



Social Drivers and Algorithmic Mechanisms on Digital Media

Hannah Metzler^{1,2,3}  and David Garcia^{2,4} 

¹Center for Medical Data Science, Medical University of Vienna; ²Complexity Science Hub Vienna, Austria;

³Institute for Globally Distributed Open Research and Education, Vienna, Austria; and ⁴Department of Politics and Public Administration, University of Konstanz

Abstract

On digital media, algorithms that process data and recommend content have become ubiquitous. Their fast and barely regulated adoption has raised concerns about their role in well-being both at the individual and collective levels. Algorithmic mechanisms on digital media are powered by social drivers, creating a feedback loop that complicates research to disentangle the role of algorithms and already existing social phenomena. Our brief overview of the current evidence on how algorithms affect well-being, misinformation, and polarization suggests that the role of algorithms in these phenomena is far from straightforward and that substantial further empirical research is needed. Existing evidence suggests that algorithms mostly reinforce existing social drivers, a finding that stresses the importance of reflecting on algorithms in the larger societal context that encompasses individualism, populist politics, and climate change. We present concrete ideas and research questions to improve algorithms on digital platforms and to investigate their role in current problems and potential solutions. Finally, we discuss how the current shift from social media to more algorithmically curated media brings both risks and opportunities if algorithms are designed for individual and societal flourishing rather than short-term profit.

Keywords

algorithms, digital media, well-being, polarization

Introduction

Algorithms on digital media platforms clearly provide value, as reflected in the wealth they generate for the companies using them. They highlight relevant posts, news, people, and groups and have become necessary to reduce information overload (Narayanan, 2023b). The central role of algorithms in several types of online interaction has raised concerns that they may fuel large psychological and societal issues, specifically mental-health issues and political polarization. First, algorithms could contribute to increasing depression, anxiety, loneliness, body dissatisfaction, and even suicides by facilitating unhealthy social comparisons, addiction, poor sleep, cyberbullying, and harassment, especially in teenagers and girls (Ritchie, 2021; Twenge, 2020; Twenge et al., 2022). Second, they may fuel hate speech, fake news, and polarization by promoting extremist and populist content or by using algorithmic filter bubbles (Bliss et al., 2020; Lewis-Kraus, 2022).

Widespread usage of digital platforms and continuous interaction with algorithms could indeed affect individual and societal well-being in important ways (Büchi, 2021). However, direct evidence supporting these conclusions remains scarce (Bail, 2021; Ferguson, 2021; Sumpter, 2018). Researchers have investigated the potential effects of digital media and its algorithms using self-reports of social-media usage and digital traces of online behavior. Yet most existing studies cannot distinguish the effects of algorithms from the general use of digital media, social behavioral patterns, or large societal changes because their traces are intermingled in these types of data (Salganik, 2019).

We aim to illustrate how algorithmic mechanisms on digital media build on societal forces and how, in combination, they influence desirable and undesirable

Corresponding Author:

Hannah Metzler, Complexity Science Hub Vienna

Email: metzler@csh.ac.at

outcomes at the individual and collective levels. We focus on algorithms that determine how data is processed and what content is presented to users on digital media, rather than the more general concept of algorithms as a set of steps to perform a task. We describe the social drivers of online interaction and how algorithms might change these dynamics. We then summarize evidence and research gaps on social, algorithmic, and societal contributions for two sample topic areas: well-being and mental health at the individual level and polarization and misinformation at the collective level. Finally, we outline open questions and research opportunities to understand whether we can improve algorithms to contribute to human flourishing, and if so, how.

Social Drivers Underlying Individual and Group Behavior on Digital Media

Social media and its algorithms are so successful because they build on ancient human needs for connection and status (Brady et al., 2020; Meshi et al., 2015; Nadkarni & Hofmann, 2012). The twin desires to get along and to get ahead are basic human motives that were crucial for survival in our ancestral environment in social groups (Cummins, 2005; Sapolsky, 2005). Status and connection are pivotal to explaining social behavior (Abele & Wojciszke, 2013; Fiske et al., 2007; Gurtman, 2009) across many domains. Examples include face perception (Todorov et al., 2008); judgments and stereotypes (Fiske et al., 2007); relationships between individuals (Schafer & Schiller, 2018) or groups (Nadler, 2016); and cultural differences in religiosity and prosociality (Gebauer et al., 2013, 2014).

Connection and status motives also strongly shape social interaction and interaction with algorithms on digital media (Eslinger et al., 2021; Meshi et al., 2015). The need for connection motivates participation in the lives of friends, interest in peer groups, self-disclosure of one's own experiences, and renewal of old connections, as well as the pursuit of new connections, dating partners, and groups to join. Status motives influence how we broadcast content, present ourselves, receive social feedback, and observe and evaluate what others share (Burke et al., 2020; Meshi et al., 2015). Studies show that humans are susceptible to social feedback on digital media: Likes influence how quickly people post again (Lindström et al., 2021), whether or not they consider a post successful (Carr et al., 2018), and how happy, self-assured, and popular they feel after posting status updates (Rosenthal-von der Pütten et al., 2019; Zell & Moeller, 2018).

Algorithmic Mechanisms and Other Platform Influences

All of these social motives are also ubiquitous in offline contexts, so how do algorithms and platform features change social interaction? Algorithms constantly adapt to changes in human behavior and are updated as behavior on platforms, and societal discussion about them, evolves. Humans, in turn, strive for the attention and recognition of others to gain social status, which motivates them to reproduce the behaviors that algorithms reward. The eventually observable behavior thus results from interactive feedback loops between human behavior, algorithms, and other platform features (Narayanan, 2023b; Tsvetkova et al., 2017; Wagner et al., 2021). Algorithms are designed to optimize certain metrics, which are used to rank content in user feeds or to suggest relevant accounts. Yet these optimization metrics are usually chosen to maximize the profits of corporations and advertisers (Bak-Coleman et al., 2021; Narayanan, 2023b) rather to bring about psychological and societal benefits.

The history of the Facebook algorithm illustrates how changes in metrics can affect social behavior (Merrill & Oremus, 2021; Oremus et al., 2021; Wallaroo Media, 2022), but also how little control engineers actually have over eventual outcomes within such complex emerging feedback loops (Narayanan, 2023a). In its early days, the algorithm optimized for the number of clicks, likes, and comments and the total time spent on Facebook. As users and companies learned to game the algorithm, clickbait emerged. To counter this, Facebook started maximizing the time users spent reading or watching content in 2015, which led to more passive use, more professionally produced content, less social interaction, and less sharing of original content. Because of user complaints and decreases in interaction, Facebook adapted the algorithm to encourage more "meaningful social interactions." It boosted posts by friends and family, boosted highly commented posts, and weighted the emotional-reaction buttons much more than likes. This became problematic, as the most heavily commented posts also made people the angriest. Strongly weighting angry reactions may have favored toxic and low-quality news content. Responding to complaints, Facebook gradually reduced the angry emoji weight from five times the weight of likes in 2018 to weight 0 in 2020.

Most current digital-media algorithms strongly optimize for engagement (Narayanan, 2023b; Nikolov et al., 2019). However, social success and quality of content are only partly correlated (Salganik et al., 2006).

Optimizing for popularity even seems to lower the overall quality of content (Ciampaglia et al., 2018). Engagement metrics primarily promote content that fits immediate human social, affective, and cognitive preferences and biases rather than quality content (Menczer, 2021) or long-term goals and values (Narayanan, 2023b). For instance, users are more likely to like and share low-quality content that others have already liked (Avram et al., 2020). Popularity metrics can also be gamed with inauthentic behavior, including bots, organized trolls, and fake-account networks (Pacheco et al., 2021; Sen et al., 2018). Furthermore, the interval at which an algorithm rewards behavior influences how quickly it is repeated (Lindström et al., 2021). Especially variable and unpredictable rewards, such as those on platforms with strong virality, seem more addictive (Munger, 2020b).

Other relevant platform features beyond algorithms include the vastly enlarged scale of digital compared with offline social networks. This increases audience size and magnifies differences in the influence and social status of individual users (Bak-Coleman et al., 2021). It also creates unprecedented opportunities for building connections, earning recognition, and observing others, thereby supercharging motives of social status and connection (Bail, 2021; Bak-Coleman et al., 2021; Brady et al., 2020). This increases potentially available social feedback, which notifications, likes, shares, and comments make easily accessible, immediate, and quantifiable (Brady et al., 2020). Finally, algorithm recommendations may have changed the structure of networks, increasing the frequency of triangles (Salganik, 2019; Ugander et al., 2011) and enabling interaction between distant individuals (Bak-Coleman et al., 2021).

These conditions make status comparisons particularly likely and painful (Brady et al., 2020; Munger, 2017). In large online networks, personal information about individuals is limited, whereas information about social groups is still visible. Social groups thus become the main relationships in the network, making social identities highly salient (Brady et al., 2020). Finally, larger networks mean that one encounters a larger number and diversity of individuals and opinions than in real life (Gentzkow & Shapiro, 2011; Guess et al., 2018). Digital media thus allow people to observe many (potentially very different) others and offer people unprecedented freedom to present themselves, get feedback, and adapt; they have become a central tool people use to understand themselves, understand others, and understand which groups they themselves belong to (Bail, 2021; Brady et al., 2020). As contexts in which status and groups are highly salient, digital media have become places where different groups

compete for status and in-group and out-group dynamics crucially determine behavior.

When audiences are larger and more public than private, competition between groups becomes particularly strong, as discussions between political groups show, for example, on Twitter. Similarly, YouTube's reputation for toxic comments could be linked to the extremely broad demographics of its users, leading to more conflict, and to the algorithm weighting up-votes and down-votes equally (Munn, 2020). On Facebook and especially Instagram, self-presentation is more central than group competition (Cingel et al., 2022; Midgley, 2019; Storr, 2018), leading, for example, to microcelebrities (Marwick, 2015). The TikTok algorithm guarantees a small number of views for everybody, which lowers the barriers to entry compared with the more hierarchical social networks on social media. It further makes it hard to predict which TikTok videos will go viral, which could explain long unwanted scrolling experiences, more passive watching, and less social interaction overall (Munger, 2020b).

Social Drivers and Algorithmic Mechanisms Influencing Individual Well-Being

Mainstream discourse and parts of the scientific literature often fail to distinguish between social drivers, algorithmic mechanisms, and societal context because they fail to derive causal insights from correlation, present results limited to single studies and countries, take self-reports at face value, or omit the fact that effect sizes are small (see Cavanagh, 2017; Dienlin & Johannes, 2020; Orben & Przybylski, 2019b; Ritchie, 2021; Sumpter, 2018).

Concerns are often raised about algorithms on digital media harming mental health by fueling addiction, bad sleep, and social comparison (Smyth & Murphy, 2023), or about algorithms purposefully manipulating user mood (Booth, 2014). This debate usually conflates the time spent using social media with algorithmic effects. Only one study pinpointed algorithmic effects, finding that reducing positive posts in the Facebook feed reduces the likelihood of users posting positive content by 0.1% (Kramer et al., 2014). Indirect hints that social dynamics in online media may be more harmful to mental health than algorithms come from a natural experiment—the rollout of Facebook across U.S. colleges in 2004 to 2006 (Braghieri et al., 2022). At this time, recommender algorithms still played no role on Facebook. Yet the study observed that starting to use Facebook produced a moderate effect on depression and a small effect on anxiety disorders but no significant effect on eating disorders, suicidal thoughts, or

attempts. Further results hinted that the negative effects arose from unhealthy social comparisons.

Other studies on short- or long-term well-being and mental health addressed only algorithmic effects as part of social-media usage as a whole. Two randomized controlled trials testing the effects of deactivating Facebook (Allcott et al., 2020; Asimovic et al., 2021) observed small to moderate decreases in anxiety, one in depression, and one in loneliness. Many other emotions did not change, consistent with an experience-sampling study testing the effects of using Twitter (de Mello et al., 2022). Life satisfaction did not change after deactivating Facebook for 1 week (Asimovic et al., 2021), but increased after 4 weeks (Allcott et al., 2020). Furthermore, specification curve analyses showed very small negative associations with social-media usage in adolescents (Orben et al., 2019; Orben & Przybylski, 2019a).

Overall, the debate about social media and individual well-being requires more nuance. Evidence for algorithms driving or reinforcing unhealthy dynamics is very thin, supporting, at best, a small effect on mood. Instead, unfavorable social-status comparisons online may harm mental health. The direction of effects between social-media usage and mental health is unclear (Ferguson, 2021; Luhmann et al., 2022; Orben et al., 2019); the direction and size of effects depend on who uses social media and in what way (Büchi, 2021). For instance, teenage girls or already socially disadvantaged individuals may be particularly vulnerable (Allcott et al., 2020; Heffer et al., 2019; Midgley, 2019; Orben et al., 2019, 2022). Passive, extreme, or low use is related to poorer well-being, whereas active, social, and moderate use correlates with better well-being (Dienlin & Johannes, 2020). Furthermore, self-reports of usage and addiction do not reliably measure actual usage and tend to systematically overestimate them, more so in some users (such as girls) than others (Boyle et al., 2022; Mahalingham et al., 2022; Scharrow, 2016; Shaw et al., 2020).

Yet algorithmic effects could also emerge as slow trends or at higher levels of the complex system involving digital media and the offline world; studies on individuals in limited time periods cannot capture such effects. For example, the constant opportunity to express oneself could slowly affect independent emotion-regulation abilities, and algorithmic reinforcement of emotional content could change norms of emotional self-disclosure in relationships over time. In any case, potential risks to the well-being and mental health of vulnerable groups need to be taken seriously, and large corporations should be held responsible for preventing harm.

None of the cited studies says anything about the societal context, including achievement pressure and individualization increasing with neoliberalism (Levitz, 2023; Storr, 2018); increasing economic inequality and

insecurity (Wilkinson & Pickett, 2017); more single households in wealthy societies, driving loneliness; sleep irregularities and addiction, especially in younger adults (Cocco, 2022); or general uncertainties about the future (e.g., climate change; Ingle & Mikulewicz, 2020). All of these societal developments could crucially affect mental health, with algorithms reinforcing existing dynamics but not being the primary cause.

When problems have strong social or societal root causes, solutions will require difficult political, institutional, and economic changes. To address the actual causes, we need research that disentangles which, if any, of the issues currently blamed on algorithms are driven by social dynamics or societal context. The influence of societal context is particularly difficult to pin down. It would require large-scale and longitudinal studies tracing and separating multiple interacting factors and their online and offline effects over time, including algorithmic and societal changes across platforms, nations, and cultures. Data for such studies are currently not available, but the European Digital Services Act may be a step forward (Turillazzi et al., 2023).

Social Drivers and Algorithmic Mechanisms Influencing the Collective Dynamics of Political Polarization and Misinformation

Could algorithms foster echo chambers of like-minded people and polarization (Garimella et al., 2017)? Do engagement metrics promote hate speech, radicalized content, and fake news? We highlight a few studies that help dissociate algorithmic mechanisms from social drivers. For more details, see Van Bavel et al. (2021), Ferguson (2021), and Lorenz-Spreen et al. (2022).

Online echo chambers might have a more minor role than has been commonly assumed (Bakshy et al., 2015; Bruns, 2021; Guess et al., 2018; Sumpter, 2018; Törnberg, 2022) and are smaller than offline echo chambers (Gentzkow & Shapiro, 2011). Weaker online echo chambers mean that people are exposed to more people they disagree with. Similarly, digital media may increase perceived rather than actual polarization (Bail, 2021). Supporting this, a field experiment on U.S. Twitter observed increased political polarization after exposure to posts from opinion leaders of the opposing party (Bail et al., 2018) and experience sampling reveals consistent results (de Mello et al., 2022).

Increases in actual polarization are less bad than commonly assumed; there is still overlap for substantial issues in the views of political parties (Bail, 2021). Because misinformation is largely a symptom of polarization (Altay, 2022; Osmundsen et al., 2021; Petersen

et al., 2022), exposure to online misinformation might also have been overestimated. Misinformation accounts for a small proportion of digital-news consumption (Altay, Nielsen, & Fletcher, 2022) and is mostly shared by a tiny minority of users (Grinberg et al., 2019; Osmundsen et al., 2021). Additionally, misinformation has been shown not to easily change beliefs or political voting behavior (Bail et al., 2020; Guess et al., 2020).

Regarding specific algorithm effects, a study on Facebook data in 2014 (Bakshy et al., 2015) found that users' social networks determined posts in their feeds much more strongly than the algorithm. Similarly, users actively engage with more partisan news than suggested by the Google search algorithm (Robertson et al., 2023). The YouTube algorithm also does not seem to radicalize many users: Only 1 out of 100,000 who started viewing moderate content later moved to far-right content (Ribeiro et al., 2021). Most movement to far-right videos comes from outside the platform, and far-right videos are not more likely toward the end of sessions, where algorithmic recommendations matter most (Hosseinmardi et al., 2021). Instead, the demand for far-right content, with supply being easy, and the lack of more moderate conservative content may explain the increases in views of such content until mid-2017 (Munger & Phillips, 2020).

Overall, evidence neither shows that algorithms cause echo chambers, nor that echo chambers cause polarization—for example, by weakening echo chambers and exposing people to more views they disagree with. Current evidence is consistent with the view that digital media as a whole, including algorithms, fuels perceived polarization by making extremist voices more visible and hiding moderate majorities (Bail, 2021). Two randomized controlled trials support this: In the politically polarized United States, affective polarization decreased after Facebook abstinence (Allcott et al., 2020). However, not having online contact with (probably moderate) ethnic out-group members in Bosnia-Herzegovina increased affective polarization (Asimovic et al., 2021). Similar to the idea of perceived polarization increasing actual polarization, the myth of fake news being common makes people more skeptical of news in general (Altay, Berriche, & Acerbi, 2022; Fletcher & Nielsen, 2019; Guess et al., 2021). Again, most studies on digital-media effects say little about larger societal drivers of polarization. One likely driver of increasing affective polarization, and thus misinformation, is the rise of authoritarian populism in many Western countries, which itself may arise from economic insecurities or backlash to progressive cultural change (Inglehart & Norris, 2016; but see Schäfer, 2022).

Yet such societal developments can interact with algorithmic effects by affecting discourse and decisions about algorithms. Letting platforms decide how to rank content may have seemed obvious for a long time, but discussions about this are increasing. Additionally, currently polarized or populist debates may make it difficult to find common ground on algorithmic optimization metrics, making it harder to address potentially negative effects. Similar feedback loops in the positive direction could begin with algorithms that emphasize the overlap in views of political groups, which could reduce polarization. Furthermore, algorithms that emphasize nuanced content could help decrease paralyzing climate anxieties or highlight constructive perspectives that motivate action. Finally, algorithms could create more collective emotional experiences by facilitating the spreading of emotional content. This could motivate protest movements or prosocial behavior but also foster intergroup conflict and intolerance.

Research Avenues Toward Solutions and Flourishing

Digital-media companies benefit from the narrative of omnipotent algorithms, as their business model relies on their customers (i.e., advertisers) believing it (Munger, 2020a; Sumpter, 2018). For instance, Cambridge Analytica wanted its customers to believe they could shift political opinion in the crucial target group of undecided voters (Sumpter, 2018). Munger (2020a) argued that activists and society should stop buying this story. Silicon Valley corporations should carry responsibility for evaluating the potential societal consequences of their platforms. Still, blaming technology as the supposed mechanism behind a problem without looking at the drivers that power the problem is unlikely to lead to resolution. This approach directs attention away from actual root causes and potentially misleads societal discourse and policies, creating ground for further complaints.

Famous platform critics such as Francis Haugen or Elon Musk (Oremus et al., 2021; Riemer & Peter, 2022) have suggested getting rid of algorithms entirely and returning to reverse chronological ordering of posts. However, chronological order is just another kind of algorithm with its own drawbacks (Riemer & Peter, 2022): It favors more frequent posters, does not reduce information overload, and likely implies that users will miss more carefully prepared but rarer content. Getting rid of algorithms also means not using them as tools where they are indeed useful. Using algorithms well, in turn, requires developing shared visions and values—things users want algorithms to align with—which is a major important avenue for future research.

Box 1. Some Important Psychological Research Questions on Algorithms on Digital Media

Overarching questions

- With which values and purposes do we want the outcomes of our algorithms to align?
- How can psychological knowledge about social behavior and cognition help to design algorithms and platforms to best foster human well-being and flourishing at an individual and collective level?
- Do current algorithms on digital media have beneficial effects compared with media without algorithms?
- Which new risks and opportunities arise from the current shift from social to algorithmic media?
- Which platform design and algorithm features best align with different purposes? Which purposes require different digital environments, and which can be combined?
- How does giving users choices about algorithm metrics or other design features affect individual and collective well-being? Which choices should be made available, and which should be implemented broadly? How should these choices be assessed?

Individual and interindividual well-being, happiness, and flourishing

- What are specific algorithmic effects on emotions and mental health, independent from social-media usage as a whole?
- How could algorithms and other platform design features . . .
 - foster short- or long-term well-being and flourishing of individuals (including pleasure, happiness, life satisfaction, and connection)?
 - emphasize connection over social comparison between individuals?
 - foster new and deepen existing important relationships?
 - reward interindividual empathy, support, and prosocial behavior?
 - foster successful emotion regulation or collective emotional experiences?
- How would allowing users to choose what they want to see affect their well-being?

Social cohesion, nuanced political discussion, and high-quality information

- Do algorithms contribute to increasing affective polarization by fueling perceived polarization—for example, by suggesting more extreme political content?
 - How could algorithms and platforms . . .
 - make silent moderate majorities more visible?
 - reduce the visibility of extremist, toxic, outraging, or regrettable content?
 - promote nuanced and high-quality content?
 - reward expressions of empathy and understanding?
 - reward cooperation rather than status competition between groups?
 - promote constructive online intergroup contact?
 - How would allowing individual user or community choices on algorithmic ranking affect polarization and the spread of misinformation at the macroscale on digital platforms?
-

Future researchers need to develop and test theories about the role of algorithms (see Box 1), including potentially positive contributions and the mechanisms and outcomes of feedback loops with social behavior. Algorithms could even help to solve problems to which they currently contribute, and they can be intentionally designed to foster short- and long-term well-being and flourishing (Steinert & Dennis, 2022). This requires developing a vision for digital-media design and algorithm design beyond those proposed by existing for-profit companies (Bail, 2021; Büchi, 2021).

Although problem audits of algorithms are rare, studies on beneficial effects are even rarer. Some A/B tests on beneficial outcomes exist for interface design (e.g.,

Zhang et al., 2022), content manipulations and connection recommendations (e.g., Rajkumar et al., 2022), or for achieving collective outcomes with the help of random bots (Shirado & Christakis, 2017). However, more experiments comparing different optimization algorithms and comparing platforms with and without algorithms are needed.

Testing the effects of current and possible future algorithm and platform design requires platforms that allow experimental manipulation while obtaining users' consent. Computational social scientists have begun developing such bespoke social-media platforms to test the effect of concealing political affiliation or gender identity (Combs, Tierney, Alqabandi, et al., 2022;

Combs, Tierney, Guay, et al., 2022), social-engagement metrics (Avram et al., 2020), or anti-addictive design features (Zhang et al., 2022). Collaborations between academia and existing platforms are another promising approach (Stray & Hadfield, 2023).

Ideas for algorithm and platform design to foster flourishing

Algorithms can improve digital-media platforms in two ways: by using different optimization metrics to rank content, or by prompting interventions upon detecting problematic content. Current digital-media platforms show that engagement metrics that optimize for entertainment are unlikely to foster rational debates. Optimal design choices will thus likely depend on the purpose of a platform and potentially on user preferences. We may want to create different platforms to foster nuanced political discussion, amplify entertainment and short-term pleasure, promote regular contact between friends and relatives, deepen personal relationships, or build communities (e.g., for mental-health support). Future research on which design choices work best to achieve each purpose, and which ones require separate platforms or subspaces on existing platforms, would be very valuable.

Platforms for fostering nuanced political discussions that strengthen social cohesion, moderate voices, and diversity will have to focus on reducing perceived polarization. This requires reducing the visibility of strongly partisan and triggering content (Rose-Stockwell, 2018), perhaps with algorithm metrics that prioritize content popular on both sides of the political spectrum. This would highlight moderate voices and reveal the opinion overlap for the important issues where it actually exists (Bail, 2021), and could promote more trustworthy news sources (Bhadani et al., 2022). Similar algorithms could highlight which principles or practical approaches resonate with people on both sides of other belief spectrums, such as those relating to climate change or alternative medicine.

Algorithm rankings could further foster intergroup contact and understanding by presenting posts that are not too distant from a user's own position (Levendusky, 2018; Sherif, 1963). In this way, algorithms could support small steps toward understanding alternative views. Using algorithmic estimates of users' positions on a dimension, platforms could further label extreme voices as such, give users feedback about their own position, or show how moderate and extreme users on both sides have responded to an account or post (Bail, 2021). Other suggestions include toning down status incentives by hiding or reducing the visibility of engagement metrics for certain types of posts (Avram et al.,

2020) or adding cues that spotlight passive user behaviors (e.g., how many scrolled over a post; Lorenz-Spreen et al., 2020). Twitter introduced view counts for tweets early in 2023, creating research opportunities to explore how this affects social-reward experiences or the spreading of polarizing and untrustworthy content. Finally, anonymity is a promising nonalgorithmic design feature for reducing conflicts rooted in social identity, with the potential to make discussions on controversial issues kinder (Combs, Tierney, Guay, et al., 2022).

Optimizing algorithms for metrics such as civility (Lewandowsky & Kozyreva, 2022; Oremus et al., 2021) would require defining what counts as civil and how civility fosters democratic discourse and diversity. When a minority is unjustly neglected, or an elite unfairly privileged, for example, angry responses are appropriate and necessary. Rather than deciding which values should guide the choice of algorithm metrics, platforms could also let users define values themselves (Lewandowsky & Kozyreva, 2022; Lorenz-Spreen et al., 2020). Facebook has tested such an approach in its "breaking the glass" experiments, deploying an algorithm that emphasized posts that users considered to be "good for the world" (Roose et al., 2021). Although this reduced low-quality content, it also lowered how often users opened Facebook and was therefore implemented only in a weakened version.

A second way to use algorithms is to detect certain posts or activities and then trigger interventions. The simplest of all interventions is adding friction, that is, increasing the time or effort it takes to share content (Brady et al., 2020; Lorenz-Spreen et al., 2020; Menczer, 2021). Adding friction seems particularly useful to prevent impulsive sharing of sensational news and outraged or toxic comments. In some cases, simply adding a time gap before allowing users to post or share might suffice. In others, additional prompts could encourage reflection before sharing (Rose-Stockwell, 2018). Empathic and humanizing prompts have been shown to reduce affective polarization (Saveski et al., 2022) and racist harassment (Hangartner et al., 2021; Munger, 2017). Undo prompts after posting hateful comments, default options to turn comments from public to private, or ideological prompts explaining that posts with moral-emotional language are unlikely to reach the other side, could all reduce hateful content (Rose-Stockwell, 2018). Interventions that effectively reduce affective polarization provide further inspiration (Hartman et al., 2022; Voelkel et al., 2022).

To foster mental health and healthy usage of digital media, algorithms can detect linguistic markers of symptoms or certain activity patterns. Trying to detect users at risk of mental-health issues, with the goal of then providing contact points for support, is a popular

research field, which, however, urgently requires methodological-validation efforts (Chancellor & De Choudhury, 2020). To reduce unwanted addictive use, algorithms can encourage users to disengage by providing reminder bots after excessive periods of scrolling or providing usage statistics (Zhang et al., 2022). Interventions such as reading-progress indicators, feed filters and content blockers for specific types of content, and separate topic-focused feeds instead of one main feed, seem even more effective (Zhang et al., 2022), and could be improved via algorithmic suggestions. Finally, we do not know of any research on how changing algorithm metrics could support individual well-being. Algorithms that reduce the visibility of toxic, regrettable, and outraged content may help reduce content that negatively affects well-being (Rose-Stockwell, 2018). Research on algorithms that prioritize content from important personal contacts, expressions of empathy and connection, or prosocial behaviors could contribute to positive well-being outcomes.

Choosing values, validating metrics, and evaluating their effect on outcomes

As the above discussion illustrates, many different values are potentially justifiable candidates for algorithmic optimization. Choosing such values, validating the metrics to optimize for them, and testing their effect on various outcomes will require research in cultural, moral, social, political, affective, and clinical psychology, as well as computational, sociological, political, and economic approaches. Such research needs to determine for which values societal consensus is possible, and where digital media have to accommodate different needs, visions, and goals within the same or between different platforms. It should further explore which values and goals individuals prioritize and how social and cultural norms affect these processes in communities and societies.

Research could also compare different ways of assessing these value preferences: Avoiding perpetuating the influence of social and cognitive biases will probably require asking for user decisions in advance and at an abstract level, rather than measuring immediate preferences when users thoughtlessly scroll through their feeds. Preliminary research shows that most U.S. users across political and demographic groups opt for seeing more accurate, nuanced, friendly, positive, and educational content (Rathje et al., 2022), although such content currently does not typically go viral by itself. Researchers need to test whether users would actually make the choices they report preferring on the platforms they regularly use, and they then need to determine whether this would reduce misinformation and

polarization at the macroscale of digital platforms. They further need to explore where individual user or community choices on algorithmic rankings or interventions are possible and beneficial (Lewandowsky & Kozyreva, 2022; Lorenz-Spreen et al., 2020), and where they need to be restricted. For example, letting users opt for only partisan content is dangerous, as contact with moderate voices from other groups may be necessary to reduce polarization.

Once certain values are agreed upon, methodological research can be employed to validate which metrics could actually represent those values, relying on the digital trace data available to algorithms. Finally, empirical research should be used to investigate how different metrics would affect the various outcomes, including affective well-being, mental health, societal cohesion, and nuanced political discussions. Given that social media are complex systems with emerging feedback loops between social drivers and algorithms, this research needs to incorporate methods from complexity science and computational social science, such as network analysis or agent-based modeling (Borsboom et al., 2021; Jackson et al., 2017; Smith & Conrey, 2007; Vlasceanu et al., 2018) to address these many open questions.

Shift from social to algorithmic media

Twitter, Facebook, and Instagram could be referred to as *traditional social media*, as their information-distribution mechanism relies primarily on social networks (Mignano, 2022; Munger, 2020b). In contrast, other platforms like YouTube and TikTok mostly rely on recommendation algorithms instead of social links. On traditional social media, social drivers can have a much larger influence on interactions and spreading dynamics. In contrast, on *algorithmic media*, the platform itself has much more power to determine presented content through the recommender system and content feed (Mignano, 2022; Narayanan, 2023b). Algorithms are more economically competitive as information-distribution mechanisms because social graphs are now easily available (Mignano, 2022). Likely for this reason, Facebook and Instagram have started following TikTok's example by adding short recommendation-based video feeds. This trend may entirely change our current conclusions about the algorithmic effects that have been limited so far. Algorithmic media might worsen problems like addiction or propaganda. Munger (2020b) has argued that the immersive nature of TikTok's mobile-first design, its higher capacity to evoke emotions via both visual and audio information, the ease of posting content, and the unpredictable virality of its algorithm might make it more addictive and its users more vulnerable to political persuasion.

However, the shift from social to algorithmic media may also present an opportunity for the endeavor of designing digital media that foster human flourishing. Because algorithmic content distribution gives greater control to the platform compared with popular users, algorithms could select content on the basis of metrics that foster well-being of individuals and societies. They could highlight overlap between opposed groups (Bail, 2021), prioritize news a user actually wants to see (Rathje et al., 2022), or simply limit how far fake news spreads (Bak-Coleman et al., 2022). Both new risks and opportunities arising from algorithmic media are important avenues for future research in psychology and computational social science.

If algorithms make societally relevant decisions, it becomes pivotal who takes these decisions, and in what way. Making sure these decisions benefit society will require transparency about algorithmic design (Kozyreva et al., 2021; Wagner et al., 2021). The recent release of the code of the Twitter algorithm illustrates that in order to actually evaluate effects we need not only information about how the algorithms weigh types of content and interactions but also information about the machine-learning models that make suggestions for individual users (Narayanan, 2023a). Although we see potential in beneficially using algorithms on digital media, we must acknowledge the barriers that exist for this kind of research. Because the vast majority of online media are proprietary for-profit platforms, the designs and targets we presented are likely at odds with profit-making to a certain extent. Testing, implementing, and adopting solutions will therefore likely require regulation (Gal, 2022). Given the unique role of digital media in creating a public sphere in a globalized world, researchers and activists have even discussed whether digital media should become a public good (Fournier-Tombs, 2022). However, we are also still very early in the history of digital-media platforms, with large shifts of users to new platforms every couple of years (Bail, 2021). Over time, market dynamics could still play out in ways that better satisfy user preferences beyond short-term rewards.

Conclusion

We have outlined different ways in which algorithms on digital media could promote positive emotions, mental health, social cohesion, and nuanced discourse. In the context of a globalized world, polarized democracies, and increasingly individualized societies, efforts to design algorithms that foster intergroup contact via digital media may make valuable contributions to reduce social, ethnic, political, and cultural barriers.

Transparency

Action Editor: Melanie Mitchell

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 work was supported by two grants from the Vienna Science and Technology Fund: Grant No. VRG16-005 to D. Garcia and ICT20-028 to H. Metzler.

ORCID iDs

Hannah Metzler  <https://orcid.org/0000-0001-9254-3675>

David Garcia  <https://orcid.org/0000-0002-2820-9151>

References

- Abele, A. E., & Wojciszke, B. (2013). The Big Two in social judgment and behavior. *Social Psychology, 44*, 61–62. <https://doi.org/10.1027/1864-9335/a000137>
- Allcott, H., Braghieri, L., Eichmeyer, S., & Gentzkow, M. (2020). The welfare effects of social media. *American Economic Review, 110*(3), 629–676. <https://doi.org/10.1257/aer.20190658>
- Altay, S. (2022). *How effective are interventions against misinformation?* PsyArXiv. <https://doi.org/10.31234/osf.io/sm3vk>
- Altay, S., Berriche, M., & Acerbi, A. (2022). *Misinformation on misinformation: Conceptual and methodological challenges.* PsyArXiv. <https://doi.org/10.31234/osf.io/edqc8>
- 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*, 1–29. <https://doi.org/10.51685/jqd.2022.020>
- Asimovic, N., Nagler, J., Bonneau, R., & Tucker, J. A. (2021). Testing the effects of Facebook usage in an ethnically polarized setting. *Proceedings of the National Academy of Sciences, USA, 118*(25), Article e2022819118. <https://doi.org/10.1073/pnas.2022819118>
- Avram, M., Micallef, N., Patil, S., & Menczer, F. (2020). Exposure to social engagement metrics increases vulnerability to misinformation. *Harvard Kennedy School Misinformation Review, 1*(5), 1–11. <https://doi.org/10.37016/mr-2020-033>
- Bail, C. A. (2021). *Breaking the social media prism: How to make our platforms less polarizing.* Princeton University Press.
- Bail, C. A., Argyle, L. P., Brown, T. W., Bumpus, J. P., Chen, H., Hunzaker, M. B. F., Lee, J., Mann, M., Merhout, F., & Volfovsky, A. (2018). Exposure to opposing views on social media can increase political polarization. *Proceedings of the National Academy of Sciences, USA, 115*(37), 9216–9221. <https://doi.org/10.1073/pnas.1804840115>
- Bail, C. A., Guay, B., Maloney, E., Combs, A., Hillygus, D. S., Merhout, F., Freelon, D., & Volfovsky, A. (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, USA*, 117(1), 243–250. <https://doi.org/10.1073/pnas.1906420116>
- 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., Kao, A. B., Moran, R. E., Romanczuk, P., Rubenstein, D. I., Tombak, K. J., Van Bavel, J. J., & Weber, E. U. (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>
- Bak-Coleman, J. B., Kennedy, I., Wack, M., Beers, A., Schafer, J. S., Spiro, E. S., Starbird, K., & West, J. D. (2022). Combining interventions to reduce the spread of viral misinformation. *Nature Human Behaviour*. Advance online publication. <https://doi.org/10.1038/s41562-022-01388-6>
- Bakshy, E., Messing, S., & Adamic, L. A. (2015). Exposure to ideologically diverse news and opinion on Facebook. *Science*, 348(6239), 1130–1132. <https://doi.org/10.1126/science.aaa1160>
- Bhadani, S., Yamaya, S., Flammini, A., Menczer, F., Ciampaglia, G. L., & Nyhan, B. (2022). Political audience diversity and news reliability in algorithmic ranking. *Nature Human Behaviour*. Advance online publication. <https://doi.org/10.1038/s41562-021-01276-5>
- Bliss, N., Bradley, E., Garland, J., Menczer, F., Ruston, S. W., Starbird, K., & Wiggins, C. (2020). *An agenda for disinformation research* (arXiv:2012.08572). arXiv. <https://doi.org/10.48550/arXiv.2012.08572>
- Booth, R. (2014, June 29). Facebook reveals news feed experiment to control emotions. *The Guardian*. <https://www.theguardian.com/technology/2014/jun/29/facebook-users-emotions-news-feeds>
- Borsboom, D., Deserno, M. K., Rhemtulla, M., Epskamp, S., Fried, E. I., McNally, R. J., Robinaugh, D. J., Perugini, M., Dalege, J., Costantini, G., Isvoranu, A.-M., Wysocki, A. C., van Borkulo, C. D., van Bork, R., & Waldorp, L. J. (2021). Network analysis of multivariate data in psychological science. *Nature Reviews Methods Primers*, 1(1), Article 1. <https://doi.org/10.1038/s43586-021-00055-w>
- Boyle, S. C., Baez, S., Trager, B. M., & LaBrie, J. W. (2022). Systematic bias in self-reported social media use in the age of platform swinging: Implications for studying social media use in relation to adolescent health behavior. *International Journal of Environmental Research and Public Health*, 19(16), Article 16. <https://doi.org/10.3390/ijerph19169847>
- Brady, W. J., Crockett, M. J., & Van Bavel, J. J. (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*, 15(4), 978–1010. <https://doi.org/10.1177/1745691620917336>
- Braghieri, L., Levy, R., & Makarin, A. (2022). *Social media and mental health* (SSRN Scholarly Paper No. 3919760). <https://doi.org/10.2139/ssrn.3919760>
- Bruns, A. (2021). Echo chambers? Filter bubbles? The misleading metaphors that obscure the real problem. In M. Pérez-Escobar & J. M. Noguera-Vivo (Eds.), *Hate speech and polarization in participatory society* (pp. 33–48). Routledge.
- Büchi, M. (2021). Digital well-being theory and research. *New Media & Society*. Advance online publication. <https://doi.org/10.1177/14614448211056851>
- Burke, M., Cheng, J., & de Gant, B. (2020). Social comparison and Facebook: Feedback, positivity, and opportunities for comparison. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (pp. 1–13). <https://doi.org/10.1145/3313831.3376482>
- Carr, C. T., Hayes, R. A., & Sumner, E. M. (2018). Predicting a threshold of perceived Facebook post success via likes and reactions: A test of explanatory mechanisms. *Communication Research Reports*, 35(2), 141–151. <https://doi.org/10.1080/08824096.2017.1409618>
- Cavanagh, S. R. (2017, August 6). No, smartphones are not destroying a generation. *Psychology Today*. <https://www.psychologytoday.com/us/blog/once-more-feeling/201708/no-smartphones-are-not-destroying-generation>
- Chancellor, S., & De Choudhury, M. (2020). Methods in predictive techniques for mental health status on social media: A critical review. *npj Digital Medicine*, 3(1), Article 1. <https://doi.org/10.1038/s41746-020-0233-7>
- Ciampaglia, G. L., Nematzadeh, A., Menczer, F., & Flammini, A. (2018). How algorithmic popularity bias hinders or promotes quality. *Scientific Reports*, 8(1), Article 1. <https://doi.org/10.1038/s41598-018-34203-2>
- Cingel, D. P., Carter, M. C., & Krause, H.-V. (2022). Social media and self-esteem. *Current Opinion in Psychology*, 45, Article 101304. <https://doi.org/10.1016/j.copsyc.2022.101304>
- Cocco, F. (2022, November 24). Are we ready for the approaching loneliness epidemic? *Financial Times*. <https://www.ft.com/content/c3aef690-b5a5-4f0d-9da5-2bf4c560c4f4>
- Combs, A., Tierney, G., Alqabandi, F., Cornell, D., Varela, G., Araujo, A. C., Argyle, L., Bail, C. A., & Volfovsky, A. (2022). *Perceived gender and political persuasion: A social media field experiment during the 2020 Democratic National Primary*. SocArXiv. <https://doi.org/10.31235/osf.io/537qn>
- Combs, A., Tierney, G., Guay, B., Merhout, F., Bail, C. A., Hillygus, D. S., & Volfovsky, A. (2022). *Anonymous cross-party conversations can decrease political polarization: A field experiment on a mobile chat platform*. SocArXiv. <https://doi.org/10.31235/osf.io/cwgu5>
- Cummins, D. (2005). Dominance, status, and social hierarchies. In D. M. Buss (Ed.), *The handbook of evolutionary psychology* (pp. 676–697). Wiley.
- de Mello, V. O., Cheung, F., & Inzlicht, M. (2022). *Twitter use in the everyday life: Exploring how Twitter use predicts well-being, polarization, and sense of belonging*. PsyArXiv. <https://psyarxiv.com/4x5em/>
- Dienlin, T., & Johannes, N. (2020). The impact of digital technology use on adolescent well-being. *Dialogues in Clinical Neuroscience*, 22(2), 135–142. <https://doi.org/10.31887/DCNS.2020.22.2/dienlin>
- Eslinger, P. J., Anders, S., Ballarini, T., Boutros, S., Krach, S., Mayer, A. V., Moll, J., Newton, T. L., Schroeter, M. L., de Oliveira-Souza, R., Raber, J., Sullivan, G. B., Swain,

- J. E., Lowe, L., & Zahn, R. (2021). The neuroscience of social feelings: Mechanisms of adaptive social functioning. *Neuroscience & Biobehavioral Reviews*, *128*, 592–620. <https://doi.org/10.1016/j.neubiorev.2021.05.028>
- Ferguson, C. J. (2021). Does the Internet make the world worse? Depression, aggression and polarization in the social media age. *Bulletin of Science, Technology & Society*, *41*(4), 116–135. <https://doi.org/10.1177/02704676211064567>
- Fiske, S. T., Cuddy, A. J. C., & Glick, P. (2007). Universal dimensions of social cognition: Warmth and competence. *Trends in Cognitive Sciences*, *11*(2), 77–83. <https://doi.org/10.1016/j.tics.2006.11.005>
- Fletcher, R., & Nielsen, R. K. (2019). Generalised scepticism: How people navigate news on social media. *Information, Communication & Society*, *22*(12), 1751–1769. <https://doi.org/10.1080/1369118X.2018.1450887>
- Fournier-Tombs, E. (2022, May 4). Elon Musk's proposed takeover of Twitter raises questions about its role in the digital social infrastructure. *The Conversation*. <http://theconversation.com/elon-musks-proposed-takeover-of-twitter-raises-questions-about-its-role-in-the-digital-social-infrastructure-182271>
- Gal, U. (2022, June 10). 'Transparency reports' from tech giants are vague on how they're combating misinformation. It's time for legislation. *The Conversation*. <http://theconversation.com/transparency-reports-from-tech-giants-are-vague-on-how-theyre-combating-misinformation-its-time-for-legislation-184476>
- Garimella, K., De Francisci Morales, G., Gionis, A., & Mathioudakis, M. (2017). Reducing controversy by connecting opposing views. In *Proceedings of the Tenth ACM International Conference on Web Search and Data Mining* (pp. 81–90). <https://doi.org/10.1145/3018661.3018703>
- Gebauer, J. E., Paulhus, D. L., & Neberich, W. (2013). Big two personality and religiosity across cultures: Communals as religious conformists and agentics as religious contrarians. *Social Psychological and Personality Science*, *4*(1), 21–30. <https://doi.org/10.1177/1948550612442553>
- Gebauer, J. E., Sedikides, C., Lüdtke, O., & Neberich, W. (2014). Agency-communion and interest in prosocial behavior: Social motives for assimilation and contrast explain socio-cultural inconsistencies. *Journal of Personality*, *82*, 452–466. <https://doi.org/10.1111/jopy.12076>
- Gentzkow, M., & Shapiro, J. M. (2011). Ideological segregation online and offline. *The Quarterly Journal of Economics*, *126*(4), 1799–1839. <https://doi.org/10.1093/qje/qjr044>
- Grinberg, N., Joseph, K., Friedland, L., Swire-Thompson, B., & Lazer, D. (2019). Fake news on Twitter during the 2016 U.S. presidential election. *Science*, *363*(6425), 374–378. <https://doi.org/10.1126/science.aau2706>
- Guess, A. M., Barberá, P., Munzert, S., & Yang, J. (2021). The consequences of online partisan media. *Proceedings of the National Academy of Sciences, USA*, *118*(14), Article e2013464118. <https://doi.org/10.1073/pnas.2013464118>
- Guess, A. M., Lockett, D., Lyons, B., Montgomery, J. M., Nyhan, B., & Reifler, J. (2020). "Fake news" may have limited effects beyond increasing beliefs in false claims. *Harvard Kennedy School Misinformation Review*, *1*(1). <https://doi.org/10.37016/mr-2020-004>
- Guess, A. M., Nyhan, B., Lyons, B., & Reifler, J. (2018). *Avoiding the echo chamber about echo chambers: Why selective exposure to like-minded political news is less prevalent than you think* (No. 2, pp. 1–25). Knight Foundation.
- Gurtman, M. B. (2009). Exploring personality with the interpersonal circumplex. *Social and Personality Psychology Compass*, *3*(4), 601–619. <https://doi.org/10.1111/j.1751-9004.2009.00172.x>
- Hangartner, D., Gennaro, G., Alasiri, S., Bahrnich, N., Bornhoft, A., Boucher, J., Demirci, B. B., Derksen, L., Hall, A., Jochum, M., Munoz, M. M., Richter, M., Vogel, F., Wittwer, S., Wüthrich, F., Gilardi, F., & Donnay, K. (2021). Empathy-based counterspeech can reduce racist hate speech in a social media field experiment. *Proceedings of the National Academy of Sciences, USA*, *118*(50), Article e2116310118. <https://doi.org/10.1073/pnas.2116310118>
- Hartman, R., Blakey, W., Womick, J., Bail, C. A., Finkel, E., Schroeder, J., Sheeran, P., Bavel, J. J. V., Willer, R., & Gray, K. (2022). *Interventions to reduce partisan animosity*. PsyArXiv. <https://doi.org/10.31234/osf.io/ha2tf>
- Heffer, T., Good, M., Daly, O., MacDonell, E., & Willoughby, T. (2019). The longitudinal association between social-media use and depressive symptoms among adolescents and young adults: An empirical reply to Twenge et al. (2018). *Clinical Psychological Science*, *7*(3), 462–470. <https://doi.org/10.1177/2167702618812727>
- Hosseinmardi, H., Ghasemian, A., Clauset, A., Mobius, M., Rothschild, D. M., & Watts, D. J. (2021). Examining the consumption of radical content on YouTube. *Proceedings of the National Academy of Sciences, USA*, *118*(32), Article e2101967118. <https://doi.org/10.1073/pnas.2101967118>
- Ingle, H. E., & Mikulewicz, M. (2020). Mental health and climate change: Tackling invisible injustice. *The Lancet Planetary Health*, *4*(4), e128–e130. [https://doi.org/10.1016/S2542-5196\(20\)30081-4](https://doi.org/10.1016/S2542-5196(20)30081-4)
- Inglehart, R. F., & Norris, P. (2016). *Trump, Brexit, and the rise of populism: Economic have-nots and cultural backlash* (SSRN Scholarly Paper No. 2818659). <https://doi.org/10.2139/ssrn.2818659>
- Jackson, J. C., Rand, D., Lewis, K., Norton, M. I., & Gray, K. (2017). Agent-based modeling: A guide for social psychologists. *Social Psychological and Personality Science*, *8*(4), 387–395. <https://doi.org/10.1177/1948550617691100>
- Kozyreva, A., Lorenz-Spreen, P., Hertwig, R., Lewandowsky, S., & Herzog, S. M. (2021). Public attitudes towards algorithmic personalization and use of personal data online: Evidence from Germany, Great Britain, and the United States. *Humanities and Social Sciences Communications*, *8*(1), Article 1. <https://doi.org/10.1057/s41599-021-00787-w>
- Kramer, A. D. I., Guillory, J. E., & Hancock, J. T. (2014). Experimental evidence of massive-scale emotional contagion through social networks. *Proceedings of the National Academy of Sciences, USA*, *111*(24), 8788–8790. <https://doi.org/10.1073/pnas.1320040111>
- Levendusky, M. S. (2018). When efforts to depolarize the electorate fail. *Public Opinion Quarterly*, *82*(3), 583–592. <https://doi.org/10.1093/poq/nfy036>
- Levitz, E. (2023, March 27). 4 explanations for the teen mental-health crisis. *Intelligencer*. <https://nymag.com/>

- intelligencer/2023/03/4-explanations-for-the-teen-mental-health-crisis.html
- Lewandowsky, S., & Kozyreva, A. (2022, April 7). Algorithms, lies, and social media. *Nieman Lab*. <https://www.niemanlab.org/2022/04/algorithms-lies-and-social-media/>
- Lewis-Kraus, G. (2022, June 3). How harmful is social media? *The New Yorker*. <https://www.newyorker.com/culture/annals-of-inquiry/we-know-less-about-social-media-than-we-think>
- Lindström, B., Bellander, M., Schultner, D. T., Chang, A., Tobler, P. N., & Amodio, D. M. (2021). A computational reward learning account of social media engagement. *Nature Communications*, *12*(1), Article 1311. <https://doi.org/10.1038/s41467-020-19607-x>
- Lorenz-Spreen, P., Lewandowsky, S., Sunstein, C. R., & Hertwig, R. (2020). How behavioural sciences can promote truth, autonomy and democratic discourse online. *Nature Human Behaviour*. Advance online publication. <https://doi.org/10.1038/s41562-020-0889-7>
- Lorenz-Spreen, P., Oswald, L., Lewandowsky, S., & Hertwig, R. (2022). A systematic review of worldwide causal and correlational evidence on digital media and democracy. *Nature Human Behaviour*. Advance online publication. <https://doi.org/10.1038/s41562-022-01460-1>
- Luhmann, M., Buecker, S., & Rüsberg, M. (2022). Loneliness across time and space. *Nature Reviews Psychology*, *2*, 9–23. <https://doi.org/10.1038/s44159-022-00124-1>
- Mahalingham, T., McEvoy, P. M., & Clarke, P. J. F. (2022). Assessing the validity of self-report social media use: Evidence of no relationship with objective smartphone use. *Computers in Human Behavior*, *140*, Article 107567. <https://doi.org/10.1016/j.chb.2022.107567>
- Marwick, A. E. (2015). Instafame: Luxury selfies in the attention economy. *Public Culture*, *27*(1), 137–160. <https://doi.org/10.1215/08992363-2798379>
- Marwick, A. E., & boyd, d. (2011). I tweet honestly, I tweet passionately: Twitter users, context collapse, and the imagined audience. *New Media & Society*, *13*(1), 114–133. <https://doi.org/10.1177/1461444810365313>
- Menczer, F. (2021, September 13). *How “engagement” makes you vulnerable to manipulation and misinformation on social media*. Nieman Lab, 1.
- Merrill, J. B., & Oremus, W. (2021, October 26). Five points for anger, one for a ‘like’: How Facebook’s formula fostered rage and misinformation. *Washington Post*. <https://www.washingtonpost.com/technology/2021/10/26/facebook-angry-emoji-algorithm/>
- Meshi, D., Tamir, D. I., & Heekeren, H. R. (2015). The emerging neuroscience of social media. *Trends in Cognitive Sciences*, *19*(12), 771–782. <https://doi.org/10.1016/j.tics.2015.09.004>
- Midgley, C. (2019). *When every day is a high school reunion: Social media comparisons and self-esteem* (Thesis). Retrieved from <https://tspace.library.utoronto.ca/handle/1807/95911>
- Mignano, M. (2022, August 8). The end of social media and the rise of recommendation media. *Medium*. <https://mignano.medium.com/the-end-of-social-media-a88fed21f86>
- Munger, K. (2017). Tweetment effects on the tweeted: Experimentally reducing racist harassment. *Political Behavior*, *39*(3), 629–649. <https://doi.org/10.1007/s11109-016-9373-5>
- Munger, K. (2020a, October 26). Susan Wojcicki wants you to think that YouTube’s algorithm is all-powerful [Substack newsletter]. *Never Met a Science*. <https://kevinmunger.substack.com/p/susan-wojcicki-wants-you-to-think>
- Munger, K. (2020b, July 26). Theorizing TikTok [Substack newsletter]. *Never Met a Science*. <https://kevinmunger.substack.com/p/theorizing-tiktok>
- Munger, K., & Phillips, J. (2020). Right-wing YouTube: A supply and demand perspective. *The International Journal of Press/Politics*, *27*(1), 186–219. <https://doi.org/10.1177/1940161220964767>
- Munn, L. (2020). Angry by design: Toxic communication and technical architectures. *Humanities and Social Sciences Communications*, *7*(1), Article 1. <https://doi.org/10.1057/s41599-020-00550-7>
- Nadkarni, A., & Hofmann, S. G. (2012). Why do people use Facebook? *Personality and Individual Differences*, *52*(3), 243–249. <https://doi.org/10.1016/j.paid.2011.11.007>
- Nadler, A. (2016). Intergroup helping relations. *Current Opinion in Psychology*, *11*, 64–68.
- Narayanan, A. (2023a, April 10). Twitter showed us its algorithm. What does it tell us? Knight First Amendment Institute. <http://knightcolumbia.org/blog/twitter-showed-us-its-algorithm-what-does-it-tell-us>
- Narayanan, A. (2023b, March 9). *Understanding social media recommendation algorithms*. Knight First Amendment Institute. <https://knightcolumbia.org/content/understanding-social-media-recommendation-algorithms>; <https://perma.cc/F3NP-FEQX>
- Nikolov, D., Lalmas, M., Flammini, A., & Menczer, F. (2019). Quantifying biases in online information exposure. *Journal of the Association for Information Science and Technology*, *70*(3), 218–229. <https://doi.org/10.1002/asi.24121>
- Orben, A., Dienlin, T., & Przybylski, A. K. (2019). Social media’s enduring effect on adolescent life satisfaction. *Proceedings of the National Academy of Sciences, USA*, *116*(21), 10226–10228. <https://doi.org/10.1073/pnas.1902058116>
- Orben, A., & Przybylski, A. K. (2019a). The association between adolescent well-being and digital technology use. *Nature Human Behaviour*, *3*(2), 173–182. <https://doi.org/10.1038/s41562-018-0506-1>
- Orben, A., & Przybylski, A. K. (2019b). Screens, teens, and psychological well-being: Evidence from three time-use-diary studies. *Psychological Science*, *30*(5), 682–696. <https://doi.org/10.1177/0956797619830329>
- Orben, A., Przybylski, A. K., Blakemore, S.-J., & Kievit, R. A. (2022). Windows of developmental sensitivity to social media. *Nature Communications*, *13*(1), Article 1. <https://doi.org/10.1038/s41467-022-29296-3>
- Oremus, W., Alcantara, C., Merrill, J. B., & Galocha, A. (2021, October). How Facebook shapes your feed. *Washington Post*. <https://www.washingtonpost.com/technology/interactive/2021/how-facebook-algorithm-works/>
- 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(3), 999–1015. <https://doi.org/10.1017/S0003055421000290>
- Pacheco, D., Hui, P.-M., Torres-Lugo, C., Truong, B. T., Flammini, A., & Menczer, F. (2021). Uncovering coordinated networks on social media: Methods and case studies. *Proceedings of the International AAAI Conference on Web and Social Media*, 15, 455–466. <https://doi.org/10.1609/icwsm.v15i1.18075>
- Petersen, M. B., Osmundsen, M., & Arceneaux, K. (2022). The “need for chaos” and motivations to share hostile political rumors. *American Political Science Review*. Advance online publication. <https://doi.org/10.31234/osf.io/6m4ts>
- Rajkumar, K., Saint-Jacques, G., Bojinov, I., Brynjolfsson, E., & Aral, S. (2022). A causal test of the strength of weak ties. *Science*, 377(6612), 1304–1310. <https://doi.org/10.1126/science.abl4476>
- Rathje, S., Robertson, C., Brady, W. J., & Bavel, J. J. V. (2022). *People think that social media platforms do (but should not) amplify divisive content*. PsyArXiv. <https://doi.org/10.31234/osf.io/gmun4>
- Ribeiro, M. H., Ottoni, R., West, R., Almeida, V. A. F., & Meira, W. (2021). *Auditing radicalization pathways on YouTube* (arXiv:1908.08313). arXiv. <https://doi.org/10.48550/arXiv.1908.08313>
- Riener, K., & Peter, S. (2022). Wrong, Elon Musk: The big problem with free speech on platforms isn't censorship. It's the algorithms. *The Conversation*. <http://theconversation.com/wrong-elon-musk-the-big-problem-with-free-speech-on-platforms-isnt-censorship-its-the-algorithms-182433>
- Ritchie, S. (2021, September 21). Is Instagram really bad for teenagers? *Unherd*. <https://unherd.com/2021/09/face-books-bad-science/>
- 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*. Advance online publication. <https://doi.org/10.1038/s41586-023-06078-5>
- Roose, K., Isaac, M., & Frenkel, S. (2021, January 7). Facebook struggles to balance civility and growth. *The New York Times*. <https://www.nytimes.com/2020/11/24/technology/facebook-election-misinformation.html>
- Rosenthal-von der Pütten, A. M., Hastall, M. R., Köcher, S., Meske, C., Heinrich, T., Labrenz, F., & Ocklenburg, S. (2019). “Likes” as social rewards: Their role in online social comparison and decisions to like other People's selfies. *Computers in Human Behavior*, 92, 76–86. <https://doi.org/10.1016/j.chb.2018.10.017>
- Rose-Stockwell, T. (2018). How to design better social media. *Medium*. <https://medium.com/s/story/how-to-fix-what-social-media-has-broken-cb0b2737128>
- Salganik, M. J. (2019). *Bit by bit: Social research in the digital age*. Princeton University Press. <https://www.bitbybitbook.com>
- Salganik, M. J., Dodds, P. S., & Watts, D. J. (2006). Experimental study of inequality and unpredictability in an artificial cultural market. *Science*, 311(5762), 854–856. <https://doi.org/10.1126/science.1121066>
- Sapolsky, R. M. (2005). The influence of social hierarchy on primate health. *Science*, 308(5722), 648–652. <https://doi.org/10.1126/science.1106477>
- Saveski, M., Gillani, N., Yuan, A., Vijayaraghavan, P., & Roy, D. (2022). Perspective-taking to reduce affective polarization on social media. *Proceedings of the International AAAI Conference on Web and Social Media*, 16, 885–895.
- Schäfer, A. (2022). Cultural backlash? How (not) to explain the rise of authoritarian populism. *British Journal of Political Science*, 52(4), 1977–1993. <https://doi.org/10.1017/S0007123421000363>
- Schafer, M., & Schiller, D. (2018). Navigating social space. *Neuron*, 100(2), 476–489. <https://doi.org/10.1016/j.neuron.2018.10.006>
- Scharkow, M. (2016). The accuracy of self-reported Internet use—A validation study using client log data. *Communication Methods and Measures*, 10(1), 13–27. <https://doi.org/10.1080/19312458.2015.1118446>
- Sen, I., Aggarwal, A., Mian, S., Singh, S., Kumaraguru, P., & Datta, A. (2018). Worth its weight in likes: Towards detecting fake likes on Instagram. In *Proceedings of the 10th ACM Conference on Web Science* (pp. 205–209). <https://doi.org/10.1145/3201064.3201105>
- Shaw, H., Ellis, D. A., Geyer, K., Davidson, B. I., Ziegler, F. V., & Smith, A. (2020). Quantifying smartphone “use”: Choice of measurement impacts relationships between “usage” and health. *Technology, Mind, and Behavior*, 1(2). <https://doi.org/10.1037/tmb0000022>
- Sherif, C. W. (1963). Social categorization as a function of latitude of acceptance and series range. *The Journal of Abnormal and Social Psychology*, 67, 148–156. <https://doi.org/10.1037/h0043022>
- Shirado, H., & Christakis, N. A. (2017). Locally noisy autonomous agents improve global human coordination in network experiments. *Nature*, 545(7654), Article 7654. <https://doi.org/10.1038/nature22332>
- Smith, E. R., & Conrey, F. R. (2007). Agent-based modeling: A new approach for theory building in social psychology. *Personality and Social Psychology Review*, 11(1), 87–104. <https://doi.org/10.1177/1088868306294789>
- Smyth, J., & Murphy, H. (2023, March 26). The teen mental health crisis: A reckoning for Big Tech. *Financial Times*. <https://www.ft.com/content/77d06d3e-2b9f-4d46-814f-da2646fea60c>
- Steinert, S., & Dennis, M. J. (2022). Emotions and digital well-being: On social media's emotional affordances. *Philosophy & Technology*, 35(2), Article 36. <https://doi.org/10.1007/s13347-022-00530-6>
- Storr, W. (2018). *Selfie: How we became so self-obsessed and what it's doing to us*. Picador. <https://www.scribd.com/book/510974252/Selfie-How-We-Became-So-Self-Obsessed-and-What-It-s-Doing-to-Us>
- Stray, J., & Hadfield, G. (2023, March 30). A unique experiment that could make social media better. *Wired*. <https://www.wired.com/story/platforms-engagement-research-meta/>
- Sumpter, D. (2018). *Outnumbered: From Facebook and Google to fake news and filter-bubbles – The algorithms that control our lives*. Bloomsbury Publishing.

- Todorov, A., Said, C. P., Engell, A. D., & Oosterhof, N. N. (2008). Understanding evaluation of faces on social dimensions. *Trends in Cognitive Sciences*, *12*(12), 455–460. <https://doi.org/10.1016/j.tics.2008.10.001>
- Törnberg, P. (2022). How digital media drive affective polarization through partisan sorting. *Proceedings of the National Academy of Sciences, USA*, *119*(42), Article e2207159119. <https://doi.org/10.1073/pnas.2207159119>
- Tsvetkova, M., Yasseri, T., Meyer, E. T., Pickering, J. B., Engen, V., Walland, P., Lüders, M., Følstad, A., & Bravos, G. (2017). Understanding human-machine networks: A cross-disciplinary survey. *ACM Computing Surveys*, *50*(1), Article 12. <https://doi.org/10.1145/3039868>
- Turillazzi, A., Taddeo, M., Floridi, L., & Casolari, F. (2023). The digital services act: An analysis of its ethical, legal, and social implications. *Law, Innovation and Technology*, *15*(1), 83–106. <https://doi.org/10.1080/17579961.2023.2184136>
- Twenge, J. M. (2020). Increases in depression, self-harm, and suicide among U.S. adolescents after 2012 and links to technology use: Possible mechanisms. *Psychiatric Research and Clinical Practice*, *2*(1), 19–25. <https://doi.org/10.1176/appi.prcp.20190015>
- Twenge, J. M., Haidt, J., Lozano, J., & Cummins, K. M. (2022). Specification curve analysis shows that social media use is linked to poor mental health, especially among girls. *Acta Psychologica*, *224*, Article 103512. <https://doi.org/10.1016/j.actpsy.2022.103512>
- Ugander, J., Karrer, B., Backstrom, L., & Marlow, C. (2011). *The anatomy of the Facebook social graph* (arXiv:1111.4503). arXiv. <https://doi.org/10.48550/arXiv.1111.4503>
- Van Bavel, J. J., Rathje, S., Harris, E., Robertson, C., & Sternisko, A. (2021). How social media shapes polarization. *Trends in Cognitive Sciences*, *25*(11), 913–916. <https://doi.org/10.1016/j.tics.2021.07.013>
- Vlasceanu, M., Enz, K., & Coman, A. (2018). Cognition in a social context: A social-interactionist approach to emergent phenomena. *Current Directions in Psychological Science*, *27*(5), 369–377. <https://doi.org/10.1177/0963721418769898>
- Voelkel, J. G., Stagnaro, M. N., Chu, J., Pink, S., Mernyk, J. S., Redekopp, C., Cashman, M., Druckman, J. N., Rand, D., & Willer, R. (2022). *Megastudy identifying successful interventions to strengthen Americans' democratic attitudes* (Working Paper No. 22-38). Institute for Policy Research, Northwestern University. <https://www.strengtheningdemocracychallenge.org/paper>
- Wagner, C., Strohmaier, M., Olteanu, A., Kıcıman, E., Contractor, N., & Eliassi-Rad, T. (2021). Measuring algorithmically infused societies. *Nature*, *595*(7866), 197–204. <https://doi.org/10.1038/s41586-021-03666-1>
- Walloo Media. (2022). *Facebook news feed algorithm history*. <https://wallaroomedia.com/facebook-newsfeed-algorithm-history/>
- Wilkinson, R., & Pickett, K. (2017). Inequality and mental illness. *The Lancet Psychiatry*, *4*(7), 512–513. [https://doi.org/10.1016/S2215-0366\(17\)30206-7](https://doi.org/10.1016/S2215-0366(17)30206-7)
- Zell, A. L., & Moeller, L. (2018). Are you happy for me . . . on Facebook? The potential importance of “likes” and comments. *Computers in Human Behavior*, *78*, 26–33. <https://doi.org/10.1016/j.chb.2017.08.050>
- Zhang, M. R., Lukoff, K., Rao, R., Baughan, A., & Hiniker, A. (2022). Monitoring screen time or redesigning it? Two approaches to supporting intentional social media use. In S. Barbosa, C. Lampe, C. Appert, D. A. Shamma, S. Drucker, J. Williamson, & K. Yatani (Eds.), *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems* (pp. 1–19). <https://doi.org/10.1145/3491102.3517722>