their design affordances, which also make these platforms prone to manipulation.** We argue against suggestions aimed at reducing freedom of speech in cyberspace and identify in inadequate regulation and lack of ethical design as the leading causes of social media-enabled group dysfunctions, highlighting research areas that can support the creation of higher quality online discursive spaces.

**9.1.19** Newman, Fletcher, Kalogeropolous, Levy, & Nielsen (2017). Reuters Institute digital news report 2017. *Reuters Institute for the Study of Journalism.*

EXECUTIVE SUMMARY:

- **The internet and social media may have exacerbated low trust and 'fake news', but we find that in many countries the underlying drivers of mistrust are as much to do with deep-rooted political polarisation and perceived mainstream media bias.**
- **Echo chambers and filter bubbles are undoubtedly real for some, but we also find that – on average – users of social media, aggregators, and search engines experience more diversity than non-users.**

- Though the economic outlook for most media companies remains extremely difficult, not all the indicators are getting worse. The growth of ad-blocking has stopped while online subscriptions and donations are picking up in some countries. Our focus groups provide some encouragement that more might be prepared to pay in the future if content is sufficiently valuable, convenient, and relevant.

With data covering more than 30 countries and five continents, this research is a reminder that the digital revolution is full of contradictions and exceptions. Countries started in different places, and are not moving at the same pace. These differences are captured in individual country pages that can be found towards the end of this report. They contain critical industry context written by experts as well as key charts and data points. The overall story around the key trends is captured in this executive summary with additional analysis on some subject areas in a separate section.

ADDITIONAL FINDINGS:

- Only a quarter (24%) of our respondents think social media do a good job in separating fact from fiction, compared to 40% for the news media. Our qualitative data suggest that users feel the combination of a lack of rules and viral algorithms are encouraging low quality and 'fake news' to spread quickly.
- There are wide variations in trust across our 36 countries. The proportion that says they trust the news is highest in Finland (62%), but lowest in Greece and South Korea (23%).
- In most countries, we find a strong connection between distrust in the media and perceived political bias. This is particularly true in countries with high levels of political polarisation like the United States, Italy, and Hungary.
- Almost a third of our sample (29%) say they often or sometimes avoid the news. For many, this is because it can have a negative effect on mood. For others, it is because they can't rely on news to be true.

**9.1.20** Yesilada & Lewandowsky (2020). Systematic review: YouTube recommendations and problematic content. *EconStor.*

ABSTRACT: There has been much concern that social media, in particular YouTube, may facilitate radicalisation and polarisation of online audiences. This systematic review aimed to determine whether the YouTube recommender system facilitates pathways to problematic content such as extremist or radicalising material. The review conducted a narrative synthesis of the papers in this area. It assessed the eligibility of 1,187 studies

and excluded studies using the PRISMA process for systematic reviews, leaving a final sample of 23 studies. Overall, **14 studies implicated the YouTube recommender system in facilitating problematic content pathways, seven produced mixed results, and two did not implicate the recommender system.** The review's findings indicate that the YouTube recommender system could lead users to problematic content. However, due to limited access and an incomplete understanding of the YouTube recommender system, the models built by researchers might not reflect the actual mechanisms underlying the YouTube recommender system and pathways to problematic content.

**9.1.21** González-Bailón & Lelkes (2022). Do social media undermine social cohesion? A critical review. *Social Issues and Policy Review*.

ABSTRACT: We evaluate the empirical evidence interrogating the question of whether social media erodes social cohesion. We look at how networks, information exchange, and norms operate on these platforms. We also evaluate the conditions under which social media can be conducive to forming social capital and encouraging prosocial behavior. We discuss the psychological mechanisms that operate at the individual level and assess whether social media can create the environment and incentives to sustain cooperation and constructive exchange. Our discussion of the literature centers on how attitudes, perceptions, and beliefs are formed during the type of online interactions encouraged by platforms, their design, and affordances. We consider the policy implications of existing research, focusing on how empirical studies may inform regulatory efforts and platform interventions.

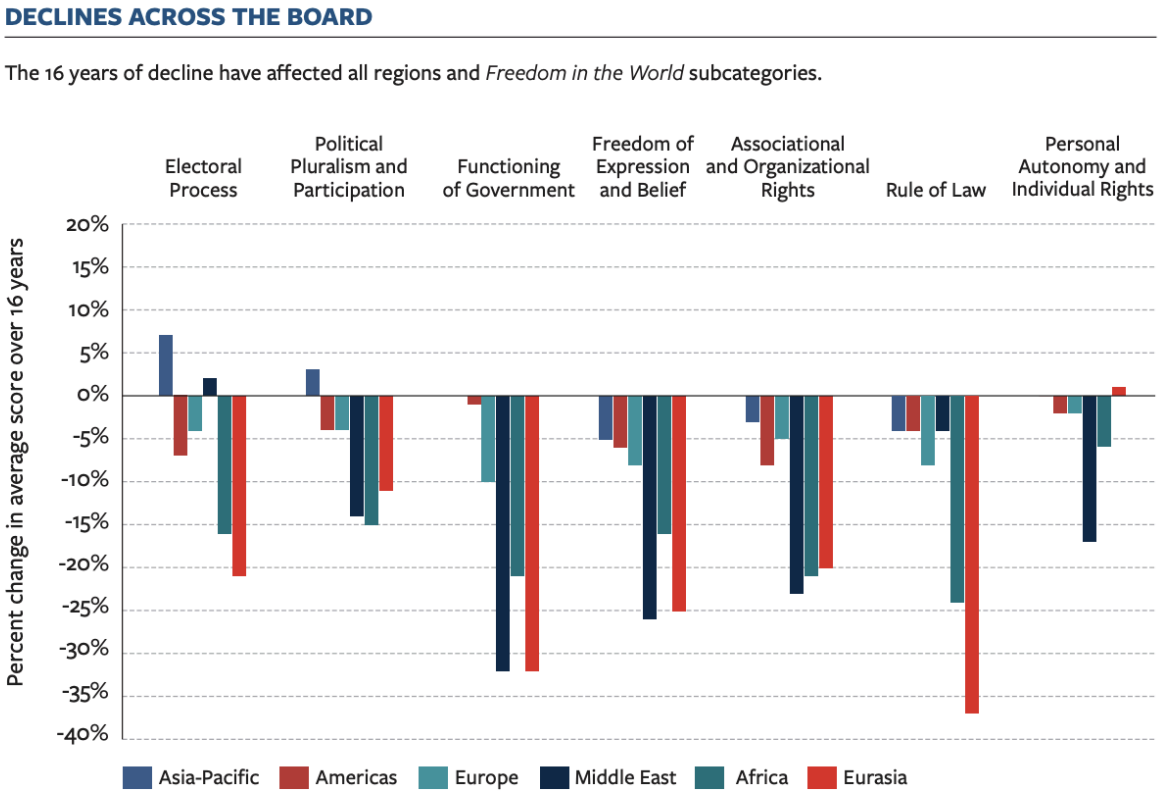
**9.1.22** Repucci & Slipowitz (2022). *The Global Expansion of Authoritarian Rule* (Freedom In The World 2022). Freedom House.

EXCERPT: Global freedom faces a dire threat. Around the world, the enemies of liberal democracy—a form of self-government in which human rights are recognized and every individual is entitled to equal treatment under law—are accelerating their attacks. Authoritarian regimes have become more effective at co-opting or circumventing the norms and institutions meant to support basic liberties, and at providing aid to others who wish to do the same. In countries with long-established democracies, internal forces have exploited the shortcomings in their systems, distorting national politics to promote hatred, violence, and unbridled power. Those countries that have struggled in the space between democracy and authoritarianism, meanwhile, are increasingly tilting

toward the latter. The global order is nearing a tipping point, and if democracy's defenders do not work together to help guarantee freedom for all people, the authoritarian model will prevail.

**The present threat to democracy is the product of 16 consecutive years of decline in global freedom. A total of 60 countries suffered declines over the past year, while only 25 improved. As of today, some 38 percent of the global population live in Not Free countries, the highest proportion since 1997. Only about 20 percent now live in Free countries.**

FIGURE:




**9.1.23** Little, & Meng (Working Paper). *Subjective and Objective Measurement of Democratic Backsliding* (SSRN Scholarly Paper No. 4327307).

ABSTRACT: Despite the general narrative that we are in a period of global democratic decline, there have been surprisingly few empirical studies to assess whether this is

systematically true. Most existing studies of backsliding rely heavily, if not entirely, on subjective indicators which rely on expert coder judgement. We survey other more objective indicators of democracy (such as incumbent performance in elections), and find little evidence of global democratic decline over the last decade. To explain the discrepancy between trends in subjective and objective indicators, we develop formal models that consider the role of coder bias and leaders strategically using more subtle undemocratic action. The simplest explanation is that recent declines in average democracy scores are driven by changes in coder bias. **While we cannot rule out the possibility that the world is experiencing major democratic backsliding almost exclusively in ways which require subjective judgement to detect, this claim is not justified by existing evidence.**

[Others? What have we missed?]

* * * * * * * * * * * * * * *

# 10. BOOKS BY SCHOLARS

In this section we include books by scholars that draw on empirical research to offer analysis of the effects of social media, or suggestions for improvements. We do not include books on polarized politics in general, or on social media in general -- there are just too many! We focus on books that bear directly on the 7 empirical questions that structure this review.

**10.1** Settle (2018). Frenemies: How social media polarizes America. *Cambridge: Cambridge University Press.*

DESCRIPTION: Why do Americans have such animosity for people who identify with the opposing political party? Jaime E. Settle argues that in the context of increasing partisan polarization among American political elites, the way we communicate on Facebook uniquely facilitates psychological polarization among the American public. Frenemies introduces the END Framework of social media interaction. END refers to a subset of content that circulates in a social media ecosystem: a personalized, quantified blend of politically informative 'expression', 'news', and 'discussion' seamlessly interwoven into a wider variety of socially informative content. Scrolling through the News Feed triggers a cascade of processes that result in negative attitudes about those

who disagree with us politically. The inherent features of Facebook, paired with the norms of how people use the site, heighten awareness of political identity, bias the inferences people make about others' political views, and foster stereotyped evaluations of the political out-group.

**10.2** Bail (2021). Breaking the social media prism: How to make our platforms less polarizing. *Princeton University Press.*

OVERVIEW: In an era of increasing social isolation, platforms like Facebook and Twitter are among the most important tools we have to understand each other. We use social media as a mirror to decipher our place in society but, as Chris Bail explains, it functions more like a prism that distorts our identities, empowers status-seeking extremists, and renders moderates all but invisible. Breaking the Social Media Prism challenges common myths about echo chambers, foreign misinformation campaigns, and radicalizing algorithms, revealing that the solution to political tribalism lies deep inside ourselves.

Drawing on innovative online experiments and in-depth interviews with social media users from across the political spectrum, this book explains why stepping outside of our echo chambers can make us more polarized, not less. Bail takes you inside the minds of online extremists through vivid narratives that trace their lives on the platforms and off—detailing how they dominate public discourse at the expense of the moderate majority. Wherever you stand on the spectrum of user behavior and political opinion, he offers fresh solutions to counter political tribalism from the bottom up and the top down. He introduces new apps and bots to help readers avoid misperceptions and engage in better conversations with the other side. Finally, he explores what the virtual public square might look like if we could hit “reset” and redesign social media from scratch through a first-of-its-kind experiment on a new social media platform built for scientific research.

Providing data-driven recommendations for strengthening our social media connections, Breaking the Social Media Prism shows how to combat online polarization without deleting our accounts.

**10.3.** Persily,& Tucker (Eds.) (2020). *Social Media and Democracy: The State of the Field, Prospects for Reform*. Cambridge University Press. *[Open Source]*

SUMMARY: The goal of this book is to synthesize the existing research on social media and democracy. We present reviews of the literature on disinformation, polarization, echo chambers, hate speech, bots, political advertising, and new media. In addition, we canvass the literature on reform proposals to address the widely perceived threats to democracy. We seek to examine the current state of knowledge on social media and democracy, to identify the many knowledge gaps and obstacles to research in this area, and to chart a course for future research. We hope to advocate for this new field of study and to suggest that universities, foundations, private firms, and governments should commit to funding and supporting this research.

| | | |
|---|---|---|
| 12 | Democratic Transparency in the Platform Society | Gorwa & Ash |
| 13 | Conclusion: The Challenges and Opportunities for Social Media Research | Persiley & Tucker |

**10.4** Aral (2020). *The Hype Machine.* Penguin Random House.

ABSTRACT: Social media connected the world—and gave rise to fake news and increasing polarization. It is paramount, MIT professor Sinan Aral says, that we recognize the outsize effect social media has on us—on our politics, our economy, and even our personal health—in order to steer today's social technology toward its great promise while avoiding the ways it can pull us apart.

Drawing on decades of his own research and business experience, Aral goes under the hood of the most powerful social networks to tackle the critical question of just how much social media actually shapes our choices, for better or worse. He shows how the tech behind social media offers the same set of behavior influencing levers to everyone who hopes to change the way we think and act—from Russian hackers to brand marketers—which is why its consequences affect everything from elections to business, dating to health. Along the way, he covers a wide array of topics, including how network effects fuel Twitter's and Facebook's massive growth, the neuroscience of how social media affects our brains, the real consequences of fake news, the power of social ratings, and the impact of social media on our kids.

In mapping out strategies for being more thoughtful consumers of social media, *The Hype Machine* offers the definitive guide to understanding and harnessing for good the technology that has redefined our world overnight.

**10.5** Benkler, Faris, & Roberts (2018). *Network propaganda: Manipulation, disinformation, and radicalization in American politics.* Oxford University Press. [h/t Steve Feldstein]

ABSTRACT: This book examines the shape, composition, and practices of the United States political media landscape. It explores the roots of the current epistemic crisis in political communication with a focus on the remarkable 2016 U.S. president election culminating in the victory of Donald Trump and the first year of his presidency. The authors present a detailed map of the American political media landscape based on the

analysis of millions of stories and social media posts, revealing a highly polarized and asymmetric media ecosystem. Detailed case studies track the emergence and propagation of disinformation in the American public sphere that took advantage of structural weaknesses in the media institutions across the political spectrum. This book describes how the conservative faction led by Steve Bannon and funded by Robert Mercer was able to inject opposition research into the mainstream media agenda that left an unsubstantiated but indelible stain of corruption on the Clinton campaign. The authors also document how Fox News deflects negative coverage of President Trump and has promoted a series of exaggerated and fabricated counter narratives to defend the president against the damaging news coming out of the Mueller investigation. Based on an analysis of the actors that sought to influence political public discourse, **this book argues that the current problems of media and democracy are not the result of Russian interference, behavioral microtargeting and algorithms on social media, political clickbait, hackers, sockpuppets, or trolls, but of asymmetric media structures decades in the making. The crisis is political, not technological.**

**10.6** Bruns (2019). *Are filter bubbles real?* Wiley.

ABSTRACT: There has been much concern over the impact of partisan echo chambers and filter bubbles on public debate. Is this concern justified, or is it distracting us from more serious issues?

**Axel Bruns argues that the influence of echo chambers and filter bubbles has been severely overstated, and results from a broader moral panic about the role of online and social media in society. Our focus on these concepts, and the widespread tendency to blame platforms and their algorithms for political disruptions, obscure far more serious issues pertaining to the rise of populism and hyperpolarisation in democracies.**

Evaluating the evidence for and against echo chambers and filter bubbles, Bruns offers a persuasive argument for why we should shift our focus to more important problems. This timely book is essential reading for students and scholars, as well as anyone concerned about challenges to public debate and the democratic process.

**10.7** Philips & Milner (2021). *You Are Here: A Field Guide for Navigating Polarized Speech, Conspiracy Theories, and Our Polluted Media Landscape.* The MIT Press. [h/t Shane Creevy]

SUMMARY: How to understand a media environment in crisis, and how to make things better by approaching information ecologically.

Our media environment is in crisis. Polarization is rampant. Polluted information floods social media. Even our best efforts to help clean up can backfire, sending toxins roaring across the landscape. In You Are Here, Whitney Phillips and Ryan Milner offer strategies for navigating increasingly treacherous information flows. **Using ecological metaphors, they emphasize how our individual me is entwined within a much larger we, and how everyone fits within an ever-shifting network map.**

Phillips and Milner describe how our poisoned media landscape came into being, beginning with the Satanic Panics of the 1980s and 1990s—which, they say, exemplify "network climate change"—and proceeding through the emergence of trolling culture and the rise of the reactionary far right (as well as its amplification by journalists) during and after the 2016 election. They explore the history of conspiracy theories in the United States, focusing on those concerning the Deep State; e**xplain why old media literacy solutions fail to solve new media literacy problems; and suggest how we can navigate the network crisis more thoughtfully, effectively, and ethically. We need a network ethics that looks beyond the messages and the messengers to investigate toxic information's downstream effects.**

**10.8** Bennett, & Livingston (2020). *The Disinformation Age: Politics, Technology, and Disruptive Communication in the United States*. Cambridge University Press. A PDF version of the book is available for free. [h/t Steve Feldstein]

SUMMARY: The intentional spread of falsehoods – and attendant attacks on minorities, press freedoms, and the rule of law – challenge the basic norms and values upon which institutional legitimacy and political stability depend. How did we get here? The Disinformation Age assembles a remarkable group of historians, political scientists, and communication scholars to examine the historical and political origins of the post-fact information era, focusing on the United States but with lessons for other democracies. Bennett and Livingston frame the book by examining decades-long efforts by political and business interests to undermine authoritative institutions, including parties, elections, public agencies, science, independent journalism, and civil society groups. The other distinguished scholars explore the historical origins and workings of disinformation, along with policy challenges and the role of the legacy press in improving public communication.

**10.9** Forestal (2022). *Designing for Democracy: How to Build Community in Digital Environments*. Oxford University Press. [h/t Jen Forestal]

SUMMARY: How should we "fix" digital technologies to support democracy instead of undermining it? In Designing for Democracy, Jennifer Forestal argues that accurately evaluating the democratic potential of digital spaces means studying how the built environment--a primary component of our "modern public square"--structures our activity, shapes our attitudes, and supports the kinds of relationships and behaviors democracy requires.

Drawing from a wide range of disciplines, including architecture, psychology, and the history of political thought, she argues that "democratic spaces" must be designed with three environmental characteristics--boundaries, durability, and flexibility--that, taken together, afford users the ability to engage in fundamental civic practices.

Through extended analyses of Facebook, Twitter, and Reddit, Forestal shows precisely how well these digital platforms meet the criteria for democratic spaces, or whether they do so at all. The result is a more nuanced analysis of the democratic communities that form--or fail to emerge--in these spaces, as well as more concrete suggestions for how to improve them. In connecting the built environment, digital technologies, and democratic theory, *Designing for Democracy* provides blueprints for democracy in a digital age.

**10.10** Vaidhyanathan (2018). *Antisocial media: How Facebook disconnects us and undermines democracy.* Oxford University Press.

BOOK SUMMARY: If you wanted to build a machine that would distribute propaganda to millions of people, distract them from important issues, energize hatred and bigotry, erode social trust, undermine respectable journalism, foster doubts about science, and engage in massive surveillance all at once, you would make something a lot like Facebook. Of course, none of that was part of the plan. In this fully updated paperback edition of Antisocial Media, including a new chapter on the increasing recognition of--and reaction against--Facebook's power in the last couple of years, Siva Vaidhyanathan explains how Facebook devolved from an innocent social site hacked together by Harvard students into a force that, while it may make personal life just a little more pleasurable, makes democracy a lot more challenging. It's an account of the

hubris of good intentions, a missionary spirit, and an ideology that sees computer code as the universal solvent for all human problems. And it's an indictment of how "social media" has fostered the deterioration of democratic culture around the world, from facilitating Russian meddling in support of Trump's election to the exploitation of the platform by murderous authoritarians in Burma and the Philippines. Both authoritative and trenchant, Antisocial Media shows how Facebook's mission went so wrong.

**10.11** Gershberg & Illing (2022). The Paradox of Democracy: Free Speech, Open Media, and Perilous Persuasion.

SUMMARY: All over the world, from India to Hungary to Turkey to Brazil to the United States, democratic cultures have been disordered. What we're witnessing is a convergence of various forces unleashed by novel media and populist rhetorical styles that implode democracy from within.

**10.12** Bartlett (2018). The People vs Tech: How the Internet is Killing Democracy (and How We Save It). *New York: Penguin.*

SUMMARY: The internet was meant to set us free. But have we unwittingly handed too much away to shadowy powers behind a wall of code, all manipulated by a handful of Silicon Valley utopians, ad men, and venture capitalists? And, in light of recent data breach scandals around companies like Facebook and Cambridge Analytica, what does that mean for democracy, our delicately balanced system of government that was created long before big data, total information, and artificial intelligence? In this urgent polemic, Jamie Bartlett argues that through our unquestioning embrace of big tech, the building blocks of democracy are slowly being removed. The middle class is being eroded, sovereign authority and civil society is weakened, and we citizens are losing our critical faculties, maybe even our free will.

The People Vs Tech is an enthralling account of how our fragile political system is being threatened by the digital revolution. Bartlett explains that by upholding six key pillars of democracy, we can save it before it is too late. We need to become active citizens, uphold a shared democratic culture, protect free elections, promote equality, safeguard competitive and civic freedoms, and trust in a sovereign authority. This essential book shows that the stakes couldn't be higher and that, unless we radically alter our course, democracy will join feudalism, supreme monarchies and communism as just another political experiment that quietly disappeared.

[Other books? What have we missed?]

* * * * * * * * * * * * * * *

# 11. PROPOSALS FOR IMPROVING SOCIAL MEDIA

[New section, very incomplete; currently being populated in summer 2022. Once it grows it will be moved to its own Collaborative Review doc, curated by the Center for Humane Technology]

## 11.1 On the need for and legitimacy of federal regulation

**11.1.1** Jones & Samples (forthcoming 2022). On the Systemic Importance of Digital Platforms. *University of Pennsylvania Journal of Business Law.* (h/t Tim Samples)

CONDENSED ABSTRACT FROM TIM SAMPLES: Proposes a theoretical basis for imposing a prudential regulatory regime for digital platforms based on their systemic importance, drawing parallels with the framework for systemically important financial institutions (SIFIs) in the Dodd-Frank Act.

**11.1.2** Werbach & Zaring (forthcoming 2022). Systemically Important Technology. *Texas Law Review*. (h/t Tim Samples)

CONDENSED ABSTRACT FROM TIM SAMPLES: This article addresses the risks of failure within the connective tissue of systemically important network institutions.

**11.1.3** Griffin (forthcoming 2021). Systemically Important Platforms, *Cornell Law Review*. (h/t Tim Samples)

CONDENSED ABSTRACT FROM TIM SAMPLES: This article proposes a special designation for systemically important platforms centered on their use of manipulative technologies.

**11.1.4** Öhman & Aggarwal (2020). What if Facebook Goes Down? Ethical and Legal Considerations for the Demise of Big Tech. *Internet Policy Review*.

CONDENSED ABSTRACT FROM TIM SAMPLES: This article explores the failure risks of Facebook, coins the term systemically important technological institutions (SITIs), and proposes more research in that area.

## 11.2 User Authentication

One of the main reasons that social media platforms are toxic to democracy is that they are a gift to trolls, Russian intelligence agents, political operatives, swindlers, and anyone else acting in bad faith who can create one or thousands of accounts. Many reform proposals (including those from Elon Musk, Jonathan Haidt, Jamie Dimon, …) talk about the benefits of requiring some form of user authentication. But what does that mean? First, it is crucial to note that authentication does NOT mean that people must post using their real names. Rather, under most authentication schemes anyone can still open an account, instantly, on platforms such as Facebook or Twitter, with a pseudonym and no authentication, if they simply want to view the posts of others. But then, as a second step, for those who want to post their own content and gain algorithmic amplification to a potentially vast audience, users would be required to take a subsequent step of authentication, likely carried out by a 3rd party company or non-profit. There are (at least) three levels of authentication.

> **Level 0 = No authentication**. This is what we have now. Any person or automated system can create unlimited fake accounts every day.
>
> **Level 1 = authenticate humans**:  users must pass a captcha, to show that they are a human and not a bot. But each human could still create and run hundreds of troll accounts, or create them and turn them over to AI to run.
>
> **Level 2 = authenticate unique identity once and untraceably**. This would be carried out by a non-profit or for-profit company, using a variety of methods. A user at Facebook (for example) who wants to be able to post would get sent over to this third party. Any methods that require showing a government ID, or giving biometric information, would then wipe out the information after authentication, when sending back the approval to the platform requesting authentication. These schemes allow each person to create only one account.

Examples of companies or non-profits who are developing such schemes:

- Human-id.org
- World Coin
- Proofofexistence.xyz

**Level 3 = authenticate identity to a 3rd party, who keeps the information**.

- A company like Clear is well situated to do this, as it already does for air travel, sporting events, and many other situations where there is a need for security balanced with privacy.
- India's Aadhar platform that authenticates people in real-time. Aadhar stores encrypted biometric data. Aadhar is maintained by "The Unique Identification Authority of India (UIDAI).

***Question: What about protecting dissidents in repressive countries?***
**Answer:** Why does the whole world need to be on a single platform? That was a dream ten years ago, but now it appears that we might need one kind of platform optimized for the "public square" of advanced or stable democracies, with incentives for constructive dialogue, and a very different set of platforms designed for life in the more dangerous "public square" of authoritarian countries, where the design imperative is for untraceability and protection of dissidents. It would be trivially easy to connect the two platforms: journalists or human rights organizations on the democratic platforms can simply re-post content from dissidents and whistleblowers on the high security platforms, without even knowing their real identities.

***Question: What about whistle blowers or political groups who want a second account?* Is everyone limited to one authenticated account?**
**Answer:** There would be provisions for accounts beyond the regular single-person accounts. Companies and non-profit organizations would certainly have accounts, and there would be provisions for authenticating them. Whistle blowers would still have hundreds of ways to get news out to the world, anonymously, via blogs, journalists, anonymous hotlines, and non-profit accounts that could be set up for the purpose. It's not clear why critics and whistleblowers must each have their own individual un-authenticated Twitter or Instagram account to be effective.

**To learn more about user authentication**

- See this essay by Scott Galloway, on the necessity of identification in the online world
- Listen to this episode of Brave New World, a conversation between Vasant Dhar and Jonathan Haidt. (Discussion of KYC is towards the end of the episode).
- Tom Newton Dunn: We must bite the bullet on online anonymity to defeat the trolls (*Evening Standard)*.

## 11.3 Age Restrictions and Age Appropriate Design

First, read the history of How 13 Became the Internet's Age of Adulthood, back in 1998. It was supposed to be 16, but lobbyists for e-commerce companies got it lowered. There was no consideration of mental health; this was about when children can sign contracts with companies to give away their data and their rights, without any parental permission. 25 years later, the internet is very different and studies show that young teens (11-15) are the most badly harmed by spending time on social media. The age should be raised, but how to enforce it, rather than relying on the honor system as we do today? Jon Haidt suggests that companies that need to enforce a minimum age should be required to offer a menu of methods by which customers could prove that they were old enough, rapidly and reliably. One option can be posing for a selfie with one's drivers license or other government-issued ID, as some companies do now, but there are so many other ways, for people who do not want to share their ID, or even their real name, with the platform. For example:

- There are already many companies devoted to checking the age of potential customers, rapidly and conveniently. There are so many of them now that they have their own trade association: The Age Verification Providers Association. Examples include AgeChecker.net, or Yoti.
- Clear (which you know from airports) already handles age verification rapidly and conveniently, e.g, for customers who want to buy beer at sporting events.
- See multiple proposals here: Chris Griswold (2022) Protecting Children from Social Media. National Affairs. E.g.: "One possibility would be for the SSA [Social Security Administration] to offer a service through which an American could type his Social Security number into a secure federal website and receive a temporary, anonymized code via email or text, like the dual-authentication methods already in widespread use. Providing this code to an online platform could allow it to confirm instantly with the SSA whether the user exceeds a certain age without further personal data reaching the platform or the government."
- See Yuval Levin's NYT essay: How Changing One Law Could Protect Kids From Social Media.
- Facebook developing AI, new ways to detect users under age 13.
- See the UK Age appropriate design code. See also Age Verification: State of Play and Key Developments in the EU and UK

- Meta is testing a new age verification system, offering users three ways to prove they are the age they say they are. BUT: it seems that they only do this if a user tries to change her age to make herself older. If users lie about their age when they create the account, they are OK.

## 11.4 Platform accountability and transparency

- Platform Accountability and Transparency Act sets up a system where independent researchers submit research proposals to the NSF to be approved to access platform data
- Algorithmic Accountability Act of 2022 enhances capacity of FTC to oversee and provide guidelines for private sector to assess impact of algorithms

## 11.5 Architectural changes to reduce virality

- Frances Haugen on limiting the number of people one can invite to join a Facebook group in a given week
- Modify the share button on Facebook as discussed on Frances Haugen Your Undivided Attention and #OneClickSafer

## 11.6 Changing incentives to reduce trolling and antisocial behavior

- Social media platforms have essentially become the public squares of democracy, yet they are overrun with bots, fake accounts, trolls, and normal people who respond to incentives to be nasty. This creates public squares where most citizens do not want to participate, and where there is little real dialogue. In a real public square, people who assault others would be removed. People who yell and scream and never listen to others would be shunned. Social norms would incentivize some degree of civility. Is there any way to make platforms such as Twitter become public squares in which social norms encourage productive conversation, rather than aggression?
- A first step should be user authentication (see section 11.2), which would greatly reduce the number and reach of anonymous trolls, although some people are trolls using their real names.
- An additional step to reduce antisocial behavior is to evaluate every user across all of their posts on a variable we might call “trollishness” or “toxicity.” Suppose a platform used at least three methods for evaluating its users, to allow cross-checking and reduce efforts to manipulate ratings: AI, ratings or reports from other users, and human ratings by platform staff. Next, suppose that a platform allowed all users to move a slider switch on a trollishness filter, which made the X% most trollish users disappear from the users feed, while at the same time making the user invisible to the most trollish. Suppose that

by default the filter was set to 1%, to remove the most trollish 1%, but users could choose to set it to 0 (to remove nobody) or to some higher number, perhaps as high as 20%. Instantly, the incentive structure of the platform would change profoundly. Nasty behavior that used to pay off handsomely will now backfire, leading to a reduction in one's audience, one's reach. This is not censorship: anyone can still say anything. This is more like the real world in which being a complete jerk leads to less reach, not more.

## 11.7 Changing parameters to reduce the noise/signal ratio

- Ellen Goodman (2020). Digital Information Fidelity and Friction: Crafting a systems-level approach to transparency

## 11.8 Miscellaneous additional reforms

- Offer users "Attention Settings," so that they can opt out of persuasive design tricks, such as autoplay, like counts, and suggested content. From Welf von Horen, who writes about The Liberation of Human Attention. See also The Humane Tech Library, a co-curated collection of designs and resources aimed at protecting human attention

* * * * * * * * * * * * * * *

# 12. CONCLUSION

[To come. In September, after receiving critiques and additional studies from other researchers, we'll summarize what we believe the academic literature says in response to the 7 questions]

* * * * * * * * * * * * * * *

# APPENDICES

## APPENDIX A: TIMELINE OF PLATFORM CHANGES

Drawing dates from Wikipedia: Facebook timeline, Twitter timeline, Youtube timeline, Instagram timeline, Reddit timeline, Tumblr history, Gab, Discord, Parler, Twitch, & Pinterest, Truth Social. See also this on FB's newsfeed algorithms

*Abbreviations: APL = Apple; FB = Facebook; TW = Twitter; IG = Instagram; YT = YouTube; Snap = Snapchat; TikTok; Reddit; Twitch; Gab; Parler; Pin = Pinterest; Truth = Truth Social, O-AI = OpenAI*

| YEAR | PLATFORM AND CHANGE |
|---|---|
| 2003 | MySpace and LinkedIn founded |
| 2004 | **FB:** Founded |
| 2005 | **YT:** Founded<br>**Reddit:** Founded |
| 2006 | **FB:** Launches news feed; Opens membership to anyone<br>**TW:** Founded |
| 2007 | **Tumblr:** Founded |
| 2008 | **Reddit:** Users can create custom reddits (or subreddits)<br>**Pin:** Founded |
| 2009 | **FB:** Adds like button and share button. Re-orders feed based on popularity, rather than reverse-chronological order<br>**TW:** Adds like button and retweet<br>**APL:** Launches push notifications. |
| 2010 | **FB:** Adds option to like individual comments; launches redesign that emphasizes bio, photos, education, and relationships<br>**TW:** Announces that it will start allowing for advertising in the form of *promoted tweets*<br>**APL:** iPhone 4 released, with front-facing camera, for selfies<br>**IG:** Launches |
| 2011 | **Snap:** Launches<br>**Twitch:** Founded |

| | |
|---|---|
| | **FB:** Launches Messenger<br>**TW:** Overhauls its website to feature the "Fly" design, which the service says is easier for new users to follow and promotes advertising. In addition to the *Home* tab, the *Connect* and *Discover* tabs are introduced along with a redesigned profile and timeline of Tweets<br>**IG:** Introduces 'filters', allowing users to easily alter their photos |
| 2012 | **FB:** Starts showing advertisements in news feed ("featured posts"); acquires Instagram. Goes public.<br>**YT:** Launched their new interface and altered the platform's algorithm from a view-based to a watch time-based system. |
| 2013 | **IG:** Introduces sponsored post advertising targeting US users<br>**FB:** Introduces threaded comments (anyone can "reply" to a comment, which facilitates multi-round arguments under other people's posts) |
| 2014 | **IG:** Photo editing becomes far more sophisticated<br>**TW:** Gamergate harassment campaign takes place in part on Twitter<br>**TW:** Announces a new suite of anti-harassment tools and promises faster response times for abuse complaints |
| 2015 | **FB:** Starts using information on how long people hover on a particular item in their news feed to gauge their level of interest in the item, in addition to the more explicit signals it currently uses (likes, comments, shares).<br>**TW:** Added Quote Tweet feature<br>**Snap:** Introduced selfie and geo-location filters, and a new way to view content from selected influencers. |
| 2016 | **IG:** Photo feed moves from chronological to algorithm-driven; Instagram Stories launch (disappear after 24 hours). Boomerang was added, users could tag each other, save posts, and post live streams.<br>**FB:** Launches Trust Indicators, a tool to help users determine how each particular publication works; Announces a set of news feed updates to combat the problem of fake news and hoaxes; Announces algorithm changes that penalize "clickbait" titles, based on a score assigned by a machine-learned model; Releases Facebook Reactions to the general public. The feature allows people to use five additional reactions beyond just the "like" action to convey their reaction to a post. The new reactions are "Love", "Haha", "Wow", "Sad", and "Angry." introduced FB live streaming. Facebook's Messenger adopted the 'stories' feature. FB Marketplace launched.<br>**TW:** Rolls out a change to its feed, making recommended tweets the default option, rather than the reverse chronological format that it had used |

| | |
|---|---|
| | since launch; added ability to retweet oneself.<br>**Reddit:** Launches a new blocking tool in an attempt to curb online harassment.<br>**Gab:** Founded<br>**Discord:** Founded<br>**Mastodon**: Founded |
| 2017 | **TikTok:** Founded<br>**TW:** Twitter increases tweets' character limit from 140 to 280 for all accounts; Redesign of user interface icons such as "like", "retweet", "reply", and circular profile pictures; ability to post tweet threads<br>**Reddit:** Bans the "altright" subreddit for violating its terms of service<br>**Snap:** Lens Studio Launches<br>**FB:** Launches Augmented Reality tool, Spark AR. |
| 2018 | **IG:** Launch of IGTV. Introduces Augmented Reality filters (see history)<br>**TikTok:** Becomes globally available<br>**Parler:** Founded |
| 2019 | **YT:** Updated its terms of service to state they are “under no obligation to host or serve content,” meaning content and channels can be removed at their discretion.<br>**FB:** Changes name to Meta |
| 2020 | **IG:** Launch of Reels<br>**Reddit:** In response to the George Floyd protests, Reddit announces a plan to revise its content policy to combat hate and racism on the site.<br>**Parler:** Parler had fewer than a million users until early 2020.In the last week of June 2020, it was estimated that the Parler app had more than 1.5 million daily users.<br>**BeReal:** Founded<br>**FB:** Launched subscriptions |
| 2021 | **Parler:** Removed from Apple and Google<br>**YT:** Removal of public dislike count<br>**Truth:** Founded<br>**TW:** Twitter Blue subscription service launched<br>**TikTok:** Becomes world’s most visited website |
| 2022 | **TW:** Elon Musk takes over. Proposes moderation council; tweets that people should be able to “choose your desired experience”.... The “Twitter Files” are released. Tweet “view count” becomes public. Numerous suspended accounts are revived (e.g., Donald Trump, Robert Malone).<br>**IG:** Enables users to revert to chronological newsfeed<br>**TikTok:** Age limit for live video hosting increases from 16 to 18. |

| | |
|---|---|
| | **O-AI:** Launches ChatGPT. Large language models begin to be incorporated into various social media platforms. |

# APPENDIX B: PNAS SPECIAL ISSUE ON POLARIZATION AND COMPLEX SYSTEMS

On December 14, 2021, PNAS devoted a large section to a special feature titled: Dynamics of Political Polarization. Only a few of these 11 essays deal directly with social media. But we include all of the essays in this appendix because together they do a great job of giving readers a perspective on complex dynamical systems, and their reactivity to small changes in key parameters. If social media is bad for democracy, it is likely to be because of such parameter changes, rather than by simple linear effects.

**B.1** Levin, Milner, & Perrings (2021). The dynamics of political polarization. *Proceedings of the National Academy of Sciences*. [Introduction to the series, gives a summary of each article]

EXCERPT: The main goal of the Special Feature is to deepen our understanding of the dynamics of political polarization and related trends, and especially the interplay among these processes at multiple scales, from the local to the international. The papers … pose a number of key questions. Do the dynamics of such systems follow a natural progression of polarization and collapse, similar to Schumpeter's economic theories (1)? How do migration, globalization, and new technologies, such as the internet, affect the trends? Does an extension of Duverger's Law (2) foreshadow a natural tendency toward polarization in nations with two-party systems, like that in the United States, undercutting Madison's dream (3)? Duverger's Law argues that a system like that of the United States, based on a plurality rule on a single ballot, will lead to a two-party system, while Madison hoped for a system that would "break and control the violence of faction" (3).… The Special Feature includes 11 individual articles, incorporating both novel research and Perspectives.

**B.2** Axelrod, Daymude, & Forrest (2021). Preventing extreme polarization of political attitudes. *Proceedings of the National Academy of Sciences*.

ABSTRACT: Extreme polarization can undermine democracy by making compromise impossible and transforming politics into a zero-sum game. "Ideological

polarization"—the extent to which political views are widely dispersed—is already strong among elites, but less so among the general public [N. McCarty, *Polarization: What Everyone Needs to Know*, 2019, pp. 50–68]. Strong mutual distrust and hostility between Democrats and Republicans in the United States, combined with the elites' already strong ideological polarization, could lead to increasing ideological polarization among the public. The paper addresses **two questions: 1) Is there a level of ideological polarization above which polarization feeds upon itself to become a runaway process? 2) If so, what policy interventions could prevent such dangerous positive feedback loops?** To explore these questions, we present an agent-based model of ideological polarization that differentiates between **the tendency for two actors to interact ("exposure") and how they respond when interactions occur, positing that interaction between similar actors reduces their difference, while interaction between dissimilar actors increases their difference**. Our analysis explores the effects on polarization of different levels of tolerance to other views, responsiveness to other views, exposure to dissimilar actors, multiple ideological dimensions, economic self-interest, and external shocks. The results suggest strategies for preventing, or at least slowing, the development of extreme polarization.

**B.3** Kawakatsu, Lelkes, Levin, & Tarnita (2021). Interindividual cooperation mediated by partisanship complicates Madison's cure for "mischiefs of faction." *Proceedings of the National Academy of Sciences*.

ABSTRACT: Political theorists have long argued that enlarging the political sphere to include a greater diversity of interests would cure the ills of factions in a pluralistic society. While the scope of politics has expanded dramatically over the past 75 y, polarization is markedly worse. Motivated by this paradox, **we take a bottom–up approach to explore how partisan individual-level dynamics in a diverse (multidimensional) issue space can shape collective-level factionalization via an emergent dimensionality reduction**. We extend a model of cultural evolution grounded in evolutionary game theory, in which individuals accumulate benefits through pairwise interactions and imitate (or learn) the strategies of successful others. The degree of partisanship determines the likelihood of learning from individuals of the opposite party. This approach captures the coupling between individual behavior, partisan-mediated opinion dynamics, and an interaction network that changes endogenously according to the evolving interests of individuals. We find that while expanding the diversity of interests can indeed improve both individual and collective outcomes, **increasingly high partisan bias promotes a reduction in issue dimensionality via party-based assortment that leads to increasing polarization**.

When party bias becomes extreme, it also boosts interindividual cooperation, thereby further entrenching extreme polarization and creating a tug-of-war between individual cooperation and societal cohesion. **These dangers of extreme partisanship are highest when individuals' interests and opinions are heavily shaped by peers and there is little independent exploration**. Overall, our findings highlight the urgency to study polarization in a coupled, multilevel context.

**B.4** Leonard, Lipsitz, Bizyaeva, Franci, & Lelkes (2021). The nonlinear feedback dynamics of asymmetric political polarization. *Proceedings of the National Academy of Sciences*.

ABSTRACT: Using a general model of opinion dynamics, we conduct a systematic investigation of key mechanisms driving elite polarization in the United States. We demonstrate that the self-reinforcing nature of elite-level processes can explain this polarization, with voter preferences accounting for its asymmetric nature. Our analysis suggests that **subtle differences in the frequency and amplitude with which public opinion shifts left and right over time may have a differential effect on the self-reinforcing processes of elites, causing Republicans to polarize more quickly than Democrats. We find that as self-reinforcement approaches a critical threshold, polarization speeds up. Republicans appear to have crossed that threshold while Democrats are currently approaching it**.

**B.5** Perrings, Hechter, & Mamada (2021). National polarization and international agreements. *Proceedings of the National Academy of Sciences*.

THIS ESSAY IS LESS RELEVANT FOR OUR REVIEW:
ABSTRACT: The network of international environmental agreements (IEAs) has been characterized as a complex adaptive system (CAS) in which the uncoordinated responses of nation states to changes in the conditions addressed by particular agreements may generate seemingly coordinated patterns of behavior at the level of the system. Unfortunately, since the rules governing national responses are ill understood, it is not currently possible to implement a CAS approach. Polarization of both political parties and the electorate has been implicated in a secular decline in national commitment to some IEAs, but the causal mechanisms are not clear. In this paper, we explore the impact of polarization on the rules underpinning national responses. We identify the degree to which responsibility for national decisions is shared across political parties and calculate the electoral cost of party positions as national obligations

under an agreement change. We find that polarization typically affects the degree but not the direction of national responses. Whether national commitment to IEAs strengthens or weakens as national obligations increase depends more on the change in national obligations than on polarization per se. Where the rules governing national responses are conditioned by the current political environment, so are the dynamic consequences both for the agreement itself and for the network to which it belongs. Any CAS analysis requires an understanding of such conditioning effects on the rules governing national responses.

**B.6** Chu, Donges, Robertson, & Pop-Eleches (2021). The microdynamics of spatial polarization: A model and an application to survey data from Ukraine. *Proceedings of the National Academy of Sciences*.

ABSTRACT: Although spatial polarization of attitudes is extremely common around the world, we understand little about the mechanisms through which polarization on divisive issues rises and falls over time. We develop a theory that explains how political shocks can have different effects in different regions of a country depending upon local dynamics generated by the preexisting spatial distribution of attitudes and discussion networks. **Where opinions were previously divided, attitudinal diversity is likely to persist after the shock. Meanwhile, where a clear precrisis majority exists on key issues, opinions should change in the direction of the predominant view. These dynamics result in greater local homogeneity in attitudes but at the same time exacerbate geographic polarization across regions and sometimes even within regions.** We illustrate our theory by developing a modified version of the adaptive voter model, an adaptive network model of opinion dynamics, to study changes in attitudes toward the European Union (EU) in Ukraine in the context of the Euromaidan Revolution of 2013 to 2014. Using individual-level panel data from surveys fielded before and after the Euromaidan Revolution, we show that **EU support increased in areas with high prior public support for EU integration but declined further where initial public attitudes were opposed to the EU, thereby increasing the spatial polarization of EU attitudes in Ukraine.** Our tests suggest that the predictive power of both network and regression models increases significantly when we incorporate information about the geographic location of network participants, which highlights the importance of spatially rooted social networks.

**B.7** Macy, Ma, Tabin, Gao, & Szymanski (2021). Polarization and tipping points. *Proceedings of the National Academy of Sciences*.

ABSTRACT: Research has documented increasing partisan division and extremist positions that are more pronounced among political elites than among voters. Attention has now begun to focus on **how polarization might be attenuated**. We use a general model of opinion change to see if **the self-reinforcing dynamics of influence and homophily may be characterized by tipping points that make reversibility problematic. The model applies to a legislative body or other small, densely connected organization**, but does not assume country-specific institutional arrangements that would obscure the identification of fundamental regularities in the phase transitions. Agents in the model have initially random locations in a multidimensional issue space consisting of membership in one of two equal-sized parties and positions on 10 issues. **Agents then update their issue positions by moving closer to nearby neighbors and farther from those with whom they disagree, depending on the agents' tolerance of disagreement and strength of party identification compared to their ideological commitment to the issues.** We conducted computational experiments in which we manipulated agents' tolerance for disagreement and strength of party identification. Importantly, we also introduced **exogenous shocks corresponding to events that create a shared interest against a common threat (e.g., a global pandemic).** Phase diagrams of political polarization reveal **difficult-to-predict transitions that can be irreversible due to asymmetric hysteresis trajectories.** We conclude that future empirical research needs to pay much closer attention to the identification of tipping points and the effectiveness of possible countermeasures.

**B.8** Santos, Lelkes, & Levin (2021). Link recommendation algorithms and dynamics of polarization in online social networks. *Proceedings of the National Academy of Sciences*.

ABSTRACT: The level of antagonism between political groups has risen in the past years. Supporters of a given party increasingly dislike members of the opposing group and avoid intergroup interactions, leading to homophilic social networks. While new connections offline are driven largely by human decisions, **new connections on online social platforms are intermediated by link recommendation algorithms, e.g., "People you may know" or "Whom to follow" suggestions. The long-term impacts of link recommendation in polarization are unclear, particularly as exposure to opposing viewpoints has a dual effect: Connections with out-group members can lead to opinion convergence and prevent group polarization or further separate opinions. Here, we provide a complex adaptive–systems perspective on the**

**effects of link recommendation algorithms**. While several models justify polarization through rewiring based on opinion similarity, **here we explain it through rewiring grounded in structural similarity—defined as similarity based on network properties. We observe that preferentially establishing links with structurally similar nodes (i.e., sharing many neighbors) results in network topologies that are amenable to opinion polarization. Hence, polarization occurs not because of a desire to shield oneself from disagreeable attitudes but, instead, due to the creation of inadvertent echo chambers. When networks are composed of nodes that react differently to out-group contacts, either converging or polarizing, we find that connecting structurally dissimilar nodes moderates opinions. Overall, our study sheds light on the impacts of social-network algorithms and unveils avenues to steer dynamics of radicalization and polarization in online social networks.**

**B.9** Stewart, Plotkin, & McCarty (2021). Inequality, identity, and partisanship: How redistribution can stem the tide of mass polarization. *Proceedings of the National Academy of Sciences*.

ABSTRACT: The form of political polarization where citizens develop strongly negative attitudes toward out-party members and policies has become increasingly prominent across many democracies. Economic hardship and social inequality, as well as intergroup and racial conflict, have been identified as important contributing factors to this phenomenon known as “affective polarization.” **Research shows that partisan animosities are exacerbated when these interests and identities become aligned with existing party cleavages.** In this paper, we use a model of cultural evolution to study how these forces combine to generate and maintain affective political polarization. We show that **economic events can drive both affective polarization and the sorting of group identities along party lines, which, in turn, can magnify the effects of underlying inequality between those groups. But, on a more optimistic note, we show that sufficiently high levels of wealth redistribution through the provision of public goods can counteract this feedback and limit the rise of polarization**. We test some of our key theoretical predictions using survey data on intergroup polarization, sorting of racial groups, and affective polarization in the United States over the past 50 y.

**B.10** Tokita, Guess, & Tarnita (2021). Polarized information ecosystems can reorganize social networks via information cascades. *Proceedings of the National Academy of Sciences*.

ABSTRACT: The precise mechanisms by which the information ecosystem polarizes society remain elusive. Focusing on political sorting in networks, **we develop a computational model that examines how social network structure changes when individuals participate in information cascades, evaluate their behavior, and potentially rewire their connections to others as a result.** Individuals follow proattitudinal information sources but are more likely to first hear and react to news shared by their social ties and only later evaluate these reactions by direct reference to the coverage of their preferred source. **Reactions to news spread through the network via a complex contagion.** Following a cascade, individuals who determine that their participation was driven by a subjectively "unimportant" story adjust their social ties to avoid being misled in the future. In our model, this dynamic leads social networks to politically sort when news outlets differentially report on the same topic, even when individuals do not know others' political identities. **Observational follow network data collected on Twitter support this prediction: We find that individuals in more polarized information ecosystems lose cross-ideology social ties at a rate that is higher than predicted by chance. Importantly, our model reveals that these emergent polarized networks are less efficient at diffusing information: Individuals avoid what they believe to be "unimportant" news at the expense of missing out on subjectively "important" news far more frequently. This suggests that "echo chambers"—to the extent that they exist—may not echo so much as silence.**

[NOTE from Tokita: Our paper studies echo chamber formation on social media; however, we show/suggest that polarized media coverage is what is ultimately creating echo chambers online, as reactions to news coverage spread through social networks and cause people to adjust their social ties. We show that people in more polarized information ecosystems—that is, consuming more partisan news that is out of sync with other sources—lose social ties to people of the opposite ideology, even when they don't know each other's politics. This happens because people compare the behavior of their friends against what their preferred news outlet is reporting and break social ties with friends—some of whom might be consuming other news sources aligned with their personal politics—who appear to be acting "out of sync" with the reality presented by their news source. Therefore, we suggest that ultimately it is the information ecosystem (news coverage) that is reshaping our social networks, without us realizing it, although clearly we focus on how this is playing out on social media.]

**B.11** Vasconcelos, Constantino, Dannenberg, Lumkowsky, Weber, & Levin (2021). Segregation and clustering of preferences erode socially beneficial coordination. *Proceedings of the National Academy of Sciences*.

ABSTRACT: Polarization on various issues has increased in many Western democracies over the last decades, leading to divergent beliefs, preferences, and behaviors within societies. We develop a model to investigate the effects of polarization on the likelihood that a society will coordinate on a welfare-improving action in a context in which collective benefits are acquired only if enough individuals take that action. We examine the impacts of different manifestations of polarization: heterogeneity of preferences, segregation of the social network, and the interaction between the two. In this context, heterogeneity captures differential perceived benefits from coordinating, which can lead to different intentions and sensitivity regarding the intentions of others. Segregation of the social network can create a bottleneck in information flows about others' preferences, as individuals may base their decisions only on their close neighbors. Additionally, heterogeneous preferences can be evenly distributed in the population or clustered in the local network, respectively reflecting or systematically departing from the views of the broader society. **The model predicts that heterogeneity of preferences alone is innocuous and it can even be beneficial, while segregation can hamper coordination, mainly when local networks distort the distribution of valuations.** We base these results on a multimethod approach including an online group experiment with 750 individuals. We randomize the range of valuations associated with different choice options and the information respondents have about others. The experimental results reinforce the idea that, **even in a situation in which all could stand to gain from coordination, polarization can impede social progress.**

# APPENDIX C: CRITIQUES OF HAIDT'S "UNIQUELY STUPID" ATLANTIC ARTICLE

I (Jon Haidt) published an essay in The Atlantic on April 11, 2022 titled WHY THE PAST 10 YEARS OF AMERICAN LIFE HAVE BEEN UNIQUELY STUPID. Below are some constructive criticisms of it from scholars, industry insiders, and others who sound at least vaguely scholarly. I thank these critics, whose criticisms will help me to write a better book. Meta responded to my essay, and The Atlantic gave me the opportunity to respond to Meta with a second essay, titled Yes, social media really is undermining democracy, despite what Meta has to say.

.

C.1 Twitter thread from Tobias Dienlin, @tdienlin

C.2 Micah Sifrey, Did the Internet Break Democracy?

C.3 Twitter thread from Christian Hoffmann, @cphoffmann

C.4 Twitter thread from Daniel Kreiss, @kreissdaniel

C.5. Twitter Thread from Thomas Zeitzoff, @Zeitsoff

C.6 Twitter Thread from Mike Mazarr, @MMazarr

C.7 Samuel James. What Jonathan Haidt is Missing

C.8 Meta's official response to Haidt's essay: What the Research on Social Media's Impact on Democracy and Daily Life Says (and Doesn't Say)

**The research Meta cites in its defense (with locations in this doc):**

**Studies**

**1.2.2** Boxell, Gentzkow, & Shapiro (2021). [this was originally published 2020]
**1.2.1** Boxell, Gentzkow, & Shapiro (2017)
**3.3.18** Benkler et al. 2020

**Reports / Reviews**

**9.1.17** Digitization and Democracy working group
**9.1.19** Reuters Institute digital news report 2017

**Books**

**10.6** Bruns (2019). *Are filter bubbles real?* Wiley.

**Haidt's response to Meta's rebuttal:** Published in The Atlantic. But this whole Google doc was the basis for my Atlantic essay, so you can decide for yourself if I have mischaracterized "the preponderance" of the research. Meta seems to be right on the filter bubble question for exposure to NEWS articles, but wrong on immersion in social networks of like-minded people (homophily). And there are 6

other questions. See especially the largest review, Lorenz-Spreen et al., study **9.1.13**, which I linked to in the essay.

C.9. Ian Leslie: Is social media to blame for everything?

C.10. Robert Wright: Is Everything Falling Apart? NonZero Newsletter

C.11. Mark Mutz & Richard Gunderman: Put Not Thy Trust in Technology. At Law & Liberty

C.12. Matthew Ingram: Have the dangers of social media been overstated? Columbia Journalism Review.

C.13. Tanner Greer: Our Problems Aren't Procedural. City Journal.

C14. Nirit Weiss-Blatt. Don't Be So Certain That Social Media Is Undermining Democracy

# APPENDIX D: IS POLITICAL DYSFUNCTION INCREASING IN THE AGE OF SOCIAL MEDIA?

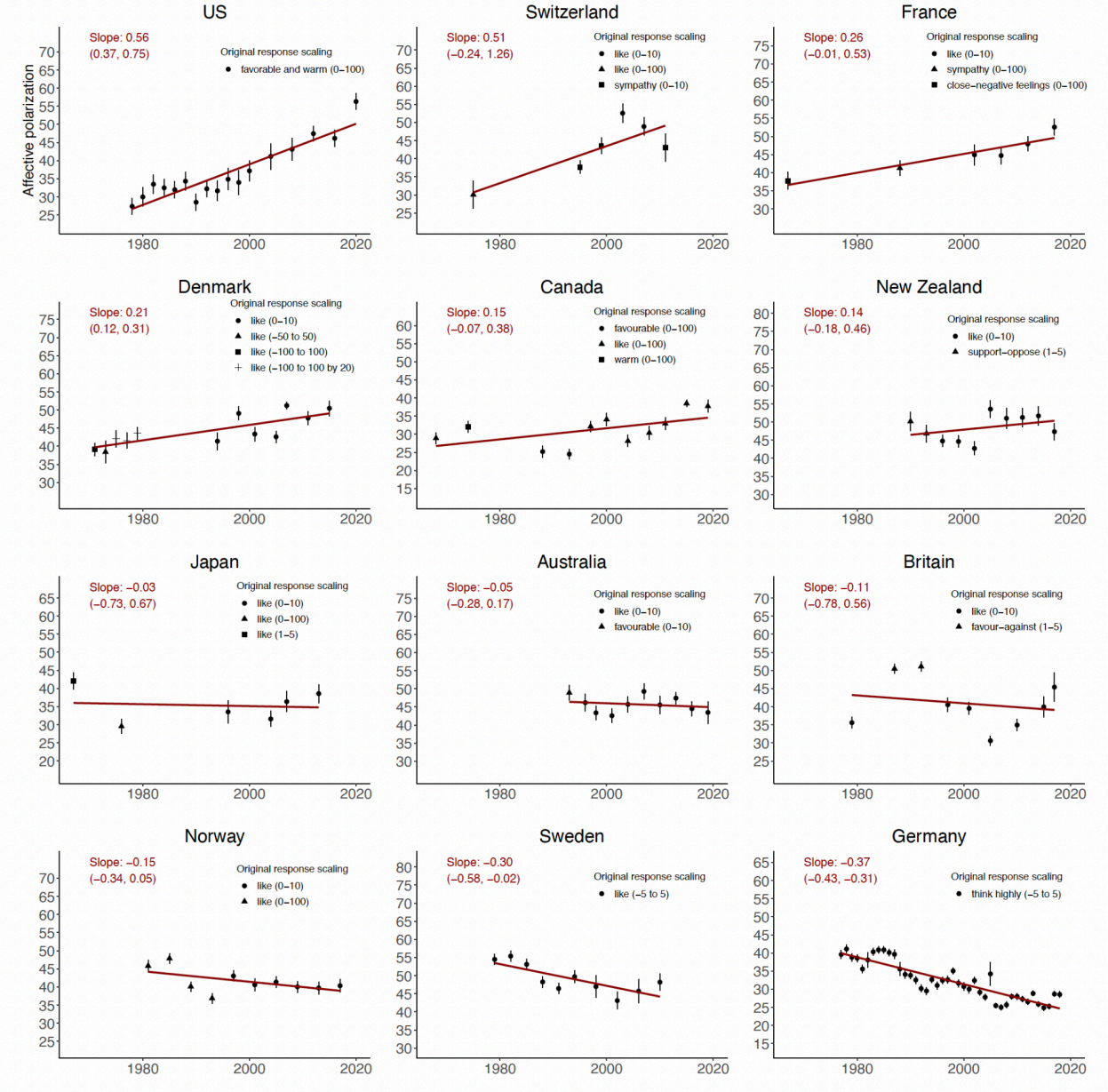
In its response to Haidt's "Uniquely Stupid" essay, Meta's Head of Research, Pratiti Raychoudhury, said:

> "Evidence simply does not support the idea that Facebook, or social media generally, is the primary cause of polarization. Research from Stanford last year looked in depth at trends in nine countries over 40 years, and found that in some countries polarization was on the rise before Facebook even existed, and in others it has been decreasing while internet and Facebook use increased."

The research she links to is already in this collaborative review doc: 1.2.2 Boxell, Gentzkow, & Shapiro (2021). *Cross-country trends in affective polarizatio*n. Raychoudhury is asserting that polarization is not rising globally, even though Facebook use was rising globally. But Boxell et al. plotted straight-line graphs *for the entire period* for which they had data, from the 1970s (for some countries) through 2020 (for some countries). Those graphs are interesting but they are not the right graphs to evaluate the specific hypothesis in Haidt's essay, which is that Facebook and Twitter pioneered architectural features from 2009 through 2012 (such as the like button, retweet/share button, and also threaded comments, which were introduced in 2013) that made the major social media platforms much more viralized, mobocratic, and effective for attacking and intimidating people. In other words, the "dart guns" of social media were only handed out globally in the early 2010s, so we should not expect to see any measurable increase in

downstream democratic dysfunction (such as rising affective polarization or democratic backsliding) for a few years after that. The most relevant graphs would therefore be ones plotted with a hinge point around 2013. Do the trends from 2013-2020 generally slope upward, compared to the trendlines from 1980-2013?

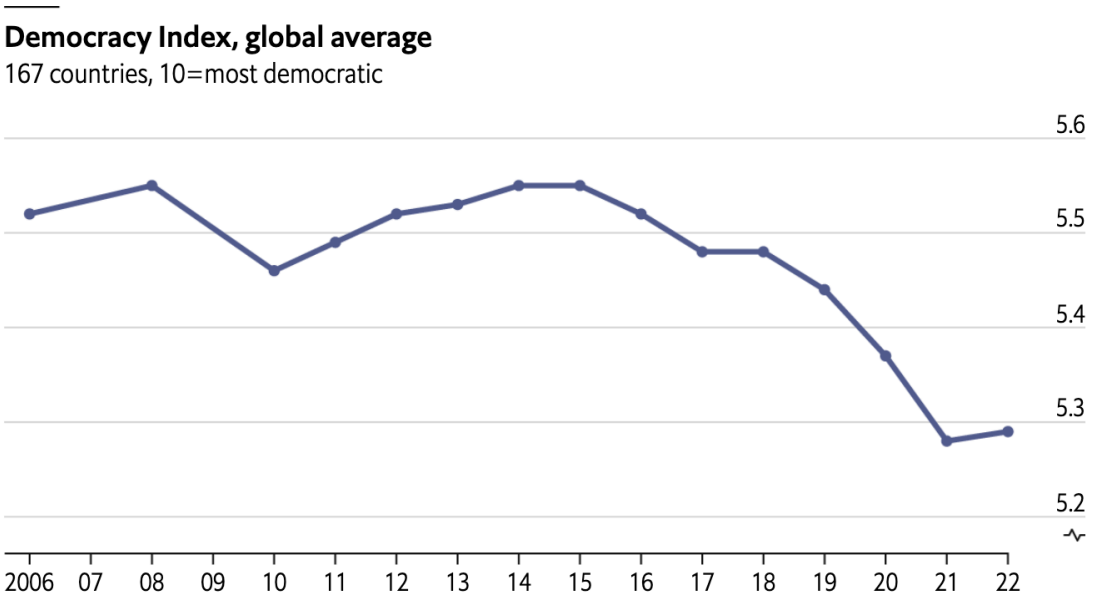
Figure 1: Trends in Affective Polarization by Country




Several institutes and academic papers have documented a global decline in the number or quality of democracies, which began to drop in the 2010s; or a global rise in polarization

D.1. Economist Intelligence Unit: A new low for global democracy (2022).
Updated report, 2023.

Updated figure with 2022 data:




D.2. V-Dem Institute, DEMOCRACY REPORT 2022: Autocratization Changing Nature?
From the Executive Summary:

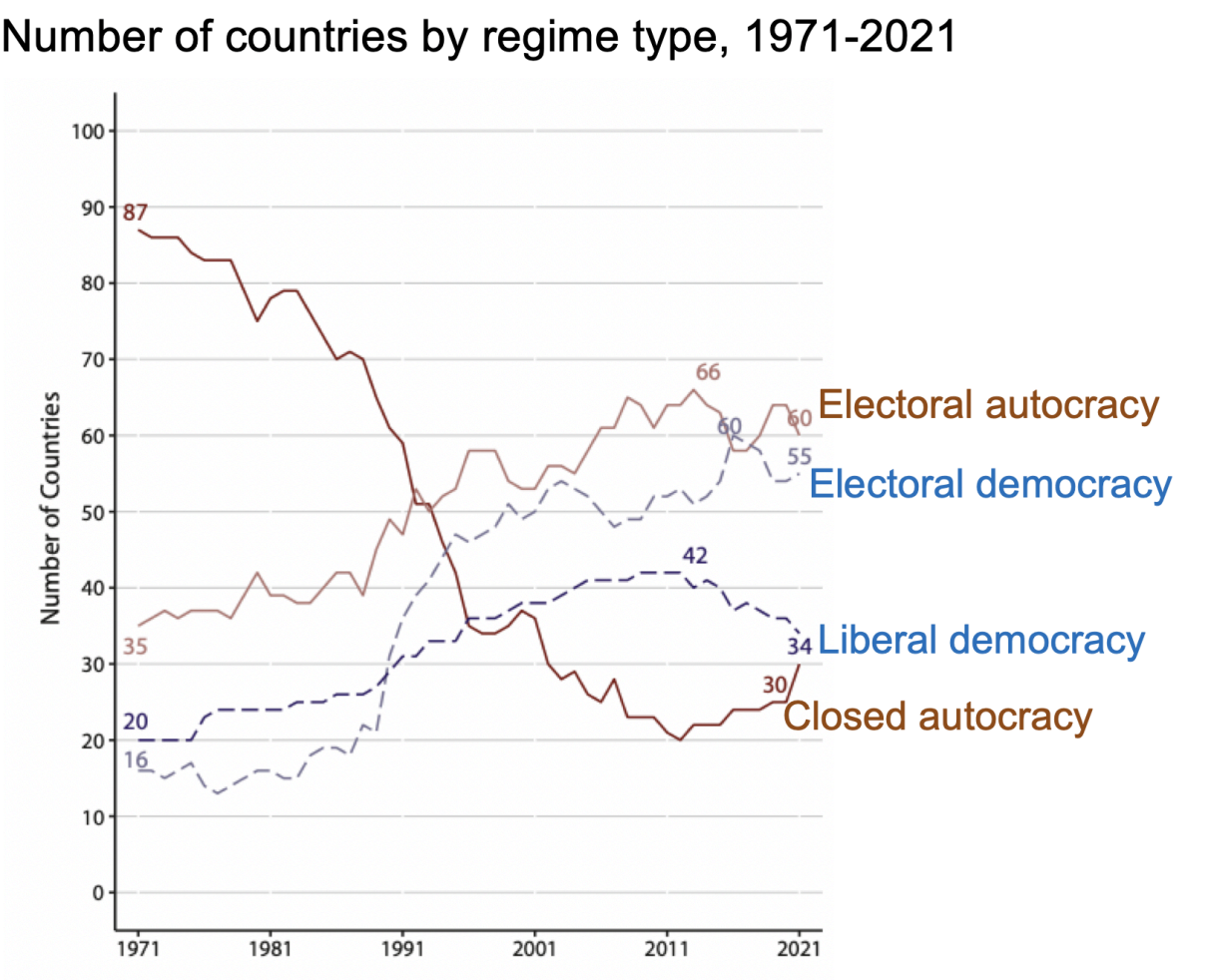
Back to 1989 Levels:

- Liberal democracies peaked in 2012 with 42 countries and are now down to the lowest levels in over 25 years: 34 nations, home to only 13% of the world population.

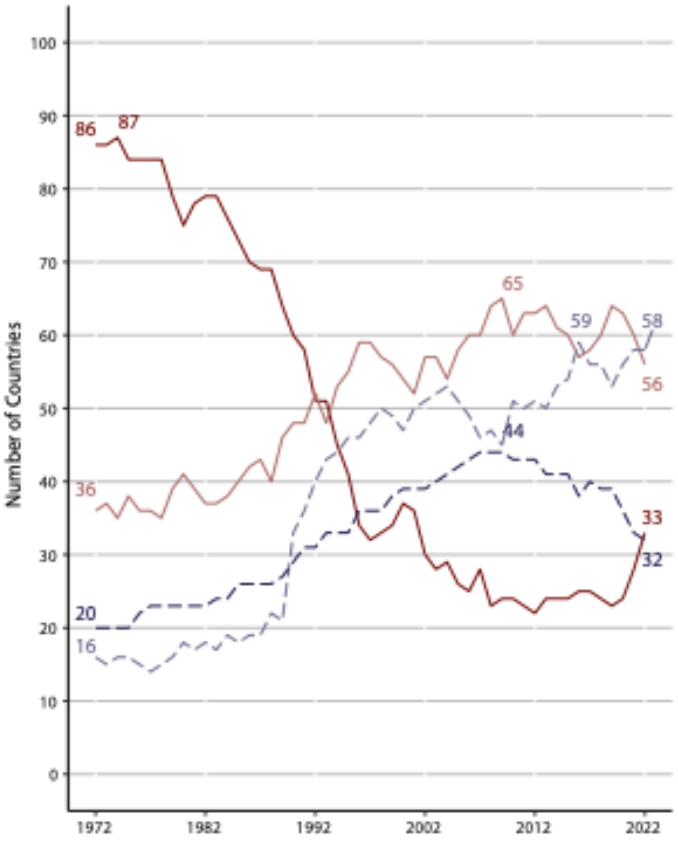
Ten Years Ago – A Different World:

- A record of 35 countries suffered significant deteriorations in freedom of expression at the hands of governments – an increase from only 5 countries 10 years ago.
- A signal of toxic polarization, respect for counterarguments and associated aspects of the deliberative component of democracy got worse in more than 32 countries – another increase from only 5 nations in 2011.

[Note the decline in liberal democracies, and the rise in closed autocracies, in the 2010s, in Figure 4 on p. 14]

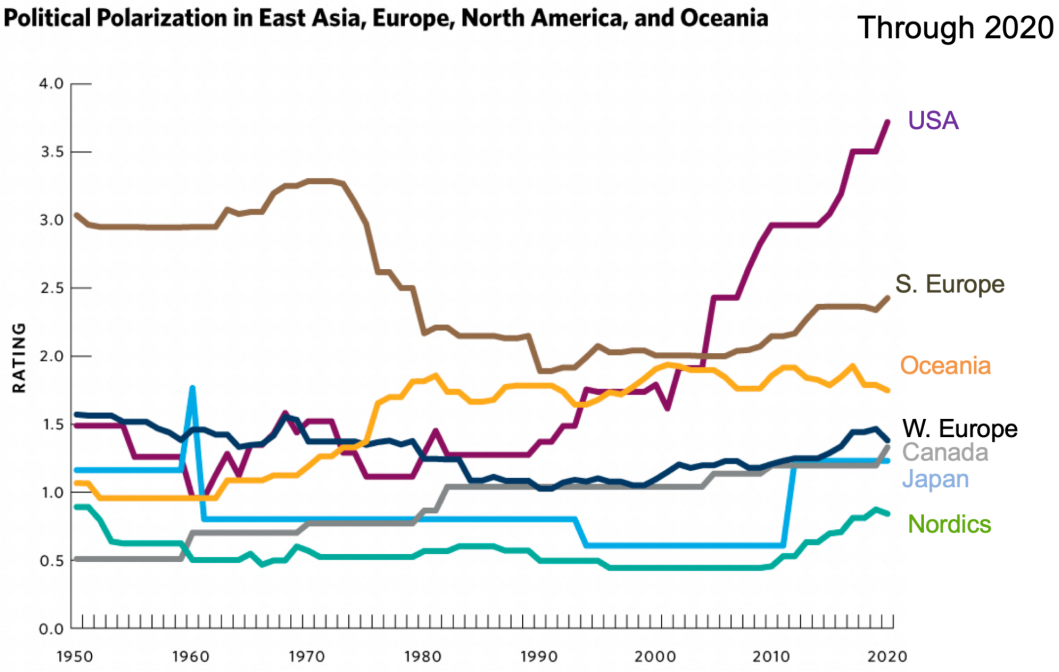



Here is the updated figure from the [2023 V-dem report]:




D.3. Carnegie Endowment for International Peace, 2020 report: [What Happens When Democracies Become Perniciously Polarized]?
Plotting polarization data from [V-dem.net] by region.

[You can see that polarization has increased the most in the USA, but it has also increased during the 2010s in Southern Europe, Western Europe, the Nordics, and Japan.]




**D.4.** Orhan (2022). The relationship between affective polarization and democratic backsliding: comparative evidence. *Democratization*.

ABSTRACT: Why do voters vote for undemocratic politicians in a democracy? My chief contention is that affective polarization has become a primary factor driving support for undemocratic politicians. Once partisan identification turns into a salient identity in the hierarchy of group affiliations, it has the potential to widen inter-party distances. Such a political environment fosters positive beliefs of their preferred party and negative beliefs of the other party, which promote political cynicism, intolerance and increase partisan loyalty. As a result, crossing party lines becomes costly, even when incumbents violate democratic principles or incumbents' economic policies do not appeal to supporters' interests. This tradeoff enables undemocratic politicians to evade electoral sanctions for undemocratic behaviour. I created an extended version of Reiljan's affective polarization application. **The new dataset covers affective polarization scores of 53 countries calculated over 170 national election surveys. I find that increasing affective polarization is highly correlated with democratic backsliding, less accountability, less freedom, fewer rights, and less deliberation in democracies.** However, ideological polarization has shown no correlation.

[Note the rise in "very high level backsliding" in the 2010s in Figure 2:]

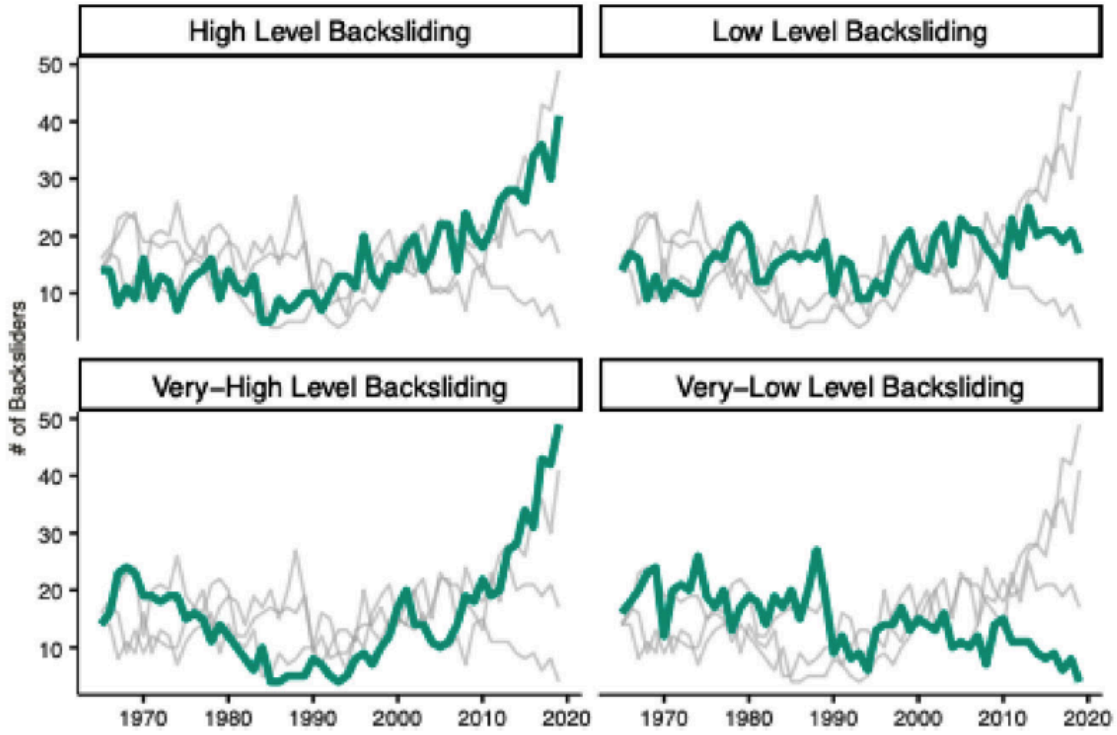



**Figure 2.** The Distribution of Backsliders Worldwide, 1965–2020.

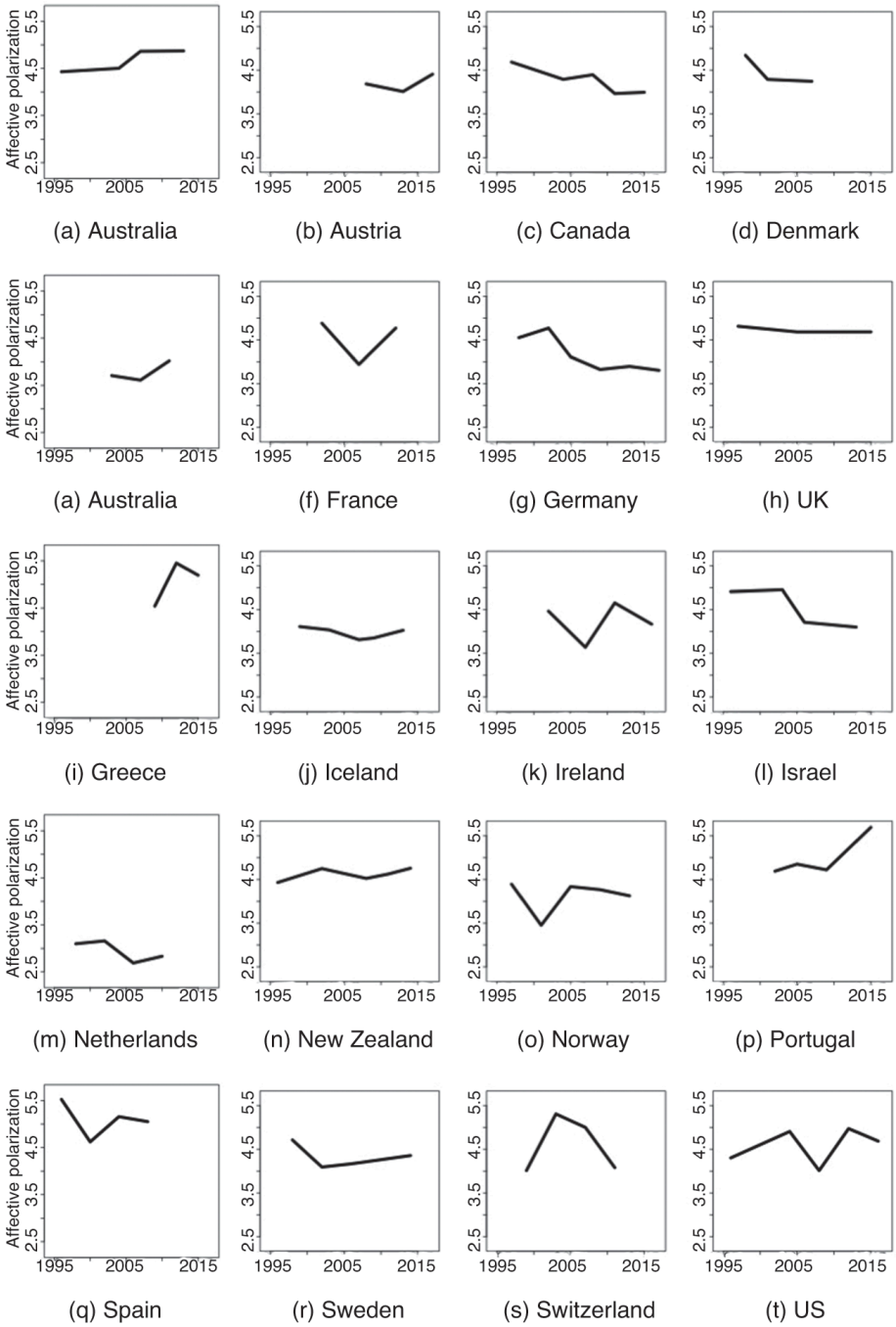
D.5. Gidron, Adams, & Horne (2020). American Affective Polarization in Comparative Perspective. Cambridge Elements: American Politics.

ABSTRACT: American political observers express increasing concern about affective polarization, i.e., partisans' resentment toward political opponents. We advance debates about America's partisan divisions by comparing affective polarization in the US over the past 25 years with affective polarization in 19 other western publics. We conclude that **American affective polarization is not extreme in comparative perspective, although Americans' dislike of partisan opponents has increased more rapidly since the mid-1990s than in most other Western publics**. We then show that affective polarization is more intense when unemployment and inequality are high; when political elites clash over cultural issues such as immigration and national identity; and in countries with majoritarian electoral institutions. Our findings situate American partisan resentment and hostility in comparative perspective, and illuminate correlates of affective polarization that are difficult to detect when examining the American case in isolation.

[Additional excerpts/notes:]

- “there is no clear over time trend of intensifying (or declining) affective polarization across Western publics.”
- “Americans’ out-party hostility has increased more sharply than what we see in most other Western democracies (although there is suggestive evidence that growing out-party dislike may be a cross-national trend).”

[Note that data for 14 of the 19 countries ends in 2015 or earlier, so we can’t evaluate what happens in the late 2010s, but here is the main figure examining temporal trends in affective polarization, and finding no overall trend:]



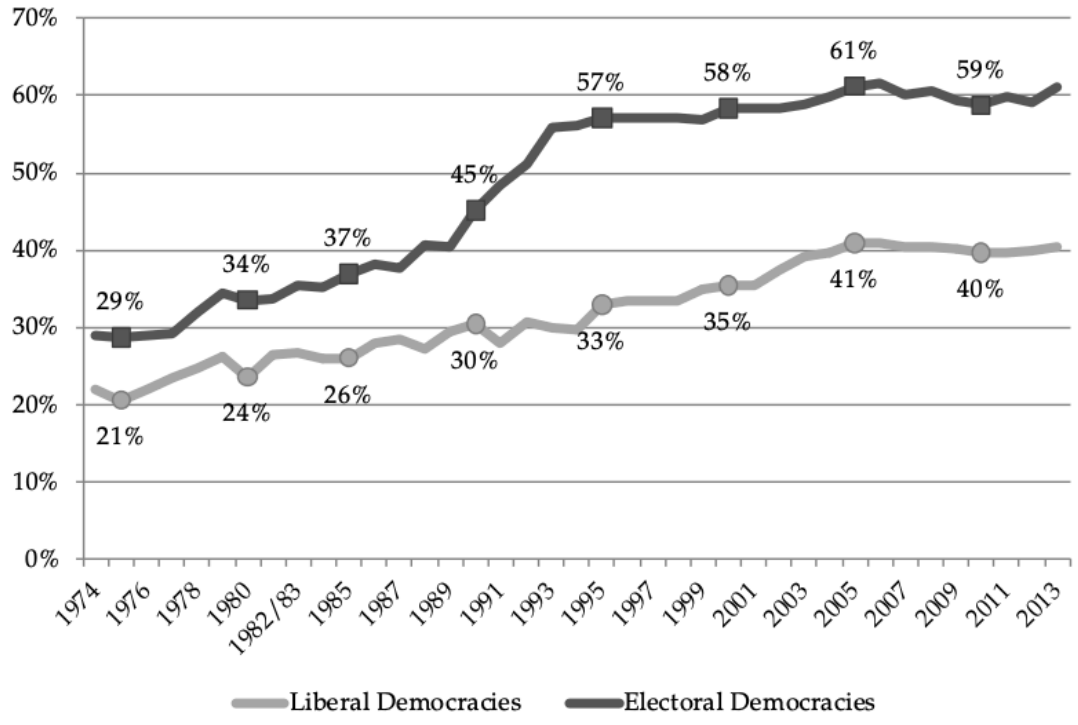

D.6. Diamond (2015). Facing Up to the Democratic Recession. *Journal of Democracy*.

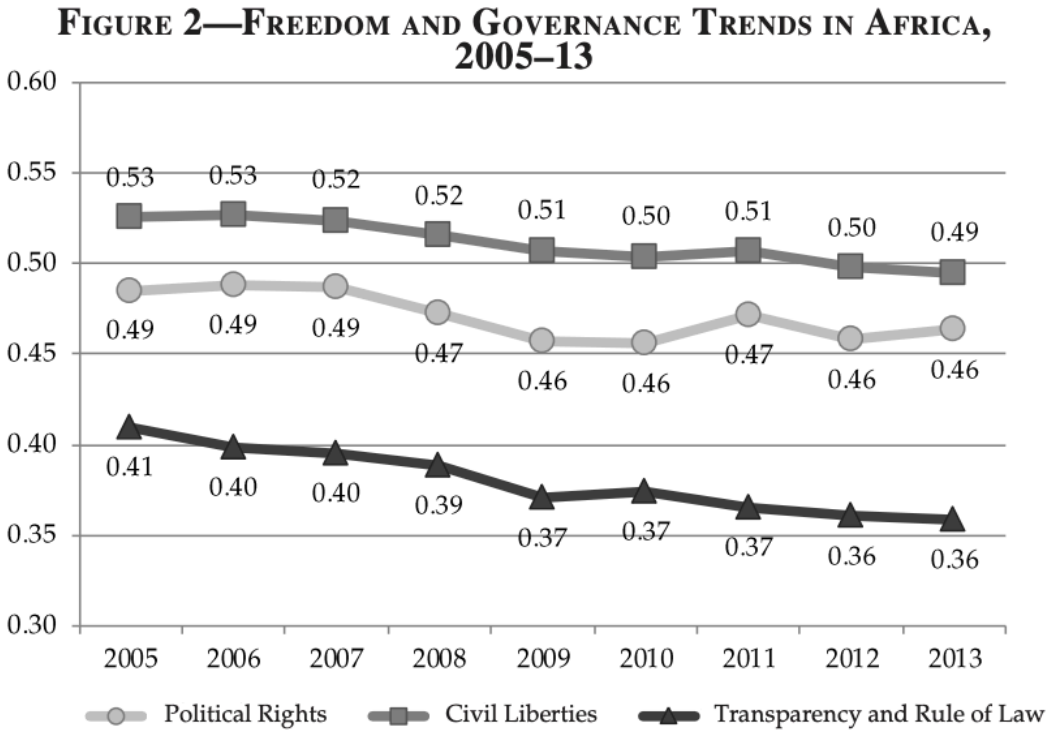
ABSTRACT: Democracy has been in a global recession for most of the last decade. Yet the picture is not entirely bleak. We have not seen “a third reverse wave.” The key imperative in the near term is to work to reform and consolidate the democracies that

have emerged during the third wave—the majority of which remain illiberal and unstable, if they remain democratic at all. It is vital that democrats in the established democracies not lose faith. Democrats have the better set of ideas. Democracy may be receding somewhat in practice, but it is still globally ascendant in peoples' values and aspirations.

FIGURES:

**FIGURE 1—THE GROWTH OF DEMOCRACIES IN THE WORLD, 1974–2013**






# APPENDIX E: EMPIRICAL STUDIES THAT BEAR ON WAYS TO IMPROVE SOCIAL MEDIA

The studies in this section are empirical studies that are often related to specific proposals in Section 11, “Proposals for improving social media.”

**E.1** Bazarova, Choi, Sosik, Cosley, & Whitlock (2015). Social sharing of emotions on Facebook. *Proceedings of the 18th ACM conference on computer supported cooperative work & social computing*.

ABSTRACT: People often share emotions with others in order to manage their emotional experiences. We investigate how social media properties such as visibility and directedness affect how people share emotions in Facebook, and their satisfaction after doing so. 141 participants rated 1,628 of their own recent status updates, posts they made on others' timelines, and private messages they sent for intensity, valence,

personal relevance, and overall satisfaction felt after sharing each message. For network-visible channels-status updates and posts on others' timelines-they also rated their satisfaction with replies they received. People shared differently between channels, with more intense and negative emotions in private messages. People felt more satisfied after sharing more positive emotions in all channels and after sharing more personally relevant emotions in network-visible channels. Finally, people's overall satisfaction after sharing emotions in network-visible channels is strongly tied to their reply satisfaction. Quality of replies, not just quantity, matters, suggesting the need for designs that help people receive valuable responses to their shared emotions.

**E.2** Matias (2019). Preventing harassment and increasing group participation through social norms in 2,190 online science discussions. *Proceedings of the National Academy of Sciences*.

ABSTRACT: Online harassment remains a common experience despite decades of work to identify unruly behavior and enforce rules against it. Consequently, many people avoid participating in online conversations for fear of harassment. Using a large-scale field experiment in a community with 13 million subscribers, I show that it is possible to prevent unruly behavior and also increase newcomer participation in public discussions of science. Announcements of community rules in discussions increased the chance of rule compliance by >8 percentage points and increased newcomer participation by 70% on average. This study demonstrates the influence of community rules on who chooses to join a group and how they behave.

**E.3** Jaidka, Zhou, & Lelkes (2019). Brevity is the soul of Twitter: The constraint affordance and political discussion. *Journal of Communication*.

ABSTRACT: Many hoped that social networking sites would allow for the open exchange of information and a revival of the public sphere. Unfortunately, conversations on social media are often toxic and not conducive to healthy political discussions. Twitter, the most widely used social network for political discussions, doubled the limit of characters in a tweet in November 2017, which provided an opportunity to study the effect of technological affordances on political discussions using a discontinuous time series design. Using supervised and unsupervised natural language processing methods, we analyzed 358,242 tweet replies to U.S. politicians from January 2017 to March 2018. **We show that doubling the permissible length of a tweet led to less uncivil, more polite, and more constructive discussions online. However, the**

**declining trend in the empathy and respectfulness of these tweets raises concerns about the implications of the changing norms for the quality of political deliberation.**

**E.4** Nyhan (2021). Why the backfire effect does not explain the durability of political misperceptions. *Proceedings of the National Academy of Sciences*.

ABSTRACT: Previous research indicated that corrective information can sometimes provoke a so-called "backfire effect" in which respondents more strongly endorsed a misperception about a controversial political or scientific issue when their beliefs or predispositions were challenged. I show how subsequent research and media coverage seized on this finding, distorting its generality and exaggerating its role relative to other factors in explaining the durability of political misperceptions. To the contrary, an emerging research consensus finds that corrective information is typically at least somewhat effective at increasing belief accuracy when received by respondents. However, the research that I review suggests that the accuracy-increasing effects of corrective information like fact checks often do not last or accumulate; instead, they frequently seem to decay or be overwhelmed by cues from elites and the media promoting more congenial but less accurate claims. As a result, misperceptions typically persist in public opinion for years after they have been debunked. Given these realities, **the primary challenge for scientific communication is not to prevent backfire effects but instead, to understand how to target corrective information better and to make it more effective. Ultimately, however, the best approach is to disrupt the formation of linkages between group identities and false claims and to reduce the flow of cues reinforcing those claims from elites and the media.** Doing so will require a shift from a strategy focused on providing information to the public to one that considers the roles of intermediaries in forming and maintaining belief systems.

**E.5** Pennycook & Rand (2019). Fighting misinformation on social media using crowdsourced judgments of news source quality. *Proceedings of the National Academy of Sciences*.

ABSTRACT: Many people consume news via social media. It is therefore desirable to reduce social media users' exposure to low-quality news content. One possible intervention is for **social media ranking algorithms to show relatively less content from sources that users deem to be untrustworthy**. But are laypeople's judgments reliable indicators of quality, or are they corrupted by either partisan bias or lack of

information? **Perhaps surprisingly, we find that laypeople—on average—are quite good at distinguishing between lower- and higher-quality sources. These results indicate that incorporating the trust ratings of laypeople into social media ranking algorithms may prove an effective intervention against misinformation, fake news, and news content with heavy political bias.**

**E.6** Allen, Arechar, Pennycook, & Rand (2021). Scaling up fact-checking using the wisdom of crowds. *Science Advances.*

ABSTRACT: Professional fact-checking, a prominent approach to combating misinformation, does not scale easily. Furthermore, some distrust fact-checkers because of alleged liberal bias. **We explore a solution to these problems: using politically balanced groups of laypeople to identify misinformation at scale.** Examining 207 news articles flagged for fact-checking by Facebook algorithms, we compare accuracy ratings of three professional fact-checkers who researched each article to those of 1128 Americans from Amazon Mechanical Turk who rated each article's headline and lede. **The average ratings of small, politically balanced crowds of laypeople (i) correlate with the average fact-checker ratings as well as the fact-checkers' ratings correlate with each other and (ii) predict whether the majority of fact-checkers rated a headline as "true" with high accuracy.** Furthermore, cognitive reflection, political knowledge, and Democratic Party preference are positively related to agreement with fact-checkers, and identifying each headline's publisher leads to a small increase in agreement with fact-checkers.

**E.7** Pennycook, Epstein, Mosleh, Arechar, Eckles, & Rand (2021). Shifting attention to accuracy can reduce misinformation online. *Nature*.

ABSTRACT: In recent years, there has been a great deal of concern about the proliferation of false and misleading news on social media. Academics and practitioners alike have asked why people share such misinformation, and sought solutions to reduce the sharing of misinformation. Here, we attempt to address both of these questions. First, we find that the veracity of headlines has little effect on sharing intentions, despite having a large effect on judgments of accuracy. This dissociation suggests that sharing does not necessarily indicate belief. Nonetheless, most participants say it is important to share only accurate news. To shed light on this apparent contradiction, we carried out four survey experiments and a field experiment on Twitter; the results show that **subtly shifting attention to accuracy increases the quality of news that people**

**subsequently share. Together with additional computational analyses, these findings indicate that people often share misinformation because their attention is focused on factors other than accuracy—and therefore they fail to implement a strongly held preference for accurate sharing**. Our results challenge the popular claim that people value partisanship over accuracy, and provide evidence for scalable attention-based interventions that social media platforms could easily implement to counter misinformation online.

**E.8** Van Alstyne (2022). Free speech, platforms & the fake news problem (Applies mechanism design and information economics to reduce the spread of misinformation). Available at *SSRN.*

ABSTRACT: How should a platform or a society address the problem of fake news? The spread of misinformation is ancient, complex, yet ubiquitous in media concerning elections, vaccinations, and global climate policy. After examining key attributes of "fake news" and of current solutions, this article presents design tradeoffs for curbing fake news. The challenges are not restricted to truth or to scale alone. Surprisingly, there exist boundary cases when a just society is better served by a mechanism that allows lies to pass, even as there are alternate boundary cases when a just society should put friction on truth. Harm reflects an interplay of lies, decision error, scale, and externalities. Using mechanism design, this article then proposes three tiers of solutions: (1) those that are legal and business model compatible, so firms should adopt them (2) those that are legal but not business model compatible, so firms need compulsion to adopt them, and (3) those that require changes to bad law.

**E.9** Henry, Zhuravskaya, & Guriev (2022). Checking and sharing alt-facts. *American Economic Journal: Economic Policy.* (h/t Sergei Guriev)

ABSTRACT: During the 2019 European elections campaign, we exposed a random sample of French voting-age Facebook users to false statements by a far-right populist party. A randomly selected subgroup was also presented with fact-checking of these statements; another subgroup was offered a choice whether to view the fact-checking. Participants could then share these statements on their Facebook pages. **We show that (i) both imposed and voluntary fact-checking reduce sharing of false statements by about 45%; (ii) the size of the effect is similar between imposed and voluntary fact-checking; and (iii) each additional click required to share false statements substantially reduces sharing.**

**E.10** Barrera, Guriev, Henry, and Zhuravskaya (2020). Facts, alternative facts, and fact checking in times of post-truth politics. *Journal of Public Economics.* (h/t Sergei Guriev)

ABSTRACT: How effective is fact checking in countervailing "alternative facts," i.e., misleading statements by politicians? In a randomized online experiment during the 2017 French presidential election campaign, we subjected subgroups of 2480 French voters to alternative facts by the extreme-right candidate, Marine Le Pen, and/or corresponding facts about the European refugee crisis from official sources. We find that: **(i) alternative facts are highly persuasive; (ii) fact checking improves factual knowledge of voters (iii) but it does not affect policy conclusions or support for the candidate; (iv) exposure to facts alone does not decrease support for the candidate, even though voters update their knowledge**. We find evidence consistent with the view that at least part of the effect can be explained by raising salience of the immigration issue.

E.11 Prosocial Design Network [h/t Olivia Fischer]

WHAT THEY DO: We believe that digital products can be designed to help us better understand one another. That's why we are building an international network of behavioral science and design experts to articulate a better, more prosocial future online; and to disentangle the Web's most glaring drawbacks: from misunderstandings to incitements to hatred.
MISSION: We are the Prosocial Design Network, and our mission is to promote prosocial design: evidence-based design practices that bring out the best in human nature online. We inspire and co-create the infrastructure needed to turn the web into something that does not weaken, but rather strengthens, the wellness of our democracy and ourselves.

* * * * * * END OF REVIEW * * * * * * * * *