aps ASSOCIATION FOR PSYCHOLOGICAL SCIENCE

Perspectives on Psychological Science
2024, Vol. 19(5) 781–795

DOI: 10.1177/17456916231190392
www.psychologicalscience.org/PPS

Sage

People Think That Social Media Platforms Do (but Should Not) Amplify Divisive Content

Steve Rathje[1], Claire Robertson[1], William J. Brady[2], and Jay J. Van Bavel[1]

[1]Department of Psychology & Center for Neural Science, New York University, and
[2]Kellogg School of Management, Northwestern University

Abstract

Recent studies have documented the type of content that is most likely to spread widely, or go "viral," on social media, yet little is known about people's perceptions of what goes viral or what should go viral. This is critical to understand because there is widespread debate about how to improve or regulate social media algorithms. We recruited a sample of participants that is nationally representative of the U.S. population (according to age, gender, and race/ethnicity) and surveyed them about their perceptions of social media virality ($n = 511$). In line with prior research, people believe that divisive content, moral outrage, negative content, high-arousal content, and misinformation are all likely to go viral online. However, they reported that this type of content should not go viral on social media. Instead, people reported that many forms of positive content—such as accurate content, nuanced content, and educational content—are not likely to go viral even though they think this content should go viral. These perceptions were shared among most participants and were only weakly related to political orientation, social media usage, and demographic variables. In sum, there is broad consensus around the type of content people think social media platforms should and should not amplify, which can help inform solutions for improving social media.

Keywords

algorithms, social media, virality, misinformation, polarization

Almost 5 billion people—or more than half the world's population—are now on social media (Statista, 2022b), and people use social media for about 147 min each day (Statista, 2022a). The content people consume on social media is greatly influenced by news feed algorithms, making social media algorithms particularly important to study. There has been intensive speculation about how social media news feed algorithms work, what type of content they amplify, and what broader impact they have on society. Indeed, some have speculated that the creation of algorithmically curated news feed on social media (as opposed to to a "chronological" news feed) had detrimental effects on democracy (Haidt, 2022). Others have argued that social media algorithms may accelerate polarization and the spread of misinformation by amplifying divisive or false content (Harris et al., 2023; Van Bavel et al., in press; Van Bavel, Harris, et al., 2021; Van Bavel, Rathje, et al., 2021; van der Linden et al., 2021). However, others have suggested that social media algorithms have little effect on people's behavior compared with user preferences (Bakshy et al., 2015) and that algorithms have many benefits for users, such as blocking out misinformation and spam (Eckles, 2022). There has been an intensive social debate about these issues among the general public and policymakers alike, along with congressional hearings about how to improve or regulate social media algorithms (C-Span, 2021).

Unfortunately, the speculation around social media algorithms likely exceeds actual public knowledge about how social media algorithms work and what their impacts are on society. Little is known about how social media algorithms work, in part because of the proprietary

Corresponding Author:
Jay J. Van Bavel, Department of Psychology & Center for Neural Science, New York University
Email: jay.vanbavel@nyu.edu

nature of social media algorithms, lack of transparency from social media companies, and the complexity of these algorithms (Bak-Coleman et al., 2021; Narayanan, 2023). Moreover, recommendation algorithms are frequently changing, making them exceedingly difficult to study. Because algorithms usually rely on massive, complex, machine-learning models, it is possible that even people who design social media recommendation algorithms know little about how they work (Eckles, 2022). Indeed, one of tech industry's biggest open secrets is that "no one quite knows how the algorithms that govern social media actually work" (Fisher, 2022b).

Despite this lack of insider knowledge, researchers have found ways to indirectly study social media algorithms and the type of content they amplify. For example, research has documented what type of content tends to go "viral" (or is widely shared, viewed, or engaged with) online (Brady et al., 2017; Rathje et al., 2021). However, content might go viral because of any number of factors—such as design of social media recommendation algorithms (Brown et al., 2022), user behavior, social media platform design (Munn, 2020), or a combination of all these factors. Other researchers have more specifically examined the effect of algorithms by, for instance, studying the type of content that is shown on algorithmically determined versus nonalgorithmically determined feeds. This has been done through, for instance, collecting social media data via browser extensions (Milli et al., 2023) or accessing internal data from social media companies (Huszár et al., 2022). While this internal data is rarely made available by social media companies to outside researchers, Meta recently collaborated with several academics to test the causal impact moving users from algorithmic to chronological Facebook and Instagram feeds for 3 months in a large-scale, randomized control trial. This shift decreased the amount of time spent on both platforms, and decreased the amount of political and untrustworthy content seen on both platforms. However, this shift did not significantly alter self-reported polarization (Guess et al., 2023). Twitter has also recently made the source code from its recommendation algorithm public; however, it is unclear how much can be interpreted from this source code without more internal data from Twitter (Twitter, 2023).

In the current article, we first review research on what tends to go viral on social media to provide insights into the type of content that is promoted on various platforms. We then recruited a representative sample of U.S. participants to examine lay perceptions of what goes viral on social media and compared these lay perceptions with our review of the research landscape. We also explored what people think should go viral on social media and examined how this differs from what people think actually does go viral.

It is important to examine the differences in people's perceptions of what does go viral versus what should go viral online because some might assume that the content that users frequently engage with simply reflects what users want to see. Indeed, Facebook has argued that their news feed-recommendation algorithms aim to amplify content that people find "valuable" and "meaningful" (Meta Transparency Center, 2023). Even if online content is divisive, the spread of divisive content online may reflect the demands and genuine preferences of social media users. People value engaging with polarizing political debates, being informed about negative events in the world, and expressing outrage about causes they care about. To use the language of economics, the type of content people engage with online might reflect their "revealed preferences," which many economists have traditionally assumed to reflect people's true desires (Beshears et al., 2008; Richter, 1966). Scholars have also noted that online outrage can have many upsides, such as supporting collective action and social change (Spring et al., 2018; cf. Brady & Crockett, 2019). Social media could also simply be accurately reflecting people's real-world feelings and desires. For instance, Facebook has argued that discussions on social media can be "emotional and polarizing because our politics is emotional and polarizing" (Raychoudhury, 2021).

Alternatively, the type of content that goes viral on social media may reflect what is profitable for social media companies because it captures attention rather than what people or society would truly like to see (Van Bavel et al., in press). Indeed, there are many instances in which people's revealed preferences (or their behavior) do not align with their stated preferences (or what they report wanting). For example, 70% of smokers report wanting to quit smoking (Beshears et al., 2008), yet they are often unable to quit because the product itself is addictive and tobacco companies are trying to maximize their own profits rather than the welfare of their consumers. Just as people smoke when they actually do not want to or eat junk food when they would like to eat healthy food, people may engage with content online that they actively do not want to see. Social media platforms are governed by an "attention economy" whereby algorithms amplify content that draws attention and keeps users active on the platform (Simon, 1971; Williams, 2018). Divisive content is good at capturing people's attention (Brady et al., 2020) and thus might be good at keeping people on social media platforms even if people do not actually want to engage with divisive content. Supporting this perspective, research suggests that people do not like expressions of partisan animus from politicians (Costa, 2020; Frimer & Skitka, 2018), and the majority of Americans also report that they are exhausted by the news (Gottfried, 2020) and by political partisanship (Hawkins et al., 2019).

This current research helps adjudicate between these two competing perspectives by examining the potential discrepancy between what goes viral on social media, what people think goes viral on social media, and what people think should go viral on social media in an ideal world. This work has direct implications for improving social media. For instance, if most people report being unhappy with the type of content that tends to be amplified by social media platforms, news feed algorithms could be adjusted to prioritize other outcomes beyond engagement (e.g., accuracy or nuance). To specifically examine how people think social media could be improved, we also measured support for certain solutions, such as making social media algorithms more transparent, giving users more control over algorithms, or regulating social media algorithms. Many of these potential solutions have been discussed widely or introduced in proposed legislation, such as the Filter Bubble Transparency Act (Thune, 2021).

What Goes Viral on Social Media?

A number of studies have identified certain features of content that are related to online virality (Berger & Milkman, 2012; Brady et al., 2017; Rathje et al., 2021; for a summary, see Table 1). Most of these studies are correlational and examined the features of social media posts that are correlated with engagement (e.g., "likes" or "shares"), which was measured as a continuous variable. In other words, most of these studies looked at factors that increase the chances that people like or share a post. For example, several studies have found that social media posts containing moral and emotional words, such as "hate" or "blame," tend to be shared 15% to 20% more in the context of online political debates, among ordinary citizens and political elites, and across several countries (Brady et al., 2017, 2019, 2021; Valenzuela et al., 2017; see also Burton et al., 2021). This may explain why people are more likely to encounter moral violations when using social media than when using any other form of media, such as television or print media (Crockett, 2017).

More broadly, negative emotional content tends to receive more engagement on social media. An analysis of 22,743 A/B tests from the website Upworthy found that news stories were more likely to be clicked on when they contained negativity in the headline (C. E. Robertson et al., 2023). Other work has found that negative emotions such as anger (Fan et al., 2020) and general negative sentiment (Hansen et al., 2011; Schöne et al., 2021) spread further on Twitter than positive or neutral sentiment.

The idea that negativity gains more traction is not unique to social media. For instance, the long-standing colloquialism "if it bleeds, it leads" has referred to the fact that negative news tends to get more attention (Pooley, 1989). In addition, psychologists have noted that humans tend to have a domain-general negativity bias and pay attention to negative information more than positive information (Baumeister et al., 2001; Rozin & Royzman, 2001). However, although the news stories that people viewed or shared on social media tended to be negative, the news stories that were at the top of *The New York Times* "most emailed" list tended to be more positive (Kraft et al., 2020). Furthermore, politicians receive more engagement on posts expressing happiness on Instagram compared with Facebook (Bossetta, 2022), suggesting that different social media platforms may have different audiences and incentive structures that influence whether negativity goes viral. Negativity is also more likely to be shared among public figures as opposed to ordinary users (Schöne et al., 2023), further illustrating the importance of context.

Others have suggested that content that evokes high-arousal emotions, whether these emotions are positive (e.g., awe) or negative (e.g., anger or anxiety), tends to be shared more. For example, *New York Times* articles that evoke both high-arousal positive emotions and high-arousal negative emotions tend to be shared more (Berger & Milkman, 2012), and emotionality predicts the sharing of science articles (Milkman & Berger, 2014). However, whether high-arousal positive emotions versus high-arousal negative emotions go viral may depend on context and culture. For instance, although high-arousal negative emotions such as anger are more "contagious" in the United States, high-arousal positive emotions such as excitement are more contagious in Japan (Hsu et al., 2021).

Divisive or polarizing content—especially about one's political out-group—tends to go viral as well. For instance, posts on Facebook and Twitter from politicians and partisan news media sources received more engagement if they referred to the political out-group (Rathje et al., 2021). For instance, each individual term referring to the political out-group increased the number of shares of a social media post by 67%. Rathje et al. (2021) found that although moral-emotional language and negative language also predicted virality, out-group language was by far the strongest predictor of virality. Posts about the political out-group received high levels of "angry" and "haha" reactions, suggesting outrage and derision. Likewise, expressions of out-party hate from politicians, although less common than expressions of in-group favoritism, received more engagement (Yu et al., 2021). In addition, controversial news (Kim & Ihm, 2020) and expressions of "indignant disagreement" among politicians receive more online engagement (Messing & Weisel, 2017). The most politically extreme politicians also have the most followers on Twitter (Hong & Kim, 2016), perhaps because they share more divisive or controversial content.

Table 1. Factors That Have Been Found to Be Associated With Social Media Virality

Driver of virality	Evidence
Moral-emotional content	• Experiencing-sampling data suggested that people were more likely to encounter moral violations on social media than in person or when consuming other forms of media (Crockett, 2017). • Each additional moral-emotional word (e.g., "hate" or "blame") added to a social media post led to an increase in the predicted number of retweets that post received among Twitter users (Brady et al., 2017) and politicians (Brady et al., 2019). • A meta-analysis of 27 studies found that each additional moral-emotional word added to a post was associated with a 12% increase in engagement (Brady & Van Bavel, 2021a). • Moral outrage, as measured by a machine-learning classifier, predicted increased engagement on Twitter (Brady et al., 2021). • Brady et al. (2020) found in a lab study that moral and emotional words captured attention faster than neutral words and that attentional capture in the lab predicted the virality of tweets online. • Context/caveats: These studies primarily looked at online political conversations.
Negative emotions	• An analysis of A/B tests from the news website Upworthy found that negativity increased news consumption (C. E. Robertson et al., 2023). • Anger was associated with increased virality among weak ties on social media in China (Fan et al., 2020). • Negative news received more engagement on Twitter than positive or neutral news (Hansen et al., 2011). • Negativity spread further on Twitter after both negative and positive political situations (Schöne et al., 2021). • Context/caveats: News stories that people privately viewed or shared on social media tended to be negative, whereas the news stories that were at the top of *The New York Times* "most emailed" list tended to be more positive, illustrating the potential contextual sensitivity of this effect (Kraft et al., 2020). Furthermore, politicians receive more engagement on posts expressing happiness on Instagram compared with Facebook (Bossetta, 2022), illustrating that there might be differences across different social media platforms. In addition, negativity was more likely to be shared among public figures than among ordinary users (Schöne et al., 2022).
High-arousal emotions	• News stories expressing high-arousal emotions were more likely to be on the top of the most-emailed list of *The New York Times* (Berger & Milkman, 2012). • Putting people into a state of arousal causally increased people's willingness to share information (Berger, 2011). • Context/caveats: Although high-arousal negative emotions are more "contagious" in the United States (or more likely to influence the Twitter behavior of followers), high-arousal positive emotions are more contagious in Japan (Hsu et al., 2021). Thus, the effects of high-arousal emotions on social media may depend on context and culture.
Divisive content/ out-group animosity	• Each additional word about the out-group added to a post increased the predicted number of shares by 68%. Out-group language strongly predicted "angry" reactions and "haha" reactions (Rathje et al., 2022). • Out-party hate, although less common than in-party love, received more engagement among U.S. politicians on social media (Yu et al., 2021). • Controversial news received more engagement (Kim & Ihm, 2020). • Expressions of "indignant disagreement" among politicians received more engagement on social media (Messing & Weisel, 2017). • Politically extreme politicians have more followers on Twitter (Hong & Kim, 2016). • Incivility is rising in tweets from American politicians, and this effect is mediated by the positive feedback uncivil tweets receive (Frimer & Skitka, 2018). • People's algorithmic (as opposed to chronological) timelines contained more posts expressing out-party animosity (Milli et al., 2023). • Context/caveats: Most of the above studies examined political elites, news sources, or other primarily political contexts instead of the postings of average users. In addition, conservative voices may be amplified more in general by social media platforms than liberal voices (González-Bailón et al., 2022; Huszár et al., 2022), although it is unclear why this is the case.
False claims (that have been fact-checked)	• Fact-checked true claims spread further than fact-checked false claims (Vosoughi et al., 2018). • Fact-checked false claims about COVID-19 spread more than fact-checked true claims about COVID-19, especially if they contained moral-emotional language (Solovev & Pröllochs, 2022). • Context/caveats: These results depend on the sample of fact-checked claims used, and fact-checkers might be more likely to fact-check already "viral" pieces of misinformation (thus, we use the words "fact-checked false claims" as opposed to misinformation more broadly). In addition, (Altay et al., 2022) found that low-quality news sites are visited rarely, although they receive more engagement on social media than browser visits. Furthermore, Bond and Garretta (2023) found that fact-checked true stories received more engagement than fact-checked false stories on Reddit, illustrating that the affordances of certain social media platforms may influence false news' propensity to go viral.

Note: Categories in this table are highly overlapping (e.g., moral outrage is often directed toward an out-group, is present in misinformation, is negative, and is high arousal).

This online incentive structure, which promotes the creation of divisive content, may have changed how politicians used social media. Indeed, one study found that the incivility in the tweets of American politicians has risen over time (Frimer et al., 2023) and that this increase in incivility was mediated by the amount of positive feedback uncivil tweets received. In other words, politicians who received more likes and retweets for incivility were more likely to post more uncivil content afterward. Algorithms may also play a role in the amplification of divisive content. Recent work found that people's algorithmic (as opposed to chronological) Twitter feeds contained more posts expressing out-party animosity and anger, indicating that certain features of social media algorithms may play a role in these patterns (Milli et al., 2023). Although both conservatives and liberals benefit roughly equally from expressing out-group animosity (Rathje et al., 2021), recent work suggests that conservative voices may be amplified more in general by social media (González-Bailón et al., 2022; Huszár et al., 2022), although it is unclear why this is.

Social media also seems to amplify the spread of misinformation and conspiracy theories (C. E. Robertson et al., 2022; van der Linden et al., 2021). False claims (that have been fact-checked) tend to be shared more than true claims (Juul & Ugander, 2021; Vosoughi et al., 2018), leading to the possibility that some types of misinformation may achieve more virality than true news. Likewise, fact-checked COVID-19 rumors were more likely to go viral than fact-checked true COVID-19 claims, especially if these rumors contained moral-emotional language (Solovev & Pröllochs, 2023). The popularity of certain types of misinformation or conspiracy theories may be related to their novelty (Vosoughi et al., 2018), their emotionality (Fong et al., 2021; Pröllochs et al., 2021), their expressions of moral outrage (McLoughlin et al., 2021), or their tendency to derogate the out-group (Osmundsen et al., 2021; C. E. Robertson et al., 2022). Although visits to untrustworthy sites make up only 2% of overall web traffic, they make up about 14% of Facebook engagement, suggesting that untrustworthy websites may receive more social media engagement than actual web visits (Altay et al., 2022).

However, many studies looked at only true and false claims that had previously been fact-checked, which may potentially bias results toward looking at instances of already viral misinformation (Altay et al., 2023). In addition, the relationship between falsity and virality might differ across different platforms, potentially because of different platform-design choices, algorithms, user bases, and social norms. For instance, a recent study found that fact-checked true claims were more likely to go viral on Reddit than fact-checked false claims, which diverges from the results of prior studies (Bond & Garretta, 2023). Recent work also suggests that people choose to engage with more false and hyper-partisan news than they are exposed to on Google search, suggesting that self-selection (as opposed to algorithmic amplification) may drive some of the engagement with misinformation (R. E. Robertson et al., 2023). The results of this study may also reflect the fact that Google search algorithms surface different types of content than social media algorithms, potentially because Google might have different incentive structures and goals than social media companies (e.g., they may be more focused on promoting relevant content as opposed to attention-grabbing content).

It is difficult to discern universal factors that drive virality, and Table 1 shows that study results often vary by culture, context, and measurement. Future work on social media virality can take advantage of more advanced methods, such as recent advances with large-language models (Rathje, Mirea, et al., 2023; Ziems et al., 2023), to better measure constructs in social media text. Cross-cultural studies could also help examine how the predictors of virality vary across culture and topic. Moreover, it is not always clear how much users drive the spread of certain types of content compared with algorithms. Indeed, these factors are often interwoven—the engagement of hyper-partisan users and political elites might be critical to trigger algorithmic amplification.

Why Does Some Content Go More Viral Than Others? Potential Psychological Processes

What drives people to share and engage with negative, moral, high-arousal, divisive, and false content online and thus make it go viral? Psychologically, this content is good at capturing people's attention, and social media companies prioritize showing content that captures people's attention. Indeed, research suggests that pepole are faster to recognize moral and emotional words than neutral words, and this increased attentional-capture helps explain why posts expressing morality and emotion go viral (Brady et al., 2020). Negativity might also be better at capturing people's attention than positivity because it has long been noted that humans have a negativity bias, or preferentially attend to negativity more than positivity (Baumeister et al., 2001; Rozin & Royzman, 2001). A 17-country study found that exposure to negative news evokes more psychophysiological arousal than exposure to positive news (Soroka et al., 2019), which could explain why people attend to and thus engage more with

negative content. Physiological arousal has also been found to promote sharing behavior. For instance, putting people into a state of physiological arousal led them to report higher intentions to share articles (Berger, 2011), suggesting that arousal might causally influence sharing behavior. These tendencies may be shaped by evolution—negative, high-arousal, group-based, or moral content could all signal some sort of physical or social threat that people need to resolve (Petersen, 2020; Van Bavel et al., in press).

Content that people engage with online might also be good at appealing to people's psychological motivations—such as identity-based motivations, status-seeking motivations, and social-bonding motivations (Brady et al., 2019; Petersen et al., 2021; Pretus et al., 2023). Experiments suggest that sharing moral and emotional language makes people appear to be loyal in-group members (Brady & Van Bavel, 2021b), so people might share this type of content to be looked on positively by their group. People are also socially reinforced on social media (through likes and shares; Brady et al., 2021) and may be motivated to share or engage with moral outrage to get social approval. Sharing content that is critical of one's out-group may also make someone appear to be a loyal in-group member and thus fulfill social-belonging motivations (Brady & Van Bavel, 2021b). People may also engage in status-seeking motivations online (Petersen et al., 2021). Indeed, people tend to share content that reflects well on them (Milkman & Berger, 2014). Status-seeking motivations can explain the sharing of both positive and false or polarizing content. For example, one study found that people high in the trait of "status-driven risk taking" were more likely to share hostile content online (Bor & Petersen, 2022). Misinformation often contains social stimuli, such as gossip (Acerbi, 2019), and mentions of political out-groups (Osmundsen et al., 2021), indicating that misinformation might be particularly good at appealing to social or identity-based motivations. Likewise, viral true news often contains social content (Al-Rawi, 2019). It is unclear how the average person perceives these dynamics on social media. We address this issue in the next section.

Lay Perceptions of Social Media

In the first section of this article, we reviewed what type of content goes viral on social media, finding that prior research suggests that moral emotions, negative emotions, high-arousal emotions, divisive content, and fact-checked false claims are all associated with increased social media "virality." Here, we examine the public's lay perceptions of what goes viral on social media to see whether it mirrors the scientific literature and examine what people think should go viral on social media.

We collected a sample of 511[1] U.S. participants from the survey platform Prolific Academic. This sample was quota-matched to be nationally representative of the general population by age, ethnicity, and gender (n = 511; age: M = 45.69 years, SD = 16.33; male = 246, female = 260, nonbinary = 5; Democrat = 342, Republican = 169). We asked participants to rate what type of content they think goes viral versus what kind of content they think should go viral on social media in a within-subjects experiment. We told people to think of the social media platform they normally use when answering these questions, and we told participants that do not use social media to make their best guess.

Participants were asked to rate on a Likert scale from 1 to 7 (1 = *strongly disagree*, 7 = *strongly agree*) the extent to which they thought the following types of content went viral on social media: content that evokes intense emotions, divisive/polarizing content, moral outrage, misinformation/conspiracy theories, content that evokes negative emotions, people criticizing their enemies, hateful content, content that evokes positive emotions, content that evokes nonintense emotions, accurate information, thoughtful/nuanced content, people praising their allies, and educational content. The first seven of these categories can be thought of as negative or harmful, and the last seven of these categories can be thought of as positive or constructive. These categories were selected on the basis of prior research on social media virality (see Table 1) and a crowdsourcing process in which Twitter, TikTok, Facebook, and LinkedIn users were asked what they think does (and should) go viral.[2] Extended methods are reported in Appendix S1 in the Supplemental Material available online. The anonymized data set, Qualtrics files, and analysis code are available on OSF: https://osf.io/mn9cb. The full question text is reported in Appendix S2 in the Supplemental Material.

There Are Stark Differences Between What People Think Goes Viral Versus Should Go Viral

We conducted paired (within-subjects) t tests to examine the differences between what participants think goes viral versus what they think should go viral. Participants believed that content that evokes intense emotions, divisive/polarizing content, moral outrage, misinformation/conspiracy theories, content that evokes negative emotions, and content featuring people criticizing their enemies goes more viral online than it should ($ps < .001$). Effect sizes ranged from d = 1.76 (for hateful content) to d = 0.27 (for content that evokes

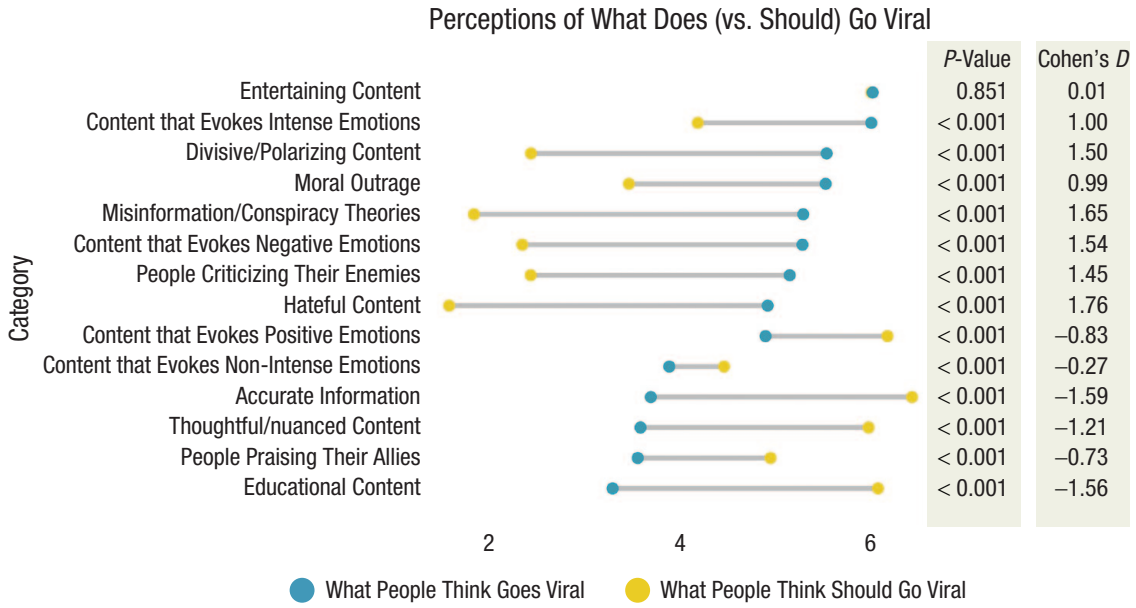

Fig. 1. There were stark differences between the content that people (n = 511) think goes viral (shown in blue) and the content people think should go viral (shown in yellow). Questions were answered on a scale from 1 (*strongly disagree*) to 7 (*strongly agree*); 4 is the exact midpoint. The p-value column represents p values from paired (within-subjects) t tests. There were significant differences for all categories except for entertaining content. Cohen's d effect sizes range from 1.76 (hateful content) to 0.27 (nonintense content) for the significant effects. For full paired t-test results and effect sizes, see Table S1 in the Supplemental Material available online.

intense emotions). In other words, people clearly believe that negative or divisive content goes much more viral online than it should.

In contrast, people reported that content that evokes positive emotions, content that evokes nonintense emotions, accurate information, thoughtful and nuanced content, content featuring people praising their allies, and educational content does not go as viral as much as people think it should (ps < .001). Effect sizes ranged from d = 1.59 (for accurate content) to d = 0.73 (for content praising one's allies). However, people believed the right amount of entertaining content goes viral on social media (p = .851). In other words, people believe that positive content goes much less viral online than they think it should. These differences are plotted visually in Figure 1, and full-paired t test results and effect sizes are shown in Table S1 in the Supplemental Material.

Perceptions of Social Media Virality Are Only Weakly Related to Ideology/Partisanship

Because many discussions about social media are politically contentious, we examined whether responses were related to political ideology and partisanship. Looking at partisan differences is important because if Republicans and Democrats disagree on the kind of content that does (and should) go viral, it may be difficult to come to consensus solutions for improving social media. This analysis allowed us to identify any potential areas of bipartisan consensus.

We conducted all paired t tests separately for Republicans and Democrats. Strikingly, all significant t tests in the main data set are significant and in the same direction when analyzed separately for Republicans or Democrats. The effect sizes for the differences between what does and should go viral are, however, descriptively larger for Democrats than Republicans. For instance, the lowest significant effect size for Republicans was d = 0.24 (for nonintense emotions), and the highest effect size for Republicans was d = 1.41 (for hateful content). The lowest significant effect size for Democrats was d = 0.29 (for nonintense emotions), and the highest effect size for Democrats was d = 2.09 (for misinformation). Yet, in general, as shown in Figure 2 and reported in Tables S2 and S3 in the Supplemental Material, differences between Democrats and Republicans tend to be very small.

There were some small noticeable differences, however, such as Republicans reporting less concern about misinformation going viral, r = .24, 95% confidence interval = [.16, .32], p < .001 (for more details and correlations, see Section S3 in the Supplemental Material). This may reflect the fact that Republicans share more misinformation (Guess et al., 2019), polarization surrounding the definitions for terms such as "misinformation," or conservatives' greater tendency to distrust institutions (Gauchat, 2012). Despite differences such

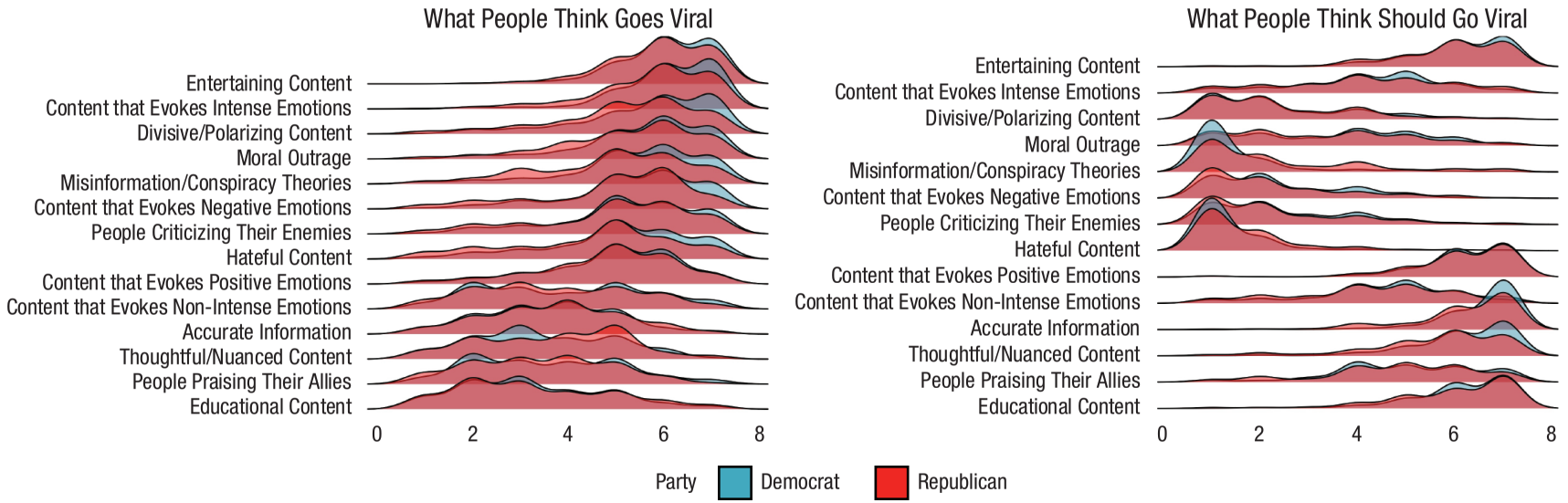

Fig. 2. Distributions of responses for Republicans (shown in red) and Democrats (shown in blue) on each measure. As shown, Republicans and Democrats report very similar beliefs about what does and should go viral on social media.

as these, liberals and conservatives showed striking similarities in their stated preferences.

We also ran a series of correlational analyses (reported in depth in Section S1 and Figs. S1–S4 in the Supplemental Material). Broadly, these analyses found that perceptions of what and should go viral were weakly and inconsistently related to age, ideology, self-reported social media usage, interest in politics, and the number of minutes people spend using social media. There is strong support for greater transparency and control over social media algorithms.

In our nationally representative sample, we also measured support for basic solutions for improving social media content-recommendation algorithms, as shown in Figure 3 (for details, see Appendix S1 in the Supplemental Material). We found that 91.98% of participants answered somewhat agree, agree, or strongly agree to the question "social media platforms should be more transparent about how algorithms work." We also found that 86.69% of participants agreed that users should have more control over how social media algorithms work. A majority of participants also agreed that "social media companies should not use algorithms that select what content to show users" (55.57%) and that "legislation should be passed to regulate social media algorithms" (53.42%). In sum, basic solutions such as greater transparency and control had near universal support in our representative sample. Solutions such as eliminating algorithms and regulating them through legislation were more controversial but still supported by the majority. Because our question about regulation was ambiguously worded and people may be skeptical about the government's ability to write effective policy about complex and rapidly changing technology, future work should explore whether people might be more supportive of some instances of regulation over others. Overall, these results suggest there are some very popular solutions to improving social media that have broad consensus.

Discussion

In a nationally representative survey of Americans, we investigated whether people's perceptions of what goes viral on social media line up with past research and with what they think should go viral on social media. People believe that many forms of divisive content—such as moral outrage, intense content, people criticizing their enemies, and misinformation—all go viral online. These lay beliefs align with past research suggesting that moral outrage (Brady et al., 2021), high-arousal content (Berger & Milkman, 2012), negative content (C. E. Robertson et al., 2023), out-group animosity (Rathje et al., 2021), and misinformation (Vosoughi et al., 2018) often go viral online. Thus, people are aware of the content that tends to be amplified in online social networks according to current research. However, the vast majority of people report that they do not think this type of divisive content should go viral online. Instead, they strongly believe content evoking positive emotions, accurate information, educational content, and thoughtful or nuanced content should go more viral than they currently do. This reveals a stark difference between people's beliefs about how social media is (and how research characterizes social media) and how it should be.

These results were strikingly similar for both Republicans and Democrats and were weakly and inconsistently correlated with other demographic characteristics. Thus, our data reveal a broad consensus in people's belief that social media should amplify very different content than it currently does. These results question the notion that content goes viral purely because most people want that content to go viral. Although some have argued for the potential upsides of moral outrage (Spring et al., 2018) and political polarization (Mac & Silverman, 2021), our results indicate that few people would be happy with social media platforms filled with outrage, misinformation, and divisive content. Social media companies, policymakers, and the general public should be aware of these gaps between how users behave on social media and users' preferences about what social media should be like.

These results introduce a paradox: Why do people engage with content online that they report not wanting to see? There are a number of possible explanations, all of which should be explored in future studies. One possible explanation is that people do not want negative content to go viral but social media algorithms are optimized for amplifying the most attention-grabbing content rather than optimizing for content people truly want to see amplified. This is likely shaped by economic incentives because social media platforms' business model depends on keeping users on their platform for as much time as possible to earn advertising revenue (Fisher, 2022a). Attention may not necessarily be a good measure for the type of content that people want to see. Indeed, some have noted that social media algorithms may play into people's more automatic (or "system one") preferences as opposed to their more carefully considered (or "system two") preferences—especially because they are trained on "mindless" behaviors, such as social media scrolling (Agan et al., 2023). Supporting this perspective, interventions that aimed to disrupt mindless social media (Allcott et al., 2020) or smartphone use (Grüning et al., 2023) led to lasting decreases in use. This has led some to conclude that a portion of social media use can be attributed to self-control failures or habit (Bayer et al., 2022) rather than intentional use.

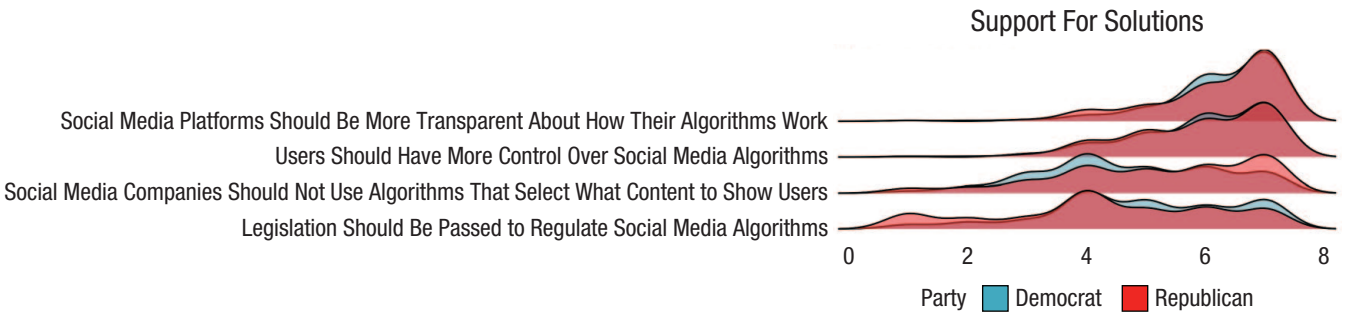

Fig. 3. Democrats (shown in blue) and Republicans (shown in red) showed strong support for greater transparency and greater control over social media algorithms. Support for legislation regulating social media algorithms and support for abolishing social media algorithms altogether were more mixed.

Another possibility is that social media content is heavily influenced by a small number of users who might not be highly representative of the general population. For instance, one report found that 10% of Twitter users produce 80% of tweets (Wojcik & Hughes, 2019). Similar work has found that a small number of Reddit users are responsible for the vast majority of toxic comments on the platform (Kumar et al., 2023), and 0.1% of Twitter users were estimated to be responsible for 80% of misinformation shares (Grinberg et al., 2019). Indeed, one estimate found that just 12 people accounted for 65% of the anti-vaccine misinformation on social media platforms during the COVID-19 pandemic (Nogara et al., 2022). Social media engagement metrics could be highly shaped by a small proportion of influential users that are highly active on social media even if the general population do not agree with the preferences of this small portion of highly active users. People who are hostile offline tend to be hostile online (Bor & Petersen, 2022), and hostile individuals may be highly visible online and drive the spread of divisive content.

These findings have several direct implications for improving social media. For example, rather than just relying on engagement data, social media platforms should pay closer attention to self-report data (like ours) about what people want to see. Indeed, Facebook has tried to integrate self-report data about people's preferences into its algorithm before. For instance, Facebook tested a feature in which they downranked posts in the news algorithm that users rated as "bad for the world." Facebook, however, decided to not implement this feature after discovering it reduced user engagement (Roose et al., 2020). Thus, even though it might be possible to improve the content people see on social media so that it aligns more with their self-reported preferences, social media platforms may be unlikely to do this if it has the prospect to reduce user engagement and undercuts the profits of social media companies. Thus, the interests of individuals and society are likely to be displaced by the economic goals of these companies.

Our results suggest that there are several highly popular solutions for improving social media from outside technology companies through regulation and other changes. The vast majority of people in our sample supported greater transparency for social media algorithms and greater personal control over social media algorithms. One potential way to institute greater transparency is to give users the ability to change the content amplified in their own news feeds and to give independent researchers access to data from social media companies so that the potential harms of social media platforms can be assessed in a nonbiased manner (Persily & Tucker, 2020; Van Bavel, Harris, et al., 2021). This approach appears to have wide, bipartisan support. However, these solutions should be empirically tested because they may come with unexpected downsides (Brady et al., 2023), such as allowing conspiracy-minded individuals to self-select into conspiratorial rabbit holes (R. E. Robertson et al., 2023)

One limitation of this work is that it is based on self-report survey responses, which should not always be taken at face value. It is also possible that people's survey responses are shaped by factors such as social desirability, or a need to present oneself in a positive fashion (Edwards, 1958). Furthermore, people's negative perceptions of what goes viral on social media might be biased by people's tendency to remember negative and emotional experiences (Kensinger & Corkin, 2003). People's responses might also be shaped by algorithm aversion (Dietvorst et al., 2015), or the general tendency for people to distrust algorithms that make errors (even more so than humans that make errors). Although our self-report data come with limitations, they strongly suggest that social media behavior might not reflect the true desires of the population or average user.

Another potential limitation of this study is that we measured whether people approved of constructs such as misinformation, hate speech, and moral outrage in the abstract even though people might not agree on what specifically counts as hate speech or misinformation. However, research suggests that both conservatives and liberals in the United States and Denmark tend to agree on what hate speech is and believe that severe hate speech should be restricted (Rasmussen, 2022). Furthermore, laypeople are reasonably good at differentiating between low- and high-quality news sources (Pennycook & Rand, 2019) and headlines (Rathje, Roozenbeek, et al., 2023). In addition, people often desire basic content-moderation decisions that censor hate speech and misinformation and do not desire unmoderated free speech online (Kozyreva et al., 2023).

Furthermore, although our study speaks to people's stated preferences, it has little to say about how these preferences are related to online behavior. Future research could potentially link social media data to survey data (Rathje, He, et al., 2022) to see whether the same people who think negative content should not go viral also engage with negative content. Although, we note that people's survey responses have been found to be correlated with online news-sharing behavior (Mosleh et al., 2020), suggesting that survey responses might be reasonable proxies for offline behavior.

Future research should also test whether a social media matching people's self-reported ideals would actually be better in practice. As Facebook found, a social media with less content that people think is "bad for the world" may be less engaging overall (Roose et al., 2020). However, building a social media that aligns with people's self-reported values may produce a more sustainable technology that optimizes human flourishing rather than monetizing attention at any cost.

Conclusions

Although Facebook[3] has argued that social media simply reflects "the good, the bad, and the ugly" (Raychoudhury, 2021), it appears that most people think—in line with social-science research—that social media too often amplifies the bad and the ugly. However, people report wanting social media platforms and algorithms to amplify more of the good and less of the bad and the ugly. Specifically, people across the political spectrum think social networks should amplify accurate, educational, and nuanced content as opposed to divisive, false, and negative content. Social media algorithms can be modified to amplify content that is more in line with people's stated preferences instead of simply prioritizing the most engaging content. Furthermore, social media algorithms can be made more transparent and users can be given more control over social media algorithms given that the overwhelming majority of people support these solutions. Although it may be challenging to change the design of social media platforms, the majority of people agree on some key points about what social media should amplify and how it should be improved. We hope our article provides a scientific roadmap for improving social media.

Transparency

Action Editor: Sudeep Bhatia
Editor: Interim Editorial Panel
Declaration of Conflicting Interests
The author(s) declared that there were no conflicts of interest with respect to the authorship or the publication of this article.
Funding
This research was supported by a Gates Cambridge Scholarship awarded to S. Rathje (Grant OPP1144), a grant from the Russell Sage foundation to S. Rathje and J. J. Van Bavel (G-2110-33990), and a grant from the John Templeton Foundation to J. J. Van Bavel.

ORCID iDs

Claire Robertson https://orcid.org/0000-0001-8403-6358
William J. Brady https://orcid.org/0000-0001-6075-5446
Jay J. Van Bavel https://orcid.org/0000-0002-2520-0442

Acknowledgments

We are grateful for helpful feedback from the New York University Social Identity and Morality Lab.

Supplemental Material

Additional supporting information can be found at http://journals.sagepub.com/doi/suppl/10.1177/17456916231190392

Notes

1. Our original planned sample size was 500, although we oversampled slightly to account for unfinished respondents and instances of nonresponse.
2. The following was posted on Twitter: https://twitter.com/jayvanbavel/status/1557385190346989568?s=20&t=0lZoJyTstZK_WvZ6CJUamQ. The following was posted on LinkedIn: https://www.linkedin.com/feed/update/urn:li:activity:6963158807183069184?updateEntityUrn=urn%3Ali%3Afs_feedUpdate%3A%28V2%2Curn%3Ali%3Aactivity%3A6963158807183069184%29. And the following was posted on TikTok: https://www.tiktok.com/@stevepsychology/video/7132223775965236486?is_copy_url=1&is_from_webapp=v1. We took responses to these social media posts in account when designing our questions in addition to considering past research on social media virality.
3. This argument was made in a statement responding to a *Washington Post* article detailing research from Rathje et al. (2021) about how out-group animosity predicts virality on Facebook and Twitter.

References

Acerbi, A. (2019). Cognitive attraction and online misinformation. *Palgrave Communications*, *5*, Article 15. https://doi.org/10.1057/s41599-019-0224-y

Agan, A. Y., Davenport, D., Ludwig, J., & Mullainathan, S. (2023). *Automating automaticity: How the context of human choice affects the extent of algorithmic bias*. National Bureau of Economic Research.

Allcott, H., Braghieri, L., Eichmeyer, S., & Gentzkow, M. (2020). The welfare effects of social media. *American Economic Review*, *110*(3), 629–676.

Al-Rawi, A. (2019). Viral news on social media. *Digital Journalism*, *7*(1), 63–79.

Altay, S., Berriche, M., & Acerbi, A. (2023). Misinformation on misinformation: Conceptual and methodological challenges. *Social Media + Society*, *9*(1). https://doi.org/10.1177/20563051221150412

Altay, S., Nielsen, R. K., & Fletcher, R. (2022). Quantifying the "infodemic:" People turned to trustworthy news outlets during the 2020 coronavirus pandemic. *Journal of Quantitative Description: Digital Media*, *2*. https://doi.org/10.51685/jqd.2022.020

Bak-Coleman, J. B., Alfano, M., Barfuss, W., Bergstrom, C. T., Centeno, M. A., Couzin, I. D., Donges, J. F., Galesic, M., Gersick, A. S., & Jacquet, J. (2021). Stewardship of global collective behavior. *Proceedings of the National Academy of Sciences, USA*, *118*(27), Article e2025764118. https://doi.org/10.1073/pnas.2025764118

Bakshy, E., Messing, S., & Adamic, L. A. (2015). Exposure to ideologically diverse news and opinion on Facebook. *Science*, *348*(6239), 1130–1132.

Bayer, J. B., Anderson, I. A., & Tokunaga, R. S. (2022). Building and breaking social media habits. *Current Opinion in Psychology*, *45*, 101303.

Baumeister, R. F., Bratslavsky, E., Finkenauer, C., & Vohs, K. D. (2001). Bad is stronger than good. *Review of General Psychology*, *5*(4), 323–370.

Berger, J. (2011). Arousal increases social transmission of information. *Psychological Science*, *22*(7), 891–893.

Berger, J., & Milkman, K. L. (2012). What makes online content viral? *Journal of Marketing Research*, *49*(2), 192–205.

Beshears, J., Choi, J. J., Laibson, D., & Madrian, B. C. (2008). How are preferences revealed? *Journal of Public Economics*, *92*(8–9), 1787–1794.

Bond, R. M., & Garretta, R. K. (2023). Engagement with fact-checked posts on Reddit. *PNAS Nexus*, *2*(3), Article pgad018. https://doi.org/10.1093/pnasnexus/pgad018

Bor, A., & Petersen, M. B. (2022). The psychology of online political hostility: A comprehensive, cross-national test of the mismatch hypothesis. *American Political Science Review*, *116*(1), 1–18.

Bossetta, S. (2022). Cross-platform emotions and audience engagement in social media political campaigning: Comparing candidates' Facebook and Instagram images in the 2020 US election. *Political Communication*, *40*, 48–68.

Brady, W. J., & Crockett, M. J. (2019). How effective is online outrage? *Trends in Cognitive Sciences*, *23*(2), 79–80.

Brady, W. J., Crockett, M. J., & Van Bavel, J. J. (2019). The MAD model of moral contagion: The role of motivation, attention, and design in the spread of moralized content online. *Perspectives on Psychological Science*, *15*, 978–1010.

Brady, W. J., Gantman, A. P., & Van Bavel, J. J. (2020). Attentional capture helps explain why moral and emotional content go viral. *Journal of Experimental Psychology: General*, *149*(4), 746–756.

Brady, W. J., Jackson, J. C., Lindström, B., & Crockett, M. J. (2023). *Algorithm-mediated social learning in online social networks*. OSF Preprints. https://doi.org/10.31219/osf.io/yw5ah

Brady, W. J., McLoughlin, K., Doan, T. N., & Crockett, M. J. (2021). How social learning amplifies moral outrage expression in online social networks. *Science Advances*, *7*(33), Article eabe5641. https://doi.org/10.1126/sciadv.abe5641

Brady, W. J., & Van Bavel, J. J. (2021a). *Estimating the effect size of moral contagion in online networks: A pre-registered replication and meta-analysis*. OSF preprint. https://osf.io/s4w2x

Brady, W. J., & Van Bavel, J. J. (2021b). *Social identity shapes antecedents and functional outcomes of moral emotion expression in online networks*. OSF preprint. https://osf.io/dgt6u/

Brady, W. J., Wills, J. A., Jost, J. T., Tucker, J. A., & Van Bavel, J. J. (2017). Emotion shapes the diffusion of moralized content in social networks. *Proceedings of the National Academy of Sciences, USA*, *114*(28), 7313–7318.

Brown, M. A., Bisbee, J., Lai, A., Bonneau, R., Nagler, J., & Tucker, J. A. (2022). *Echo chambers, rabbit holes, and algorithmic bias: How YouTube recommends content to real users*. SSRN. https://ssrn.com/abstract=4114905

Burton, J. W., Cruz, N., & Hahn, U. (2021). Reconsidering evidence of moral contagion in online social networks. *Nature Human Behaviour*, *5*(12), 1629–1635.

Costa, M. (2020). Ideology, not affect: What Americans want from political representation. *American Journal of Political Science*, *65*, 342–358.

Crockett, M. J. (2017). Moral outrage in the digital age. *Nature Human Behaviour*, *1*(11), 769–771.

C-Span. (2021). *Senate hearing on social media algorithms*. https://www.c-span.org/video/?511248-1/senate-hearing-social-media-algorithms

Dietvorst, B. J., Simmons, J. P., & Massey, C. (2015). Algorithm aversion: People erroneously avoid algorithms after seeing them err. *Journal of Experimental Psychology: General*, *144*(1), 114.

Eckles, D. (2022). *Algorithmic transparency and assessing effects of algorithmic ranking*. Testimony before the Senate Subcommittee on Communications, Media, and Broadband. https://osf.io/preprints/socarxiv/c8za6/

Edwards, A. L. (1958). The social desirability variable in personality assessment and research. *Academic Medicine*, *33*(8), 610–611.

Fan, R., Xu, K., & Zhao, J. (2020). *Weak ties strengthen anger contagion in social media*. arXiv. https://doi.org/10.48550/arXiv.2005.01924

Fisher, M. (2022a). *The chaos machine: The inside story of how social media rewired our minds and our world*. Little, Brown, and Company.

Fisher, M. (2022b, September 1). How social media influences our behavior, and vice versa. *The New York Times*. https://www.nytimes.com/2022/09/01/books/review/max-fisher-chaos-machine.html

Fong, A., Roozenbeek, J., Goldwert, D., Rathje, S., & van der Linden, S. (2021). The language of conspiracy: A psychological analysis of speech used by conspiracy theorists and their followers on Twitter. *Group Processes & Intergroup Relations*, *24*(4), 606–623.

Frimer, J. A., & Skitka, L. J. (2018). The Montagu Principle: Incivility decreases politicians' public approval, even with their political base. *Journal of Personality and Social Psychology*, *115*(5), 845–866.

Frimer, J. A., Aujla, H., Feinberg, M., Skitka, L. J., Aquino, K., Eichstaedt, J. C., & Willer, R. (2023). Incivility is rising among American politicians on Twitter. *Social Psychological and Personality Science*, *14*(2), 259–269.

Gauchat, G. (2012). Politicization of science in the public sphere: A study of public trust in the United States, 1974 to 2010. *American Sociological Review*, *77*(2), 167–187.

González-Bailón, S., d'Andrea, V., Freelon, D., & De Domenico, M. (2022). The advantage of the right in social media news sharing. *PNAS Nexus*, *1*(3), Article pgac137. https://doi.org/10.1093/pnasnexus/pgac137

Gottfried, J. (2020). *Americans' news fatigue isn't going away–about two-thirds still feel worn out*. Pew Research Center. https://www.pewresearch.org/short-reads/2020/02/26/almost-seven-in-ten-americans-have-news-fatigue-more-among-republicans/

Grinberg, N., Joseph, K., Friedland, L., Swire-Thompson, B., & Lazer, D. (2019). Fake news on Twitter during the 2016 US presidential election. *Science*, *363*(6425), 374–378.

Grüning, D. J., Riedel, F., & Lorenz-Spreen, P. (2023). Directing smartphone use through the self-nudge app one sec. *Proceedings of the National Academy of Sciences*, *120*(8), e2213114120.

Guess, A. M., Malhotra, N., Pan, J., Barberá, P., Allcott, H., Brown, T., Crespo-Tenorio, A., Dimmery, D., Freelon, D., Gentzkow, M., González-Bailón, S., Kennedy, E., Kim, Y. M., Lazer, D., Moehler, D., Nyhan, B., Rivera, C. V., Settle, J., Thomas, D. R., … Tucker, J. A. (2023). How do social media feed algorithms affect attitudes and behavior in an election campaign? *Science*, *381*(6656), 398–404.

Guess, A., Nagler, J., & Tucker, J. (2019). Less than you think: Prevalence and predictors of fake news dissemination on Facebook. *Science Advances*, *5*(1), eaau4586.

Haidt, J. (2022, April 11). Why the past 10 years of American Life have been uniquely stupid. *The Atlantic*. https://www.theatlantic.com/magazine/archive/2022/05/social-media-democracy-trust-babel/629369/

Hansen, L. K., Arvidsson, A., Nielsen, F. A., Colleoni, E., & Etter, M. (2011). Good friends, bad news - Affect and virality in Twitter. In J. J. Park, L. T. Yang, & C. Lee (Eds.), *Future information technology* (pp. 34–43). Springer.

Harris, E., Rathje, S., Robertson, C., & Van Bavel, J. J. (2023). The SPIR model of social media and polarization: Exploring the role of selection, platform design, incentives, and real-world context. *International Journal of Communication*.

Hawkins, S., Yudkin, D., Juan-Torres, M., & Dixon, T. (2019). *Hidden tribes: A study of America's polarized landscape*. More in Common. https://hiddentribes.us/media/qfpekz4g/hidden_tribes_report.pdf

Hong, S., & Kim, S. H. (2016). Political polarization on Twitter: Implications for the use of social media in digital governments. *Government Information Quarterly*, *33*(4), 777–782.

Hsu, T. W., Niiya, Y., Thelwall, M., Ko, M., Knutson, B., & Tsai, J. L. (2021). Social media users produce more affect that supports cultural values, but are more influenced by affect that violates cultural values. *Journal of Personality and Social Psychology*, *121*, 969–983.

Huszár, F., Ktena, S. I., O'Brien, C., Belli, L., Schlaikjer, A., & Hardt, M. (2022). Algorithmic amplification of politics on Twitter. *Proceedings of the National Academy of Sciences, USA*, *119*(1), Article e2025334119. https://doi.org/10.1073/pnas.2025334119

Juul, J. L., & Ugander, J. (2021). Comparing information diffusion mechanisms by matching on cascade size. *Proceedings of the National Academy of Sciences, USA*, *118*(46), Article e2100786118. https://doi.org/10.1073/pnas.2100786118

Kensinger, E. A., & Corkin, S. (2003). Memory enhancement for emotional words: Are emotional words more vividly remembered than neutral words?. *Memory & Cognition*, *31*(8), 1169–1180.

Kim, E., & Ihm, J. (2020). More than virality: Online sharing of controversial news with activated audience. *Journalism & Mass Communication Quarterly*, *97*(1), 118–140.

Kozyreva, A., Herzog, S. M., Lewandowsky, S., Hertwig, R., Lorenz-Spreen, P., Leiser, M., & Reifler, J. (2023). Resolving content moderation dilemmas between free speech and harmful misinformation. *Proceedings of the National Academy of Sciences, USA*, *120*(7), Article e2210666120. https://doi.org/10.1073/pnas.2210666120

Kraft, P. W., Krupnikov, Y., Milita, K., Ryan, J. B., & Soroka, S. (2020). Social media and the changing information environment: Sentiment differences in read versus recirculated news content. *Public Opinion Quarterly*, *84*(Suppl. 1), 195–215.

Kumar, D., Hancock, J., Thomas, K., & Durumeric, Z. (2023, May 1–5). Understanding the behaviors of toxic accounts on Reddit. In *Proceedings of the ACM Web Conference 2023 (WWW '23)*, Austin, TX, USA (p. 12). Association for Computing Machinery. https://doi.org/10.1145/3543507.3583522

Mac, R., & Silverman, C. (2021, March 13). Facebook created an employee "playbook" to respond to accusations of polarization. *BuzzFeed News*. https://www.buzzfeednews.com/article/ryanmac/facebook-execs-polarization-playbook

McLoughlin, K. L., Brady, W. J., & Crockett, M. J. (2021). The role of moral outrage in the spread of misinformation. *TMS Proceedings*. https://tmb.apaopen.org/pub/nwpo88ls

Messing, S., & Weisel, R. (2017). *Partisan conflict and congressional outreach*. Pew Research Center. https://www.pewresearch.org/politics/2017/02/23/partisan-conflict-and-congressional-outreach/

Meta Transparency Center. (2023). *Our approach to ranking*. Transparency Center. https://transparency.fb.com/features/ranking-and-content/

Milkman, K. L., & Berger, J. (2014). The science of sharing and the sharing of science. *Proceedings of the National Academy of Sciences, USA*, *111*(Suppl. 4), 13642–13649.

Milli, S., Carroll, M., Pandey, S., Wang, Y., & Dragan, A. D. (2023). *Twitter's algorithm: Amplifying anger, animosity, and affective polarization*. arXiv. https://doi.org/10.48550/arXiv.2305.16941

Mosleh, M., Pennycook, G., & Rand, D. G. (2020). Self-reported willingness to share political news articles in online surveys correlates with actual sharing on Twitter. *PLOS ONE*, *15*(2), Article e0228882. https://doi.org/10.1371/journal.pone.0228882

Munn, L. (2020). Angry by design: toxic communication and technical architectures. *Humanities and Social Sciences Communications*, *7*(1), 1–11.

Narayanan, A. (2023). *Understanding social media recommendation algorithms*. Knight First Amendment Institute.

Nogara, G., Vishnuprasad, P. S., Cardoso, F., Ayoub, O., Giordano, S., & Luceri, L. (2022). The disinformation dozen: An exploratory analysis of covid-19 disinformation proliferation on Twitter. In *14th ACM Web Science Conference* (pp. 348–358). Association for Computing Machinery.

Osmundsen, M., Bor, A., Vahlstrup, P. B., Bechmann, A., & Petersen, M. B. (2021). Partisan polarization is the primary psychological motivation behind political fake news sharing on Twitter. *American Political Science Review*, *115*, 999–1015.

Pennycook, G., & Rand, D. G. (2019). Fighting misinformation on social media using crowdsourced judgments of news source quality. *Proceedings of the National Academy of Sciences, USA*, *116*(7), 2521–2526.

Persily, N., & Tucker, J. A. (2020). *Social media and democracy: The state of the field, prospects for reform*. Cambridge University Press.

Petersen, M. B. (2020). The evolutionary psychology of mass mobilization: How disinformation and demagogues coordinate rather than manipulate. *Current Opinion in Psychology*, *35*, 71–75.

Petersen, M. B., Osmundsen, M., & Bor, A. (2021). Beyond populism: The psychology of status-seeking and extreme political discontent. In J. P. Forgas, W. D. Crano, & K. Fiedler (Eds.), *The psychology of populism* (pp. 62–80). Routledge.

Pooley, E. (1989). Grins, gore, and videotape: The trouble with local TV news. *New York Magazine*, *22*(40), 36–44.

Pretus, C., Servin-Barthet, C., Harris, E. A., Brady, W. J., Vilarroya, O., & Van Bavel, J. J. (2023). The role of political devotion in sharing partisan misinformation. *Journal of Experimental Psychology: General*. Advance online publication. https://doi.org/10.1037/xge0001436

Pröllochs, N., Bär, D., & Feuerriegel, S. (2021). Emotions explain differences in the diffusion of true vs. False social media rumors. *Scientific Reports*, *11*, Article 22721. https://doi.org/10.1038/s41598-021-01813-2

Rasmussen, J. (2022). *The hate speech consensus: Severity shapes Americans' and Danes' preferences for restricting hate speech*. Aarhus University. https://psyarxiv. com/j4nuc

Rathje, S., He, J. K., Roozenbeek, J., Van Bavel, J. J., & van der Linden, S. (2022). Social media behavior is associated with vaccine hesitancy. *PNAS Nexus*, *1*(4), Article pgac207. https://doi.org/10.1093/pnasnexus/pgac207

Rathje, S., Mirea, D.-M., Sucholutsky, I., Marjieh, R., Robertson, C., & Van Bavel, J. J. (2023). *GPT is an effective tool for multilingual psychological text analysis*.

Rathje, S., Roozenbeek, J., Van Bavel, J. J., & van der Linden, S. (2023). Accuracy and social motivations shape judgements of (mis) information. *Nature Human Behaviour*, *7*, 892–903.

Rathje, S., Van Bavel, J. J., & van der Linden, S. (2021). Outgroup animosity drives engagement on social media. *Proceedings of the National Academy of Sciences, USA*, *118*(26), Article e2024292118. https://doi.org/10.1073/pnas.2024292118

Raychoudhury, P. (2021, July 14). *Extremism is bad for our business and what we are doing about it*. Facebook Research. https://research.fb.com/blog/2021/07/extremism-is-bad-for-our-business-and-what-we-are-doing-about-it/

Richter, M. K. (1966). Revealed preference theory. *Econometrica: Journal of the Econometric Society*, *34*, 635–645.

Robertson, C. E., Pretus, C., Rathje, S., Harris, E., & Van Bavel, J. J. (2022). How social identity shapes conspiratorial belief. *Current Opinion in Psychology*, *47*, Article 101423. https://doi.org/10.1016/j.copsyc.2022.101423

Robertson, C. E., Pröllochs, N., Schwarzenegger, K., Parnamets, P., Van Bavel, J. J., & Feuerriegel, S. (2023). Negativity drives online news consumption. *Nature Human Behaviour*, *7*(5), 812–822. https://doi.org/10.1038/s41562-023-01538-4

Robertson, R. E., Green, J., Ruck, D. J., Ognyanova, K., Wilson, C., & Lazer, D. (2023). Users choose to engage with more partisan news than they are exposed to on Google search. *Nature*, *618*, 342–348.

Roose, K., Isaac, M., & Frenkel, S. (2020, November 24). Facebook struggles to balance civility and growth. *The New York Times*. https://www.nytimes.com/2020/11/24/technology/facebook-election-misinformation.html

Rozin, P., & Royzman, E. B. (2001). Negativity bias, negativity dominance, and contagion. *Personality and Social Psychology Review*, *5*(4), 296–320.

Schöne, J., Garcia, D., Parkinson, B., & Goldenberg, A. (2023). Negative expressions are shared more on Twitter for public figures than for ordinary users. *PNAS Nexus*, *2*(7), Article pgad219. https://doi.org/10.1093/pnasnexus/pgad219

Schöne, J. P., Parkinson, B., & Goldenberg, A. (2021). Negativity spreads more than positivity on Twitter after both positive and negative political situations. *Affective Science*, *2*(4), 379–390.

Simon, H. A. (1971). Designing organizations for an information-rich world. In M. Greenberg (Ed.), *Computers, communications, and the public interest* (pp. 36–72). The Johns Hopkins Press.

Solovev, K., & Pröllochs, N. (2022, April 25–29). Moral emotions shape the virality of COVID-19 misinformation

on social media. In F. Laforest, R. Troncy, E. Simperl, D. Agarwal, A. Gionis, I. Herman, L. Médini (Eds.), *Proceedings of the ACM web conference 2022*, Virtual Event, Lyon, France, (pp. 3706–3717). Association for Computing Machinery.

Solovev, K., & Pröllochs, N. (2023). Moralized language predicts hate speech on social media. *PNAS Nexus*, *2*(1), pgac281.

Soroka, S., Fournier, P., & Nir, L. (2019). Cross-national evidence of a negativity bias in psychophysiological reactions to news. *Proceedings of the National Academy of Sciences, USA*, *116*(38), 18888–18892. https://doi.org/10.1073/pnas.1908369116

Spring, V. L., Cameron, C. D., & Cikara, M. (2018). The upside of outrage. *Trends in Cognitive Sciences*, *22*(12), 1067–1069.

Statista. (2022a). *Daily social media usage worldwide*. https://www.statista.com/statistics/433871/daily-social-media-usage-worldwide/

Statista. (2022b). *Number of social network users worldwide from 2017 to 2025*. https://www.statista.com/statistics/278414/number-of-worldwide-social-network-users/

Thune, J. (2021). *S.2024–117th Congress (2021-2022): Filter Bubble Transparency Act* (2021/2022) [Legislation]. https://www.congress.gov/bill/117th-congress/senate-bill/2024

Twitter. (2023). *Twitter's recommendation algorithm*. https://blog.twitter.com/engineering/en_us/topics/open-source/2023/twitter-recommendation-algorithm

Valenzuela, S., Piña, M., & Ramírez, J. (2017). Behavioral effects of framing on social media users: How conflict, economic, human interest, and morality frames drive news sharing. *Journal of Communication*, *67*(5), 803–826. https://doi.org/10.1111/jcom.12325

Van Bavel, J. J., Harris, E. A., Pärnamets, P., Rathje, S., Doell, K. C., & Tucker, J. A. (2021). Political psychology in the digital (mis)information age: A model of news belief and sharing. *Social Issues and Policy Review*, *15*(1), 84–113.

Van Bavel, J. J., Rathje, S., Harris, E., Robertson, C., & Sternisko, A. (2021). How social media shapes polarization. *Trends in Cognitive Sciences*, *25*, 913–916.

Van Bavel, J. J., Robertson, C., del Rosario, K., Rasmussen, J., & Rathje, S. (in press). Social media and morality. *Annual Review of Psychology*.

van der Linden, S., Roozenbeek, J., Maertens, R., Basol, M., Kácha, O., Rathje, S., & Traberg, C. S. (2021). How can psychological science help counter the spread of fake news? *The Spanish Journal of Psychology*, *24*, Article e25. https://doi.org/10.1017/SJP.2021.23

Vosoughi, S., Roy, D., & Aral, S. (2018). The spread of true and false news online. *Science*, *359*(6380), 1146–1151. https://doi.org/10.1126/science.aap9559

Williams, J. (2018). *Stand out of our light: Freedom and resistance in the attention economy*. Cambridge University Press.

Wojcik, S., & Hughes, A. (2019). Sizing up Twitter users. PEW Research Center. https://www.pewresearch.org/internet/2019/04/24/sizing-up-twitter-users/

Yu, X., Wojcieszak, M., & Casas, A. (2021). Affective polarization on social media: In-party love among American politicians, greater engagement with out-party hate among ordinary users. *Political Behaviour*, 1–26. https://doi.org/10.31219/osf.io%2Frhmb9

Ziems, C., Held, W., Shaikh, O., Chen, J., Zhang, Z., & Yang, D. (2023). *Can large language models transform computational social science?* arXiv. https://arxiv.org/abs/2305.03514