***Citation:**
Brady, W.J., Jackson, J.C., Lindström, B, & Crockett, M.J. (2023). Algorithm-mediated social learning in online social networks. *Trends in Cognitive Sciences*, 27(10), 947-960.

Algorithm-Mediated Social Learning in Online Social Networks

William J. Brady[1]*, Joshua Conrad Jackson[1], Björn Lindström[2] & M.J. Crockett[3,4]

[1]Northwestern University, Kellogg School of Management
[2]Karolinska Institutet, Department of Clinical Neuroscience
[3]Princeton University, Department of Psychology
[4]Princeton University, University Center for Human Values

*Correspondence to: william.brady@kellogg.northwestern.edu,
website: williamjbrady.com
twitter: @william__brady

Keywords: algorithms, social learning, social media, social networks, norms

Abstract

Human social learning is increasingly occurring on online social platforms, such as Twitter, Facebook, and TikTok. On these platforms, algorithms exploit existing social learning biases (i.e., towards PRestigious, Ingroup, Moral, and Emotional information, or PRIME information) to sustain users' attention and maximize engagement on platforms. Here, we synthesize emerging insights into 'algorithm-mediated social learning' and propose a framework that examines its consequences in terms of functional misalignment. We suggest that when social learning biases are exploited by algorithms, PRIME information becomes amplified via human-algorithm interactions in the digital social environment in ways that cause social misperceptions, conflict, and spread misinformation. We discuss solutions for reducing functional misalignment including algorithms promoting bounded diversification and increasing transparency of algorithmic amplification.

Highlights

- Humans are increasingly interacting in environments mediated by algorithms that control the flow of social information, yet little is known about how algorithms impact social learning.

- Algorithm-mediated social learning is currently characterized by functional misalignment: Human social learning evolved to promote adaptive behaviors that foster cooperation and collective problem-solving, but content algorithms are designed to sustain attention and engagement on platforms.

- Emerging evidence suggests that content algorithms exploit social learning biases by amplifying prestigious, ingroup, moral and emotional (PRIME) information and teaching users to produce more of this content via social learning.

- In specific contexts like morality and politics, these human-algorithm interactions saturate the environment with PRIME information, which leads to social misperceptions that can promote conflict and misinformation rather than cooperation and collective problem-solving.

- The framework of functional misalignment can shed light on how to design algorithms that foster more functional social learning in digital environments.

Social learning in the digital age

Humans rely on social learning to navigate the world. We observe others, copy their behavior, infer their goals and intentions, and notice whether our own and others' behavior is punished or praised [1–3]. In the digital age, social learning is increasingly taking place in online social networks hosted on platforms such as Facebook, Twitter, and TikTok. Social learning on these platforms has an important non-human dimension because of *content algorithms* that manage what information we see and how we see it. *Algorithm-mediated* social learning means that, for the first time in history, much of what we learn and how we learn is influenced by content algorithms designed by corporations [4–6].

Algorithm-mediated social learning has far-reaching implications. Algorithms impact how we encounter moral and political issues on Twitter and Facebook[i,ii][7,8], how videos go viral on TikTok [9], and which conspiracy theories are promoted on YouTube [10,11]. Algorithms filter how we learn about current events and what we find out about our friends [12]. Perhaps the most startling factor in these trends is that we still know little about *how* algorithms are impacting our social learning. The rise of algorithm-mediated social learning has been so swift that it has outpaced the scientific study of social media [13].

Here, we propose that algorithm-mediated social learning is currently characterized by a problem we call *functional misalignment.* In brief, human social learning evolved over hundreds of thousands of years to promote adaptive behaviors that allow for cooperation and collective problem-solving [3,14,15]. Thus, a key *function* of human social learning is to promote cooperation and collective problem-solving. By contrast, content algorithms have appeared in the last decade to maximize the time people spend online, in order to maximize advertising revenue [5,16–18]. In other words, the *function* of content algorithms is to maximize engagement online. We suggest that the cooperative functions of human social learning and the engagement-maximizing functions of content algorithms are currently misaligned, because engagement-maximization does not in practice promote cooperation and collective problem solving. This misalignment can yield dysfunctional human-algorithm interactions in online social networks with maladaptive consequences including social misperceptions that can exacerbate conflict and the spread of misinformation.

After reviewing evidence for functional biases in human social learning, we conceptualize and review evidence for the problem of functional misalignment in algorithm-mediated social learning, demonstrating that it involves an interaction of algorithmic amplification and two forms of human social learning: observational learning and reinforcement learning. We then show how functional misalignment may escalate to produce maladaptive cultural evolution in contexts of morality and politics. We end by discussing psychologically informed solutions for reducing functional misalignment in human-algorithm interactions via strategies that help algorithms promote more functional social learning.

Biased social learning and the problem of functional misalignment

Social learning biases

Humans do not learn from others in a uniform manner -- there are several well-documented context and content biases that have emerged to optimize the function of social learning [19,20]. Context biases describe a tendency to learn in particular contexts, and from particular kinds of individuals [19]. For example, humans tend to copy prestigious individuals [21–23] and disproportionately learn from their in-groups [24,25]. Prestige bias fosters efficient social learning from successful individuals, since markers of prestige (e.g., a large house) can indirectly signal the value of learning from the individual [21–23]. Ingroup bias can be an effective strategy in the evolution of mutual cooperation [26], and it may also help humans acquire information which is most relevant to survival in their particular ecologies [27–29].

Content biases describe a tendency to disproportionately attend to and learn from certain types of informational content in our environments [14,30]. For example, humans have an outsized attention to moralized information [31–33] and to emotionally arousing information, specifically negatively valenced information [30,34–36]. A bias towards moralized information can help human groups regulate social norms and stigmatize norm violators [37,38], whereas a bias towards negative social information may have helped us quickly detect and communicate social threats such as deceptive individuals [39]. Attending disproportionately to moralistic and negative emotional social information is especially functional when such information is relatively rare, and this kind of information is highly diagnostic, e.g., that someone is uncooperative [39–42]. In summary, human social learning demonstrates biases toward PRestigious, Ingroup, Moralized and Emotional information (PRIME information), and these biases often promote adaptive behaviors that support cooperation and collective problem-solving.

Constraints on social learning biases

However, biases toward PRIME information are only functional contingent upon specific statistical contingencies in an environment. A bias towards prestige can be misleading in environments where prestige does not meaningfully signal success. People may drive luxury cars because of a large inheritance, own big houses because of a high-risk loan, or boast large followers on Instagram because they purchased bots. In these cases, learners may actually suffer when they are biased towards learning from prestigious individuals [43]. Cult leaders, corrupt politicians, and conmen all use reputation and prestige to spread false or malicious information since they are assumed to be knowledgeable or trustworthy [14].

Biases towards the ingroup and to negative moralized information also have functional limits. Ingroup bias may lose its functionality when environments become more heterogeneous and identities are formed by arbitrary social divisions (e.g. race [44]). Under these conditions, solely attending to ingroup information can lead to false consensus [45], discrimination [46] and more generally forgoing learning opportunities from the outgroup. Negative moralistic information may also become dysfunctional if it becomes more commonplace in social environments. If people regularly accuse one another of immorality, for example, these accusations may become less useful for diagnosing norm violations and preserving cooperation [47–49].

In summary, social learning biases that draw us toward PRIME information are generally functional, but they become maladaptive when PRIME information is overrepresented in the environment. When everyone claims to be prestigious, it is difficult to know whom to learn from. When all disagreement is framed as group division, it is easy to unreasonably escalate conflicts. And when dialogue is frequently negative, emotional, and moral, it can be difficult to distinguish the heinous from the merely disagreeable. Flooding people with PRIME information may be the best way to debase the value of our biased attention to this information. Content algorithms may be overloading people with PRIME information in just this way.

The problem of functional misalignment

We use the term “functional misalignment” to describe how algorithms exploit social learning biases to amplify PRIME information to the point where those biases are no longer useful in promoting cooperation and collective problem-solving (i.e., they are no longer functional). Although there are many cases where algorithms can promote adaptive cooperation and collaboration (Box 1), here we focus on learning contexts (e.g., discussions about morality and politics) where the amplification of PRIME information causes social misperceptions associated with increased conflict and the spread of misinformation. As we will argue, this consequence is not simply the result of algorithms amplifying PRIME information, but rather the interaction of algorithmic amplification and humans learning to produce more PRIME information by observing the outputs of algorithm amplification (observational learning) and being rewarded by others due to their own social learning biases (reinforcement learning).

In summary, when content algorithms exploit social learning biases, a feedback loop of human-algorithm interaction occurs that over-represents PRIME information in the environment and promotes social misperceptions that can lead to conflict and misinformation rather than cooperation and collective problem-solving (see **Fig. 1**).

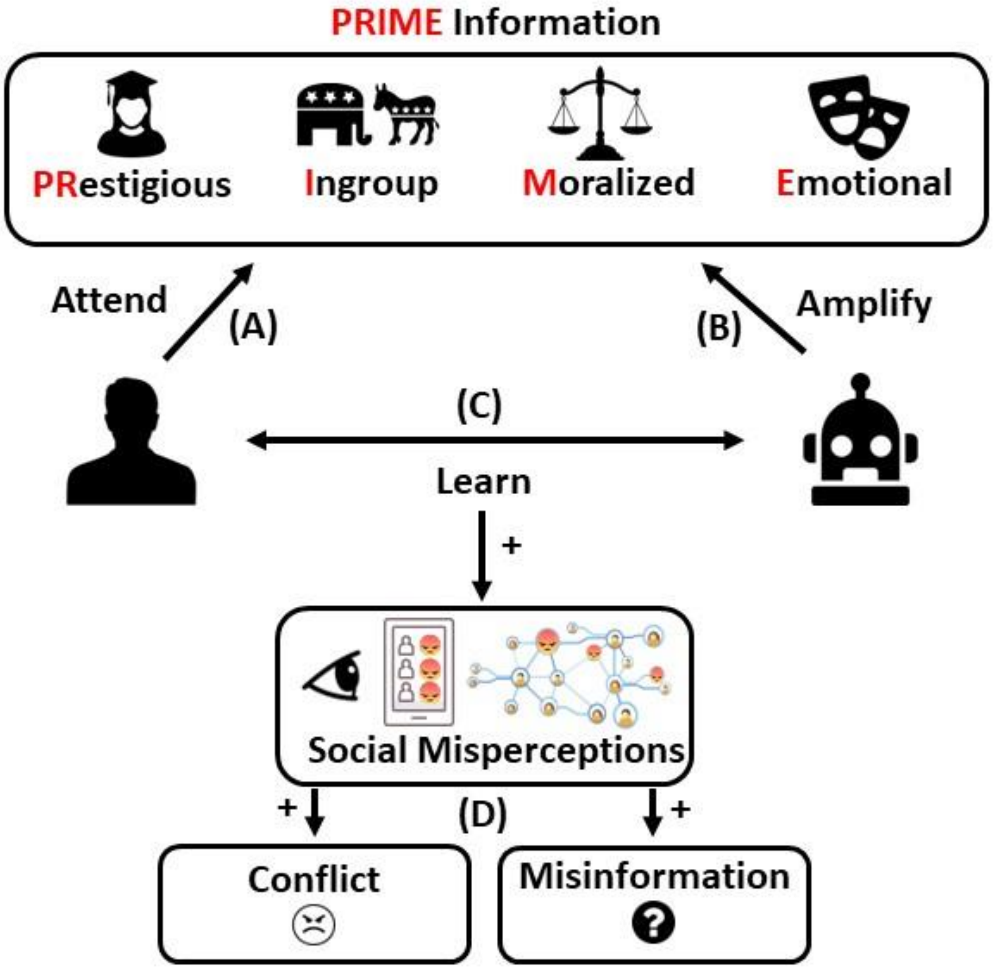


Fig. 1. Algorithm-mediated social learning in online social networks. During algorithm--mediated social learning, content algorithms on social media platforms exploit human social--learning biases and amplify prestigious, ingroup, moral and emotional ('PRIME') information as a side-effect of goals to maximize engagement on the platforms. Humans naturally attend to PRIME content (A) and learn to produce more of it when content algorithms amplify PRIME information (e.g., via observational and reinforcement learning) (B-C), thus creating a digital environment that inflates (relative to no algorithm influence) the amount of PRIME information in many contexts, such as discussions of morality and politics (D) When PRIME information oversaturates digital environments, it can increase conflict and facilitate the spread of misinformation, even when biases toward PRIME information typically facilitate cooperation and collective problem-solving

Content algorithms exploit social learning biases

As of this writing, there are now over 4 billion social media users worldwide[iii]. Facebook users upload more than 300 million photos per day and post 510,000 comments per minute, and there are 500 million Twitter posts per day[iv, v]. Users clearly do not have the time or attention to view all of these posts. Content algorithms must select the information we see and decide what kind of information to amplify [50].

Content algorithms systematically exploit human social learning biases because they are designed to optimize attentional capture and engagement time on the platform, and social learning biases strongly predict what users will want to see. Given that major social media platforms derive nearly all of their revenue from advertisements[vi, vii], algorithms are designed to amplify content that sustains user attention and keeps users on social media platforms to maximize advertising revenue profits [5,16–18]. Since humans are biased to attend to, remember and transmit PRIME information, it is not surprising that algorithms trained on human preferences end up amplifying PRIME information in the digital environment.

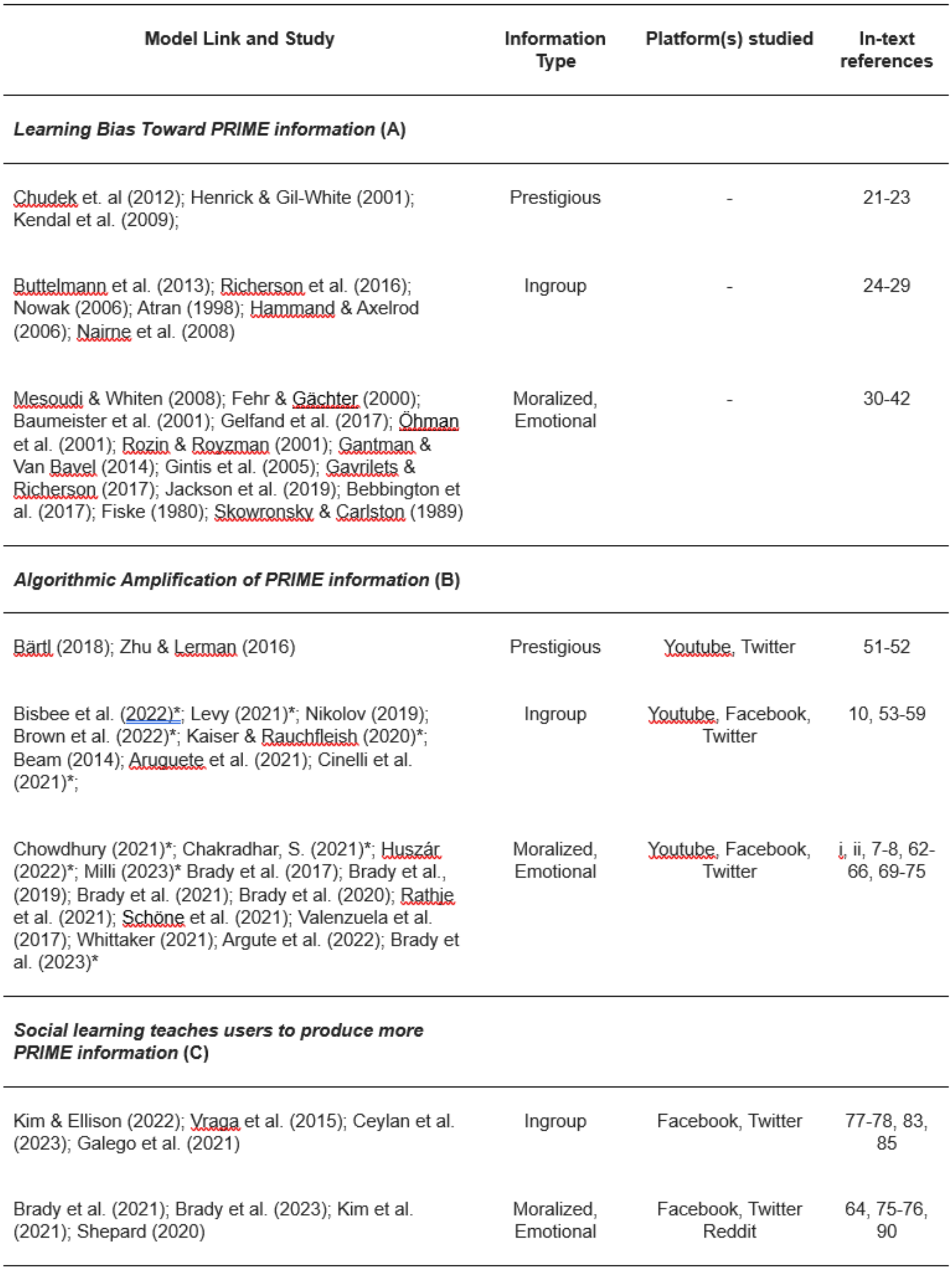
There is growing evidence of algorithmic amplification of PRIME content. Recent research suggests that algorithms amplify prestigious individuals on social media more than non-prestigious individuals. One study found that 3% of YouTube channels captured 85% of the site's total viewership [51]. Another found that the top 20% of all Twitter users own 96% of all followers, 93% of all retweets, and 93% of all mentions [52].

Several studies have also found that content algorithms amplify ingroup information. The Facebook algorithm demotes out-partisan news stories compared to an unsorted feed [53], and broader studies of multiple social media sites find amplification of in-partisan content as well [54]. YouTube users who were most skeptical of the 2020 election results were three times more likely to be recommended videos that questioned the legitimacy of the 2020 election than users who did not express skepticism [10]. Similarly, the longer that partisan users follow recommendations of the Youtube algorithm, the more their videos become biased toward their own partisan identity [55,56]. Search engine algorithms also biased users' exposure to news containing in-partisan messages compared to when users manually selected the news they would like to view [57].

In many cases, social media algorithms expose people to ingroup values and opinions that are even stronger than their own positions [58,59]. While there are several studies that find social media also exposes users to out-partisan information, thus countering the simple idea of “informational echo-chambers'' [60], many of these studies use self-report data and conflict with digital trace data studies that find consistent evidence of ingroup bias [61]. Furthermore, the out-partisan information we are exposed to is often colored with ingroup commentary [5].

A growing number of studies have also demonstrated that moralized and emotional information is highly likely to spread through online social network platforms [5,62–71]. Both independent research using randomized controlled trials and internal company research have documented how content algorithms specifically amplify moralized and extreme political content[i, ii] [7,8,72], to the point where users believe content is more socially representative than it really is [73,74].

Taken together, the studies reviewed suggest that algorithms exploit social learning biases by amplifying PRIME information (See **Table 1**). As a result, PRIME information is overrepresented in the digital environment, which also provides input to human social learning unfolding in response to algorithm behavior in online social networks.



Model Link and Study	Information Type	Platform(s) studied	In-text references
***Learning Bias Toward PRIME information* (A)**			
Chudek et. al (2012); Henrick & Gil-White (2001); Kendal et al. (2009);	Prestigious	-	21-23
Buttelmann et al. (2013); Richerson et al. (2016); Nowak (2006); Atran (1998); Hammand & Axelrod (2006); Nairne et al. (2008)	Ingroup	-	24-29
Mesoudi & Whiten (2008); Fehr & Gächter (2000); Baumeister et al. (2001); Gelfand et al. (2017); Öhman et al. (2001); Rozin & Royzman (2001); Gantman & Van Bavel (2014); Gintis et al. (2005); Gavrilets & Richerson (2017); Jackson et al. (2019); Bebbington et al. (2017); Fiske (1980); Skowronsky & Carlston (1989)	Moralized, Emotional	-	30-42
***Algorithmic Amplification of PRIME information* (B)**			
Bärtl (2018); Zhu & Lerman (2016)	Prestigious	Youtube, Twitter	51-52
Bisbee et al. (2022)*; Levy (2021)*; Nikolov (2019); Brown et al. (2022)*; Kaiser & Rauchfleish (2020)*; Beam (2014); Aruguete et al. (2021); Cinelli et al. (2021)*;	Ingroup	Youtube, Facebook, Twitter	10, 53-59
Chowdhury (2021)*; Chakradhar, S. (2021)*; Huszár (2022)*; Milli (2023)* Brady et al. (2017); Brady et al., (2019); Brady et al. (2021); Brady et al. (2020); Rathje et al. (2021); Schöne et al. (2021); Valenzuela et al. (2017); Whittaker (2021); Argute et al. (2022); Brady et al. (2023)*	Moralized, Emotional	Youtube, Facebook, Twitter	i, ii, 7-8, 62-66, 69-75
***Social learning teaches users to produce more PRIME information* (C)**			
Kim & Ellison (2022); Vraga et al. (2015); Ceylan et al. (2023); Galego et al. (2021)	Ingroup	Facebook, Twitter	77-78, 83, 85
Brady et al. (2021); Brady et al. (2023); Kim et al. (2021); Shepard (2020)	Moralized, Emotional	Facebook, Twitter Reddit	64, 75-76, 90

Table 1. Summary of empirical evidence for components of the algorithm-mediated social learning model. Letters (A) – (C) refer to model links designed in Fig. 1. *In the section, 'algorithmic amplification of PRIME information', studies marked with an asterisk attempt to disentangle correlated effects of algorithm amplification vs. user preferences through experimentation (e.g., using randomized control trials on social media or behavioral experiments in mock social media environments) or observation (e.g., observing what content is recommended to users by their personalized feeds); otherwise studies focus on macro-level patterns of engagement where algorithm amplification and user preferences are correlated.

Feedback loops in algorithm-mediated social learning

Observational learning

When content algorithms exploit human social learning biases and amplify PRIME information, they do not only increase our exposure to this information; they also change what we think is appropriate in our online networks via *observational learning* [5,64]. Observational learning describes how people alter their behavior in response to watching others, and is especially helpful for teaching us what kinds of behaviors are socially appropriate and representative (i.e., common) in a social group [1–3]. By increasing the probability that a user observes PRIME information, content algorithms should also lead users to perceive PRIME information as more normative and representative.

Several recent studies show that algorithm-mediated observational learning teaches users to transmit more PRIME information. For instance, Twitter users' outrage expression is predicted by the amount of outrage they observe in their social network , and has a causal impact on decisions to express outrage in a message [64]. Similarly, experimentally manipulating exposure to toxic comments in Facebook news pages increased a users' own toxic comments [75]. Another study across multiple platforms found that social media users' observation of political behaviors in their feed predicted their own political posting behavior as well as offline political behavior [76]. Qualitative studies of Facebook users found that their willingness to share political information was predicted by observations of norms regarding political content based on their feed [77]. These studies show that content algorithms, by encouraging us to view PRIME information, also encourage us to produce PRIME information via observational learning.

Reinforcement learning

Content algorithms also reward users who post PRIME information with more exposure. This is important, because social feedback via "likes" and "shares" drives future social media activity [64,78]. Perhaps more importantly, algorithms also deliver posts containing PRIME information to a broader audience, including people from outside the user's direct social network (e.g. "context collapse" [79]). By rewarding

PRIME information through exposure and feedback, content algorithms teach users to produce more of this content through *reinforcement learning*, or learning that occurs in response to positive or negative social feedback [80]. When a user posts PRIME information, they are more likely to get socially rewarded, and thus learn to post more PRIME information.

Several studies have now demonstrated that the algorithm-mediated delivery of social feedback on social media platforms directly shapes user's behavior via reinforcement learning [64,78,81–83]. For instance, Instagram and web forum users' posting behavior can be predicted as a direct function of social reward received in their posting history [78]. Similarly, Reddit users are more likely to post to communities that previously gave them more positive social feedback [83], and Facebook users are likely to post more regularly and at faster intervals in order to gain positive social feedback [81].

Similar reinforcement learning dynamics have been demonstrated specifically for learning to post more PRIME information over time. For example, Twitter users' moral outrage could be predicted by whether users had received positive feedback in response to their previous outrage posts [64]. Furthermore, in a mock social media environment, manipulating positive feedback for outrage posts increased users' likelihood of expressing outrage over partisan political issues [64]. Even political leaders' decisions to post PRIME information are sensitive to reinforcement learning: Spanish politicians' Twitter accounts systematically changed the political issues they discussed based on social feedback received from previous posts[viii].

Gaming

Observational and reinforcement learning can create feedback loops of amplified PRIME information without any user awareness that they are participating in this process, but some users intentionally alter their social media activity based on learned knowledge of algorithmic amplification. This behavior is referred to as *gaming*, in which users manipulate their post content so that algorithms are more likely to feature the content on news feeds or search engine results. Gaming is relatively rare because most people are not aware of how algorithms work. Those who do tend to be younger and more educated [84–86]. They also tend to be individuals engaging in personal promotion or political persuasion [87,88]. Because the knowledge required to game algorithms is rare, yet these users' content will get amplified and have an outsized influence on social media feeds, highly motivated extreme political users can exacerbate feedback loops of amplifying PRIME information. This process may be a key mechanism through which extreme political viewpoints gain exposure online [88].

In summary, emerging evidence shows how algorithm-mediated social learning directly shapes user posting behavior, including how often users post, when they choose to post, and their decisions to post PRIME information. Our discussion of feedback loops implies that algorithm-mediated social learning should continually produce more PRIME information over time, and there is some evidence for increased negativity since the advent of algorithms [89]. However, PRIME information is unlikely to make up the entire information landscape even though the feedback loops described above amplify it, see Box 2.

Algorithm-mediated social learning produces social misperceptions

When PRIME information is artificially amplified via human-algorithm interaction, a key consequence is that PRIME information is *overrepresented* relative to the true base rate in users' social networks (i.e., people see more PRIME content than they would if they randomly sampled content from people in their immediate social network). For example, several studies spanning different social media platforms found that negative emotional, moralized and politically extreme information is overrepresented on the platform as a result of the interaction of human preferences and algorithmic amplification. Because such information is promoted by algorithms, users perceive it to be more socially representative of dialogue on social media than it really is [74,90].

Overrepresentation of negative moralized content in the context of politics is a problem because this content is only generated by a minority of users [91], yet users may infer it is more common based on its representation in news feeds. For instance, recent work found that when moral outrage is overrepresented in users' social media feeds, it increases their perception of their social network's affective polarization, norms of outrage expression, and ideological extremity [74]. Overrepresentation of negative moralized content may be a key process by which conflict is exacerbated online: when we overestimate the extent to which our ingroup or outgroup feel negatively toward each other, it increases intergroup conflict [92–94]. Furthermore, recent experimental studies have demonstrated that when content algorithms overrepresent ingroup political information it increases polarization [53]; but see also [95].

If algorithm-mediated social learning leads us to misperceive PRIME information as more prevalent than it really is, this may create an environment that facilitates the spread of misinformation. Recent work found that viral misinformation exploits emotionality [96,97], and that users are less discerning of fake news and are more likely to share fake news when they rely on their immediate emotional responses rather than deliberating about the news content [98,99].

Misinformation profiteers are especially likely to exploit moral outrage and anger because it helps them spread content widely. Feeling outrage also makes people more motivated to share news regardless of its accuracy [100]. Bots and troll farms originating from several countries also specifically use moralized, emotional and ingroup information to sow discontent and misinformation across several platforms [101]. Furthermore, the advent of deepfakes - videos that use neural networks to manipulate the face of individuals - has shown how misinformation can spread by exploiting prestige bias (i.e., when users make a famous or important person produce information that is false or manipulative [102,103]). Taken together, recent work has documented that the algorithmic amplification of PRIME information can lead to social misperceptions that create environments ripe for intergroup conflict and the spread of misinformation.

Functional misalignment and cultural evolution in the digital age

In theories of cultural evolution, social learning is the engine of cultural transmission, and many theories argue that the social learning biases described above are key to cumulative cultural evolution, when culture builds on itself over time [14,104]. Throughout this article, we have suggested that algorithm-mediated social learning can oversaturate digital environments with PRIME content in ways that lead to social misperceptions. We have focused so far on the immediate consequences of promoting PRIME content for cooperation and emotional well-being, but promoting this content could also produce harmful longer-term patterns of cultural evolution.

Algorithm-mediated social learning may, for example, encourage *tipping points* that promote extreme social and political norms. Culture does not change linearly, but rather by pulses and pauses including tipping points which catalyze major cultural innovation [105]. Tipping points typically result when useful norms are adopted en masse, but even fringe norms can rapidly spread through a population when a small group of devoted individuals consistently pushes an extreme belief or narrative [106,107].

Algorithms may encourage these kinds of unhealthy tipping points by promoting and rewarding fringe information because it activates our social learning biases. There are now several documented examples of political groups using tipping points for spreading cultural narratives online (Doroshenko & Tu, 2022; Guess et al., 2020; King et al., 2017). In algorithm-mediated social learning environments, users may perceive these extreme narratives as more common and normative than they really are [74]. For example, Trump voters were much more likely to be presented with extremist views about fraud in the 2020 election by content algorithms (Bisbee et al., 2022). Seeing this promoted content, especially if it is associated with prestigious individuals (e.g., senators or House representatives), may encourage users to infer that beliefs about election fraud are widespread and held with little doubt, increasing the likelihood that users adopt these beliefs.

These algorithm-influenced tipping points are concerning because they may accelerate the spread of fringe theories and misinformation. These processes also implicate other aspects of cultural evolution, for instance raising the possibility that algorithm-mediated social learning could impact the efficacy of moralistic norms [47,111], see Outstanding Questions.

Aligning algorithms with functional social learning

Given that content algorithms are currently misaligned with functional human social learning in their design, both users and social media companies should be motivated to re-align the goals of algorithms and human social learners because it can improve users' experience. Many social media users are "exhausted" by the growing spread of PRIME information on social media [112,113], and specifically think that divisive moral and emotional content should *not* be amplified on social media [113]. In this section, we outline two strategies for honoring this preference and improving algorithm-mediated social learning.

Bounded informational diversification

One path to improving algorithm-mediated social learning could involve changing how content algorithms are designed. Content algorithms have a range of benefits (Box 1), but they may also benefit from modifications. For example, algorithms could seek to amplify more diverse information. Models of collective problem solving show that maintaining diversity can improve problem-solving quality [114–116]. One reason why algorithms perform so well at games like Chess and Go is because they have learned a diverse set of human strategies, which they can use to suggest optimal solutions [117,118]. Humans may also benefit from content algorithms which increase diverse content. Recent research shows that exposing people to viewpoints from outside their immediate social network can reduce personal bias and groupthink in mundane decision-making games about colors, especially among motivated actors [119].

However, diversifying social media may not be as simple as winning a game of Chess, and algorithms that blindly increase content diversity may backfire in several ways. These algorithms could unintentionally amplify fringe political beliefs on the feeds of politically moderate users. They could also increase disagreement and partisan sorting. Research shows that partisan conflict does not result from "echo chambers," but precisely the opposite: conflict arises when people who hold diverse political views discuss highly divisive topics in the same space [95,120]. This evidence suggests that wholesale diversification may not be a good strategy for content algorithms. Wholesale diversification may work well when people play games with well-defined structures, but it may fail in important contexts where people discuss moral and political topics.

We suggest that *bounded informational diversification* might overcome this limitation. In other words, social media algorithms should diversify content along specific theory-informed dimensions–the very same dimensions that make up our PRIME model. Throughout this paper, we have argued that algorithms implicitly amplify PRIME content by learning from human attentional biases. We suggest that an effective design-centered approach could explicitly penalize PRIME content to counteract human attentional biases. Newsfeeds could still prioritize posts from strong and weak ties within users' social network, but these posts would be less inflammatory, less tribalist, less outrage-inducing, and would be more representative of how people communicate in the real world.

A limitation of this approach is that it may still maintain "echo-chamber" interactions because of issues such as partisan sorting. Nevertheless, the alternative of amplifying cross-partisan interactions online may exacerbate conflict more than mitigate it [95,120], see also Outstanding Questions).

Transparency of algorithmic influence

In addition to *design-centered approaches*, it may be equally important to develop *person-centered* solutions that empower users to make decisions that can improve their social learning without relying on changes being made by platforms [5,121]. We suggest that a key hybrid design/person-centered solution is to increase transparency of algorithmic influence specifically to show users how algorithms influence the social information they see. This may be as simple as defining the reason why a post was promoted (e.g., because it comes from a close social tie, or because the algorithm deems a post to have engaging content).

Increasing algorithm transparency is especially important because the majority of social media users do not understand how algorithms affect the content they see [84–86], and because public demand for interpretable algorithms is particularly high for applications concerning morality and fairness [122]. A better working knowledge of how algorithms impact social information could facilitate users adjusting their social inferences developed by observational and reinforcement learning.

Private companies may never develop algorithms that fully align with human cooperation and problem-solving, because these companies are motivated by profit [128]. But algorithm transparency can mitigate this conflict of interest. An even more ambitious goal could be to give people full *control* over the algorithms that personalize their content. However, allowing people full control over algorithms could simply amplify existing social learning biases as in the case of current algorithms that maximize attentional capture, but there are ways to allow users to intentionally change algorithms without having full control such as in the case of selective filtering [124]. If users want to avoid politically polarized content, this may be as simple as selecting an algorithm that does not promote outrage-inducing posts.

Concluding Remarks

In this review, we covered emerging evidence suggesting that content algorithms exploit human social learning biases toward PRIME information to sustain attention and engagement with platforms. By promoting PRIME information, algorithms teach users to express more of this information themselves via observational learning and reinforcement learning. We argued these human-algorithmic interactions are a case of functional misalignment because they produce a digital environment that overrepresents PRIME information to the point where it fuels conflict and misinformation rather than cooperation and collective problem-solving. Our framework also suggests that human-algorithm interactions can better support functional social learning by increasing the amplification of bounded diverse information and increasing transparency of algorithm influence.

Our functional misalignment perspective generates several future research directions (see also Outstanding Questions). First, it is important to better understand exactly how much variance is explained by content algorithms vs. human social learning in their joint influence on people's behavior on social media [5,125]. We encourage cross-disciplinary studies to investigate this interaction more precisely through field experiments, laboratory experiments, and computational models [126]. We also encourage research that leverages the interaction of human social learning biases and content algorithms to enhance interactions on social media (e.g. fostering accurate social inferences and diverse interactions). More broadly, as content algorithms increasingly dominate our access to information, the functional misalignment perspective highlights how small design decisions can have large emergent consequences due to complex interactions between algorithms and human social learning mechanisms. Major efforts including academic-industry collaborations will be required to better model the dynamics of what we learn from algorithms and what algorithms learn from us.

Outstanding questions

- How does algorithm-mediated social learning impact the efficacy of moralistic norms? If moralistic information is overrepresented in the environment, it may become more difficult for us to choose which moral issues are most worthy of our efforts.
- How can algorithm-mediated social learning be leveraged to spark sustained collective action? We have seen cases such as #MeToo and #BlackLivesMatter bring attention to collective action, yet the extent to which these digital norms are sustained offline remains unclear.
- How will greater algorithmic transparency impact our social learning processes and our awareness of them? It is an open empirical question whether people will adjust their social inferences when they have a greater understanding of the information that is most likely to be amplified by algorithms.
- What are the neural underpinnings of algorithm-mediated social learning? When PRIME content saturates our digital environment, how do brain systems key for reward and motivation, like the midbrain dopaminergic system, respond to such information?

- How can we best study the interaction of algorithmic amplification and social learning in one paradigm? Studies that simply manipulate algorithm selection or study a users' learning history in isolation are unlikely to accurately estimate the effects of algorithm-mediated social learning.
- Since a users' exposure to algorithms and learning history is highly variable, which populations of individuals are best to study in order to isolate effects of algorithmic amplification vs. social learning?
- As there is large social media user heterogeneity, what types of users are most likely to have algorithm-mediated social learning lead to social misperceptions?
- What time scale is required to detect the impact of content algorithms on our attitudes and behaviors?
- How do small differences in content algorithm design and differing social norms in a network affect outcomes of algorithm-mediated social learning?

Box 1: Benefits of algorithm-mediated social learning

Although we highlight the problem of functional misalignment that arises from algorithm-mediated social learning, there are also contexts in which algorithms complement or even improve social learning in online networks. Content algorithms are very good at recommending and amplifying like-minded people who share our interests [13], and thus help us increase our *relational mobility* [127]. Greater relational mobility, when it is not centered around moralized identities, can help improve social learning outcomes such as trust, self-esteem and passion [128]. Content algorithms also amplify the most popular content and information, which can bolster the *wisdom of crowds* in specific contexts that do not typically involve moralization such as stock market decision-making [129]. Finally, although content algorithms tend to amplify PRIME information, if they are designed in the right way they can also act as key *information quality filters* to downrank divisive content and misinformation that tend to hinder our social learning [130]. By identifying current areas where algorithm-mediated social learning improves learning outcomes, both scientists and practitioners can better understand how to improve content algorithms' current design.

Box 2: Constraints on feedback loops of algorithm-mediated social learning

If human-algorithmic interactions tend to increase PRIME information via feedback loops of social learning and algorithmic amplification, one question is why social media platforms are not converging to exclusively contain PRIME information overtime. Although there is evidence that online social information governed by algorithms is increasing in negativity over time [70,89], obviously PRIME information is not the only content represented in online social networks. There are at least three processes that constrain the human-algorithmic interactions that amplify PRIME content. First, certain contexts have different *norms of expression* that make it more appropriate to express positive emotions or less divisive content, thus constraining the spread of PRIME content in those contexts. For example, certain moralized events go viral with positivity because of the celebration surrounding the event [62] (e.g., #lovewins after the U.S. Supreme Court decision to legalize same-sex marriage) and certain cultures are more influenced by positive emotions because of different emotion norms [131]. Second, although we focus on key human social learning biases, there are also other *competing attention biases* driving online behavior that can compete with biases toward PRIME information. For instance, we are often drawn to surprising content which can be positive and may not be prestigious, negative or moralistic at all [97,132]. Finally, although humans are inherently social, sometimes we fail to use social information when it is available due to *inefficient social learning*. We are less likely to use social information when we have conflicting prior beliefs or when transmission is noisy [133]. Thus, algorithm-mediated social learning alone cannot explain all information dynamics online because at times we discount social information all together. In summary, algorithm-mediated social learning often leads to the spread of PRIME information, yet this outcome is constrained by varying social norms of emotion expression, competing attention biases, and inefficient social learning.

Acknowledgements

We thank the Marketing Department at Northwestern University, Kellogg School of Management for feedback on ideas included in this manuscript. We also thank the University of Maryland Social, Decision and Organizational Sciences seminar series, members of the Center for Humans and Machines at the Max Planck Institute for Human Development, and members of Northwestern Institute on Complex Systems (NICO) for feedback. We also thank Arvind Narayanan, Katherine Glenn Bass, and the participants at the ‘Optimizing for What?’ conference hosted by the Knight First Amendment Institute at Columbia University for feedback on earlier drafts. We thank the Stanford AI group hosted by Michael Bernstein for feedback on earlier drafts. William Brady and Joshua Jackson thank the ChatGPT algorithm for complimenting our learning abilities and helping us choose the acronym “PRIME”. William Brady thanks Andy Stott and Gunnar Haslam for their music that was used during the writing of this manuscript.

Declaration of Generative AI and AI-assisted technologies in the writing process

During the preparation of this work, the authors used ChatGPT in order to help come up with the acronym “PRIME”. After using this tool, the authors reviewed and edited the acronym as needed and take full responsibility of the content of the publication.

Resources

i. Chowdhury, R. (2021) Examining algorithmic amplification of political content on Twitter[Online]. Available: https://blog.twitter.com/en_us/topics/company/2021/rml-politicalcontent. [Accessed: 01-Aug-2022]

ii. Chakradhar, S. (2021) More internal documents show how Facebook's algorithm prioritized anger and posts that triggered it. Nieman Lab.

iii. Kepios (2022) Global Social Media Statistics. DataReportal – Global Digital Insights. [Online]. Available: https://datareportal.com/social-media-users. [Accessed: 27-Nov-2022]

iv. Beveridge, C. (2022) 33 Twitter Statistics That Matter to Marketers in 2022. Social Media Marketing & Management Dashboard.

v. Zephoria (2021) Top 15 Facebook Statistics for 2020 - The Year in Review. Zephoria Inc.

vi. Dixon, S. (2022) Global Meta advertising revenue 2021. Statista. [Online]. Available: https://www.statista.com/statistics/271258/facebooks-advertising-revenue-worldwide/. [Accessed: 08-Nov-2022]

vii. Iqbal, M. (2022) Twitter Revenue and Usage Statistics (2022). Business of Apps. [Online]. Available: https://www.businessofapps.com/data/twitter-statistics/. [Accessed: 08-Nov-2022]

viii. Gallego, A. et al. (2021) Politician-citizen interactions and dynamic representation: Evidence from Twitter. Econ. Work. Pap. at <https://ideas.repec.org//p/upf/upfgen/1769.html>

References

1. Bandura, A. (1986) Social foundations of thought and action. *Englewood Cliffs NJ* 1986
2. Ho, M.K. *et al.* (2017) Social is special: A normative framework for teaching with and learning from evaluative feedback. *Cognition* 167, 91–106
3. Olsson, A. *et al.* (2020) The neural and computational systems of social learning. *Nat. Rev. Neurosci.* 21, 197–212
4. Airoldi, M. *et al.* (2016) Follow the algorithm: An exploratory investigation of music on YouTube. *Poetics* 57, 1–13
5. Brady, W.J. *et al.* (2020) The MAD Model of Moral Contagion: The Role of Motivation, Attention, and Design in the Spread of Moralized Content Online. *Perspect. Psychol. Sci.* 15, 978–1010
6. Entman, R.M. and Usher, N. (2018) Framing in a Fractured Democracy: Impacts of Digital Technology on Ideology, Power and Cascading Network Activation. *J. Commun.* 68, 298–308
7. Huszár, F. *et al.* (2022) Algorithmic amplification of politics on Twitter. *Proc. Natl. Acad. Sci.* 119, e2025334119
8. Milli, S. *et al.* Twitter's Algorithm: Amplifying Anger, Animosity, and Affective

Polarization. DOI: https://doi.org/10.1177/0956797615594620
9. Arthurs, J. *et al.* (2018) Researching YouTube. *Convergence* 24, 3–15
10. Bisbee, J. *et al.* (2022) Election Fraud, YouTube, and Public Perception of the Legitimacy of President Biden. *J. Online Trust Saf.* 1
11. Schmitt, J.B. *et al.* (2018) Counter-messages as Prevention or Promotion of Extremism?! The Potential Role of YouTube: Recommendation Algorithms. *J. Commun.* 68, 780–808
12. Latzer, M. *et al.* (2016) The economics of algorithmic selection on the Internet. *Handb. Econ. Internet*
13. Acerbi, A. (2019) *Cultural Evolution in the Digital Age*, Oxford University Press
14. Boyd, R. and Richerson, P.J. (1988) *Culture and the Evolutionary Process*, University of Chicago Press
15. Heyes, C. (2021) Is morality a gadget? Nature, nurture and culture in moral development. *Synthese* 198, 4391–4414
16. Christian, B. (2020) *The Alignment Problem: Machine Learning and Human Values*, National Geographic Books
17. Williams, J. (2018) *Stand Out of Our Light: Freedom and Resistance in the Attention Economy*, Cambridge University Press
18. Zhang, Xiaochen *et al.* (2021) Welfare properties of profit maximizing recommender systems. *MIS Q.* 45, 1–28
19. Henrich, J. and McElreath, R. (2003) The evolution of cultural evolution. *Evol. Anthropol. Issues News Rev.* 12, 123–135
20. Kendal, R.L. *et al.* (2018) Social Learning Strategies: Bridge-Building between Fields. *Trends Cogn. Sci.* 22, 651–665
21. Chudek, M. *et al.* (2012) Prestige-biased cultural learning: bystander’s differential attention to potential models influences children’s learning. *Evol. Hum. Behav.* 33, 46–56
22. Henrich, J. and Gil-White, F.J. (2001) The evolution of prestige: freely conferred deference as a mechanism for enhancing the benefits of cultural transmission. *Evol. Hum. Behav.* 22, 165–196
23. Kendal, J. *et al.* (2009) The evolution of social learning rules: Payoff-biased and frequency-dependent biased transmission. *J. Theor. Biol.* 260, 210–219
24. Buttelmann, D. *et al.* (2013) Selective imitation of in-group over out-group members in 14-month-old infants. *Child Dev.* 84, 422–428
25. Richerson, P. *et al.* (2016) Cultural group selection plays an essential role in explaining human cooperation: A sketch of the evidence. *Behav. Brain Sci.* 39, e30
26. Nowak, M.A. (2006) Five Rules for the Evolution of Cooperation. *Science* 314, 1560–1563
27. Atran, S. (1998) Folk biology and the anthropology of science: cognitive universals and cultural particulars. *Behav. Brain Sci.* 21, 547–569; discussion 569-609
28. Hammond, R.A. and Axelrod, R. (2006) The Evolution of Ethnocentrism. *J. Confl. Resolut.* 50, 926–936
29. Nairne, J.S. *et al.* (2008) Adaptive memory: the comparative value of survival processing. *Psychol. Sci.* 19, 176–180
30. Mesoudi, A. and Whiten, A. (2008) The multiple roles of cultural transmission experiments in understanding human cultural evolution. *Philos. Trans. R. Soc. B*

Biol. Sci. 363, 3489–3501

31. Fehr, E. and Gächter, S. (2000) Cooperation and Punishment in Public Goods Experiments. *Am. Econ. Rev.* 90, 980–994
32. Gantman, A.P. and Van Bavel, J.J. (2014) The moral pop-out effect: enhanced perceptual awareness of morally relevant stimuli. *Cognition* 132, 22–29
33. Gintis, H. *et al.* (2005) *Moral Sentiments and Material Interests: The Foundations of Cooperation in Economic Life*, MIT Press
34. Baumeister, R.F. *et al.* (2001) Bad is Stronger than Good. *Rev. Gen. Psychol.* 5, 323–370
35. Gelfand, M.J. *et al.* (2017) The Strength of Social Norms Across Human Groups. *Perspect. Psychol. Sci.* 12, 800–809
36. Öhman, A. *et al.* (2001) Emotion drives attention: Detecting the snake in the grass. *J. Exp. Psychol. Gen.* 130, 466–478
37. Gavrilets, S. and Richerson, P.J. (2017) Collective action and the evolution of social norm internalization. *Proc. Natl. Acad. Sci.* 114, 6068–6073
38. Jackson, J.C. *et al.* (2019) Revenge: A Multilevel Review and Synthesis. *Annu. Rev. Psychol.* 70, 319–345
39. Rozin, P. and Royzman, E.B. (2001) Negativity Bias, Negativity Dominance, and Contagion. *Personal. Soc. Psychol. Rev.* 5, 296–320
40. Bebbington, K. *et al.* (2017) The sky is falling: evidence of a negativity bias in the social transmission of information. *Evol. Hum. Behav.* 38, 92–101
41. Fiske, S.T. (1980) Attention and weight in person perception: The impact of negative and extreme behavior. *J. Pers. Soc. Psychol.* 38, 889–906
42. Skowronski, J.J. and Carlston, D.E. (1989) Negativity and extremity biases in impression formation: A review of explanations. *Psychol. Bull.* 105, 131–142
43. Mesoudi, A. (2009) How cultural evolutionary theory can inform social psychology and vice versa. *Psychol. Rev.* 116, 929–952
44. De, S. *et al.* (2015) The Inevitability of Ethnocentrism Revisited: Ethnocentrism Diminishes As Mobility Increases. *Sci. Rep.* 5, 17963
45. Marks, G. and Miller, N. (1987) Ten years of research on the false-consensus effect: An empirical and theoretical review. *Psychol. Bull.* 102, 72–90
46. Brewer, M.B. (1999) The Psychology of Prejudice: Ingroup Love and Outgroup Hate? *J. Soc. Issues* 55, 429–444
47. Brady, W.J. and Crockett, M.J. (2019) How Effective Is Online Outrage? *Trends Cogn. Sci.* 23, 79–80
48. Eriksson, K. *et al.* (2015) Bidirectional associations between descriptive and injunctive norms. *Organ. Behav. Hum. Decis. Process.* 129, 59–69
49. Lindström, B. *et al.* (2018) The role of a “common is moral” heuristic in the stability and change of moral norms. *J. Exp. Psychol. Gen.* 147, 228–242
50. Zuckerberg, M. *et al.* (2010) Dynamically providing a news feed about a user of a social network, US7669123B2
51. Bärtl, M. (2018) YouTube channels, uploads and views: A statistical analysis of the past 10 years. *Convergence* 24, 16–32
52. Zhu, L. and Lerman, K. (2016) Attention Inequality in Social Media. DOI: 10.48550/arXiv.1601.07200
53. Levy, R. (2021) Social Media, News Consumption, and Polarization: Evidence from

a Field Experiment. *Am. Econ. Rev.* 111, 831–870

54. Nikolov, D. *et al.* (2019) Quantifying Biases in Online Information Exposure. *J. Assoc. Inf. Sci. Technol.* 70, 218–229
55. Brown, M.A. *et al.* (2022) Echo Chambers, Rabbit Holes, and Algorithmic Bias: How YouTube Recommends Content to Real Users. *Available SSRN 4114905*
56. Kaiser, J. and Rauchfleisch, A. (2020) Birds of a Feather Get Recommended Together: Algorithmic Homophily in YouTube's Channel Recommendations in the United States and Germany. *Soc. Media Soc.* 6, 2056305120969914
57. Beam, M.A. (2014) Automating the News: How Personalized News Recommender System Design Choices Impact News Reception. *Commun. Res.* 41, 1019–1041
58. Aruguete, N. *et al.* (2021) News by Popular Demand: Ideological Congruence, Issue Salience, and Media Reputation in News Sharing. *Int. J. Press.* DOI: 10.1177/19401612211057068
59. Cinelli, M. *et al.* (2021) The echo chamber effect on social media. *Proc. Natl. Acad. Sci.* 118, e2023301118
60. Persily, N. and Tucker, J.A. (2020) *Social Media and Democracy: The State of the Field, Prospects for Reform*, Cambridge University Press
61. Terren, L. and Borge-Bravo, R. (2021) Echo Chambers on Social Media: A Systematic Review of the Literature. *Rev. Commun. Res.* 9, 99–118
62. Brady, W.J. *et al.* (2017) Emotion shapes the diffusion of moralized content in social networks. *Proc. Natl. Acad. Sci.* 114, 7313–7318
63. Brady, W.J. *et al.* (2019) An ideological asymmetry in the diffusion of moralized content on social media among political leaders. *J. Exp. Psychol. Gen.* 148, 1802–1813
64. Brady, W.J. *et al.* (2021) How social learning amplifies moral outrage expression in online social networks. *Sci. Adv.* 7, eabe5641
65. Brady, W.J. *et al.* (2020) Attentional capture helps explain why moral and emotional content go viral. *J. Exp. Psychol. Gen.* 149, 746–756
66. Brady, W.J. and Van Bavel, J.J. (2021) Social identity shapes antecedents and functional outcomes of moral emotion expression in online networks
67. Crockett, M.J. (2017) Moral outrage in the digital age. *Nat. Hum. Behav.* 1, 769–771
68. Goldenberg, A. and Gross, J.J. (2020) Digital Emotion Contagion. *Trends Cogn. Sci.* 24, 316–328
69. Rathje, S. *et al.* (2021) Out-group animosity drives engagement on social media. *Proc. Natl. Acad. Sci.* 118, e2024292118
70. Schöne, J.P. *et al.* (2021) Negativity Spreads More than Positivity on Twitter After Both Positive and Negative Political Situations. *Affect. Sci.* 2, 379–390
71. Valenzuela, S. *et al.* (2017) Behavioral Effects of Framing on Social Media Users: How Conflict, Economic, Human Interest, and Morality Frames Drive News Sharing. *J. Commun.* 67, 803–826
72. Whittaker, J. *et al.* (2021) Recommender systems and the amplification of extremist content. *Internet Policy Rev.* 10, 1–29
73. Arugute, N. *et al.* (2022) Network activated frames: content sharing and perceived polarization in social media. *J. Commun.* DOI: 10.1093/joc/jqac035
74. Brady, W.J. *et al.* (2023) Overperception of moral outrage in online social networks

inflates beliefs about intergroup hostility. *Nat. Hum. Behav.* DOI: 10.31219/osf.io/k5dzr

75. Kim, J.W. *et al.* (2021) The Distorting Prism of Social Media: How Self-Selection and Exposure to Incivility Fuel Online Comment Toxicity. *J. Commun.* 71, 922–946
76. Kim, D.H. and Ellison, N.B. (2022) From observation on social media to offline political participation: The social media affordances approach. *New Media Soc.* 24, 2614–2634
77. Vraga, E.K. *et al.* (2015) How individual sensitivities to disagreement shape youth political expression on Facebook. *Comput. Hum. Behav.* 45, 281–289
78. Lindström, B. *et al.* (2021) A computational reward learning account of social media engagement. *Nat. Commun.* 12, 1311
79. Marwick, A. and Ellison, N.B. (2012) “There Isn’t Wifi in Heaven!” Negotiating Visibility on Facebook Memorial Pages. *J. Broadcast. Electron. Media* 56, 378–400
80. Glimcher, P.W. (2011) Understanding dopamine and reinforcement learning: The dopamine reward prediction error hypothesis. *Proc. Natl. Acad. Sci.* 108, 15647–15654
81. Anderson, I.A. and Wood, W. (2021) Habits and the electronic herd: The psychology behind social media’s successes and failures. *Consum. Psychol. Rev.* 4, 83–99
82. Ceylan, G. *et al.* (2023) Sharing of misinformation is habitual, not just lazy or biased. *Proc. Natl. Acad. Sci.* 120, e2216614120
83. Das, S. and Lavoie, A. (2014) The effects of feedback on human behavior in social media: an inverse reinforcement learning model. in *Proceedings of the 2014 international conference on Autonomous agents and multi-agent systems*, Richland, SC, pp. 653–660
84. Gran, A.-B. *et al.* (2021) To be or not to be algorithm aware: a question of a new digital divide? *Inf. Commun. Soc.* 24, 1779–1796
85. Klawitter, E. and Hargittai, E. (2018) “It’s Like Learning a Whole Other Language": The Role of Algorithmic Skills in the Curation of Creative Goods. *Int. J. Commun.* 12, 21
86. Smith, A. (2018) Many Facebook users don’t understand how the site’s news feed works. *Pew Research Center*.
87. Cotter, K. (2019) Playing the visibility game: How digital influencers and algorithms negotiate influence on Instagram. *New Media Soc.* 21, 895–913
88. Shepherd, R.P. (2020) Gaming Reddit’s Algorithm: r/the_donald, Amplification, and the Rhetoric of Sorting. *Comput. Compos.* 56, 102572
89. Rozado, D. *et al.* (2022) Longitudinal analysis of sentiment and emotion in news media headlines using automated labelling with Transformer language models. *PLOS ONE* 17, e0276367
90. Arugute, N. *et al.* (2023) Network activated frames: content sharing and perceived polarization in social media. *J. Commun.* 73, 14–24
91. McClain, C. (2021) 70% of U.S. social media users never or rarely post or share about political, social issues. *Pew Research Center*.
92. Lees, J.M. and Cikara, M. (2020) *Understanding and Combating Misperceived Polarization*, PsyArXiv
93. Levendusky, M.S. and Malhotra, N. (2016) (Mis)perceptions of Partisan

Polarization in the American Public. *Public Opin. Q.* 80, 378–391

94. Wilson, A.E. *et al.* (2020) Polarization in the contemporary political and media landscape. *Curr. Opin. Behav. Sci.* 34, 223–228
95. Bail, C.A. *et al.* (2018) Exposure to opposing views on social media can increase political polarization. *Proc. Natl. Acad. Sci.* 115, 9216–9221
96. Bakir, V. and McStay, A. (2018) Fake News and The Economy of Emotions. *Digit. Journal.* 6, 154–175
97. Vosoughi, S. *et al.* (2018) The spread of true and false news online. *Science* 359, 1146–1151
98. Martel, C. *et al.* (2020) Reliance on emotion promotes belief in fake news. *Cogn. Res. Princ. Implic.* 5, 47
99. Bago, B. *et al.* (2020) Fake news, fast and slow: Deliberation reduces belief in false (but not true) news headlines. *J. Exp. Psychol. Gen.* 149, 1608–1613
100. McLoughlin, K.L. *et al.* (2021) The role of moral outrage in the spread of misinformation. *Technol. Mind Behav.* DOI: 10.1037/tms0000136
101. Simchon, A. *et al.* (2022) Troll and divide: the language of online polarization. *PNAS Nexus* 1, pgac019
102. Chesney, B. and Citron, D. (2019) Deep Fakes: A Looming Challenge for Privacy, Democracy, and National Security. *Calif. Law Rev.* 107, 1753
103. Groh, M. *et al.* (2022) Deepfake detection by human crowds, machines, and machine-informed crowds. *Proc. Natl. Acad. Sci.* 119, e2110013119
104. Richerson, P.J. and Boyd, R. (2008) *Not By Genes Alone: How Culture Transformed Human Evolution*, University of Chicago Press
105. De, S. *et al.* (2018) Tipping Points for Norm Change in Human Cultures. in *Social, Cultural, and Behavioral Modeling*, Cham, pp. 61–69
106. Centola, D. *et al.* (2018) Experimental evidence for tipping points in social convention. *Science* 360, 1116–1119
107. Rogers, E. *et al.* (2008) Diffusion of Innovations. In *An Integrated Approach to Communication Theory and Research* ((2nd edn)), Routledge
108. Doroshenko, L. and Tu, F. (2022) Like, Share, Comment, and Repeat: Far-right Messages, Emotions, and Amplification in Social Media. *J. Inf. Technol. Polit.* 0, 1–17
109. Guess, A.M. *et al.* (2020) The sources and correlates of exposure to vaccine-related (mis)information online. *Vaccine* 38, 7799–7805
110. King, G. *et al.* (2017) How the Chinese Government Fabricates Social Media Posts for Strategic Distraction, Not Engaged Argument. *Am. Polit. Sci. Rev.* 111, 484–501
111. Brady, W.J. and Crockett, M.J. (in press) Norm Psychology in the Digital Age: How Social Media Shapes the Cultural Evolution of Normativity. *Brains Behav. Sci.*
112. Anderson, M. and Auxier, B. (2020) 55% of U.S. social media users say they are ‘worn out’ by political posts and discussions. *Pew Research Center*.
113. Rathje, S. *et al.* (2022) People think that social media platforms do (but should not) amplify divisive content. DOI: 10.31234/osf.io/gmun4
114. Hong, L. and Page, S.E. (2004) Groups of diverse problem solvers can outperform groups of high-ability problem solvers. *Proc. Natl. Acad. Sci.* 101, 16385–16389
115. Smaldino, P.E. *et al.* (2022) Maintaining transient diversity is a general principle for

improving collective problem solving. DOI: 10.31235/osf.io/ykrv5

116. Yaniv, I. (2011) Group diversity and decision quality: Amplification and attenuation of the framing effect. *Int. J. Forecast.* 27, 41–49
117. Brinkmann, L. *et al.* (2022) Hybrid social learning in human-algorithm cultural transmission. *Philos. Trans. R. Soc. Math. Phys. Eng. Sci.* 380, 20200426
118. Strittmatter, A. *et al.* (2020) Life cycle patterns of cognitive performance over the long run. *Proc. Natl. Acad. Sci.* 117, 27255–27261
119. Hardy, M.D. *et al.* (2022) Bias amplification in experimental social networks is reduced by resampling. DOI: 10.48550/arXiv.2208.07261
120. Törnberg, P. (2022) How digital media drive affective polarization through partisan sorting. *Proc. Natl. Acad. Sci.* 119, e2207159119
121. Kozyreva, A. *et al.* (2020) Citizens Versus the Internet: Confronting Digital Challenges With Cognitive Tools. *Psychol. Sci. Public Interest* 21, 103–156
122. Nussberger, A.-M. *et al.* (2022) Public attitudes value interpretability but prioritize accuracy in Artificial Intelligence. *Nat. Commun.* 13, 5821
123. Striphas, T. (2015) Algorithmic culture. *Eur. J. Cult. Stud.* 18, 395–412
124. Alvarado, O. and Waern, A. (2018) Towards Algorithmic Experience: Initial Efforts for Social Media Contexts. in *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, New York, NY, USA, pp. 1–12
125. Metzler, H. and Garcia, D. (2022) Social drivers and algorithmic mechanisms on digital media. DOI: 10.31234/osf.io/cxa9u
126. Eckles, D. *et al.* (2018) Field studies of psychologically targeted ads face threats to internal validity. *Proc. Natl. Acad. Sci.* 115, E5254–E5255
127. Schug, J. *et al.* (2010) Relational Mobility Explains Between- and Within-Culture Differences in Self-Disclosure to Close Friends. *Psychol. Sci.* 21, 1471–1478
128. Yuki, M. and Schug, J. (2020) Psychological consequences of relational mobility. *Curr. Opin. Psychol.* 32, 129–132
129. Chen, H. *et al.* (2014) Wisdom of Crowds: The Value of Stock Opinions Transmitted Through Social Media. *Rev. Financ. Stud.* 27, 1367–1403
130. Hernández, A.D. *et al.* (2022) Addressing contingency in algorithmic misinformation detection: Toward a responsible innovation agenda. DOI: 10.48550/arXiv.2210.09014
131. Hsu, T.W. *et al.* (2021) Social media users produce more affect that supports cultural values, but are more influenced by affect that violates cultural values. *J. Pers. Soc. Psychol.* 121, 969–983
132. Berger, J. and Milkman, K.L. (2012) What Makes Online Content Viral? *J. Mark. Res.* 49, 192–205
133. Morin, O. *et al.* (2021) Social information use and social information waste. *Philos. Trans. R. Soc. B Biol. Sci.* 376, 20200052