New England Journal of Public Policy

Volume 36
Issue 1 *The Changing Character of War and Peacemaking*

Article 13

6-16-2024

Scaling Expertise: A Note on Homophily in Online Discourse and Content Moderation

Dylan Weber
Changing Character of War Centre, Pembroke College, University of Oxford; Artis International

Scaling Expertise: A Note on Homophily in Online Discourse and Content Moderation

Dylan Weber
Changing Character of War Centre, Pembroke College, University of Oxford; Artis International

Abstract

It is now empirically clear that the structure of online discourse tends toward homophily; users strongly prefer to interact with content and other users that are similar to them. I review the evidence for the ubiquity of homophily in discourse and highlight some of its worst effects including narrowed information landscape for users and increased spread of misinformation. I then discuss the current state of moderation frameworks at large social media platforms and how they are ill-equipped to deal with structural trends in discourse such as homophily. Finally, I sketch a moderation framework based on a principal of "scaling expertise" that I believe can contend with the scale of online discourse while maintaining sensitivity to context and culture.

Dr. Dylan Weber is a fellow of the Changing Character of War Centre at Pembroke College and of Artis International where he additionally serves as the director of artificial intelligence and deputy chief technology officer. He is a mathematician whose interests lie mainly in the dynamics of online discourse. This intersects several areas including statistical physics, modeling of self-organized behavior, and data science, and artificial intelligence.

The internet was and continues to be lauded by many for its role in improving discourse and democratizing information by allowing for more direct discussion among the populace and by greatly enhancing access to information generally.[1] However, at its advent and throughout its history, some have cautioned that due to well-established psychological tendencies (such as confirmation bias and selective exposure), discourse on the internet would actually result in a narrowing of the information landscape of its users. This view predicted that, given the freedom to select information from the virtually infinite array that the internet provides, users would spend their time on the internet interacting with those who they already agreed with and consuming information from sources that confirmed their prior views. In other words, this view predicted that the structure of online discourse would tend toward "echo chambers" or *homophily*.[2]

The 2016 United States presidential election prompted many empirical studies that leverage large social media data sets to look into the extent to which homophily presents on social media and what effects it might have on its users. Despite the large variety of methods across this body of work, it is largely unified in its findings; social media has a strongly homophilic structure (see, for example, Figure 1) and this structure contributes to multiple undesirable effects including the narrowing of user information landscapes, enhanced spread of misinformation in homophilic clusters, and increased ideological polarization.[3]

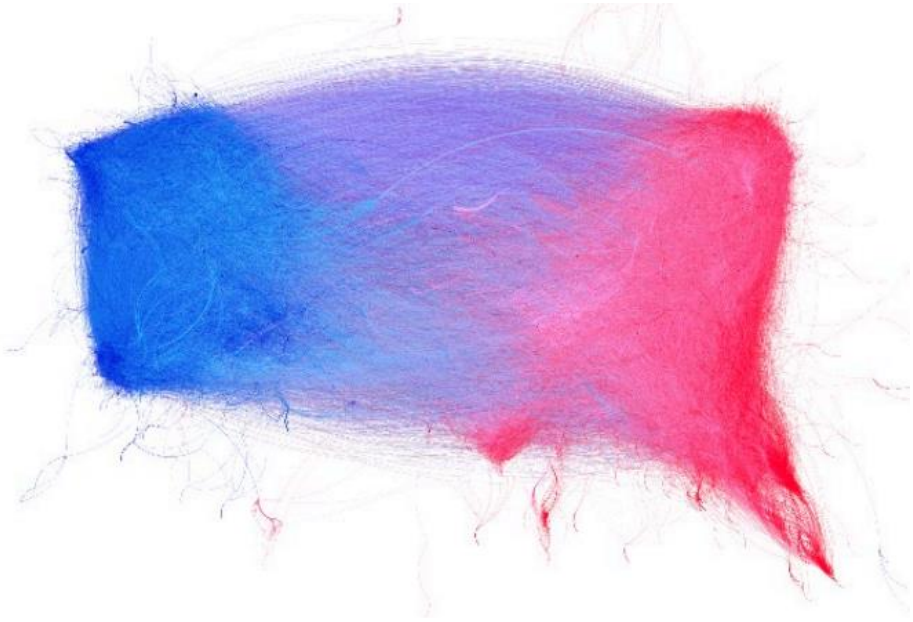

Figure 1. Visualization of the Twitter conversation from May 2023 surrounding the Black Lives Matter movement. Nodes in the network are users, edges represent aggregated interactions between users over the data collection period. Nodes are colored according to their support for or opposition to the Black Lives Matter movement: red indicates support, blue indicates opposition. Edges are directed and are colored according to the color of the source node. The network illustrates a strongly homophilic structure.

The ubiquity of homophily in online discourse and its ill effects has been largely acknowledged (usually under different names) by academia and policy makers, and to a lesser extent by social media platforms themselves. This has prompted a meta-discussion concerning mitigation of these undesirable effects through different strategies to mediate online discourse. All of the methods that have been actually implemented have been devised by social media platforms, resulting in an explosion of controversy surrounding the ethics of mediating speech in online discourse to the point that this has become a political issue itself.[4] Additionally, the implementation of moderation strategies by social media platforms has resulted in data that has allowed for an early foray into empirical measurement of the implemented strategies' effectiveness. Ethical and political discussions aside, I argue that none of these methods have been especially effective, as evidenced by the continued existence of strongly homophilic structures in online discourse and the balance of the research into implemented moderation strategies.

In the first section of the article I review the evidence for the strong tendency toward homophily in online discourse and discuss possible mechanisms for that tendency and its myriad ill effects on discourse. In the second section we discuss the current practice of content moderation on large social media platforms and some of its challenges. In the last section, I present a moderation framework that centers transparency, data, and scaled expertise.

Homophily in Online Discourse

Since the early days of social media, and indeed, the internet, there has been a vocal minority in academia warning that the disintermediated nature of online discourse could result in the formation of "echo chambers."[5] However, until the 2016 United States presidential election, there was a sparse amount of empirical investigation into whether such communities were actually emerging. In the years since, there has been an explosion of research into this question. Given the variation in methods and in the definition of "echo chamber" across these studies, we will first offer some definitions to put all the findings on common ground.

Definition 1. A network is a collection of nodes and a collection of edges that connect the nodes.

Definition 2. Given a network and a quantity defined on its nodes that is a priori independent of the network edge structure, we say that the network is homophilic with respect to the quantity if nodes are more likely to be connected in the network if they have a similar value of the quantity.

In short, a network is homophilic if "users of a feather connect together." It is important to consider that given a network, it may be homophilic with respect to one quantity and not another. For example, if we consider a network of users engaging in a debate about two political candidates we would expect that network to be homophilic with respect to support for the candidates but not with respect to support for various sports teams or food preferences. All online discourse can be viewed through the lens of network structure where users are represented by nodes of the network and edges represent their interactions online. All research about the existence of echo chambers in online discourse can be viewed as an investigation into homophily within the user interaction networks with respect to some measure of the users' ideology. From this point of view, there is a meta-methodology employed by researchers to investigate homophily within online discourse (though there is much variation in specific methodologies to accomplish each step). First, researchers collect a large social media data set on an issue or collection of issues. Next, they define a methodology for quantifying a user's ideology on the issue in question. Then, a scheme is defined for structuring the data set into a network and a metric for quantifying homophily on the

resulting network with respect to the defined ideology measure. This methodology has produced a stark answer: the structure of social media tends strongly toward homophily.

Early forays into the use of this methodology identified homophilic structures in conversations surrounding the 2012 presidential election and other political issues in the United States including the 2013 government shutdown, minimum wage, and marriage equality.[6] The presence of strong homophily within networks pertaining to specific political conversations was confirmed by Kiran Garimella et al. through an investigation of the conversations surrounding gun control, Obamacare, and abortion, and by Bjarke Mønsted and Ana Lucía Schmidt through an examination of the Twitter and Facebook conversations surrounding vaccines.[7] Several years later, Matteo Cinelli et al. looked into the abortion, gun control, and vaccine conversations using more sophisticated metrics for ideology and homophily and again found that strong homophily presented on Facebook and Twitter. Through leveraging their more sophisticated ideology measure and examining the same conversations on Reddit and Gab, they were also able to show that entire social platforms can be echo chambers with respect to certain issues.[8] However, homophily is not limited to conversations surrounding a single topic. Robert Bond and Solomon Messing, and Eytan Bakshy et al. conducted investigations of the US political conversation on Facebook writ large.[9] All three studies identify pronounced homophily using various ideology metrics. Nor is homophily restricted to political conversation; Walter Quattrociocchi et al., Alessandro Bessi et al., and Fabiana Zollo et al. examined Facebook discussion driven by either science backed information or conspiracy backed information and find homophily within the Facebook friendship network.[10] Those who consumed information from conspiracy groups were very unlikely to be friends with those who consumed information from groups centered on science and vice versa. A similar pattern was found in news consumption on Facebook by Schmidt; users tend to only interact with a small group of similarly aligned outlets in lieu of all other news sources. Users who consume the same group of news sources are much more likely to interact with each other.[11]

Though there is strong consensus on the existence of strong homophily within online discourse, the work on mechanisms that drive this effect is in early stages. Many originally pointed to the algorithmic curation of content as a likely primary driver. Proponents of this theory pointed out that algorithmic curation of content could place users in a "filter bubble" where they are mainly exposed to content for which they have previously demonstrated preference.[12] However it seems that this effect is less pronounced than feared and that homophily is more likely driven by a natural tendency to self-select aligned content and place oneself in an "engagement echo chamber."[13] Indeed in a study of political cross-cutting content, which is content that disagrees with the political ideology of the user viewing it, Bakshy et al. found that the filter bubble effect was small; "there is on average slightly less cross-cutting content: conservatives see approximately 5% less cross-cutting content compared to what friends share, while liberals see about 8% less ideologically diverse content." However, they also confirmed that individuals engage with cross-cutting content at much lower rates than ideologically aligned content.[14] Indeed, despite the evidence that the filter bubble effect is small, there is much evidence that the dynamic of selective exposure is large and is a primary driver of the observed tendency toward homophily. For example, Schmidt et al. find that Facebook users tend to initially engage with a large amount of pages but quickly converge to spending most of their time on a small collection.[15] Additionally, the most active users focused their attention on the fewest number of pages, a finding confirmed by Cinelli et al.[16]

These studies by Schmidt and Cinelli et al. demonstrate a worrying effect of homophily in online discourse; the information landscape of the average social media user is greatly diminished. This effect is also confirmed by Dimitar Nikolav et al.[17] More worrying, homophilic clusters are

very susceptible to the spread of misinformation. Michela Del Vicario et al. find that a certain degree of homophily within the user interaction network is necessary for a viral cascade to occur and that the more homophilic a network is, the larger the size of the cascades it can facilitate. Additionally, they find that content must be sufficiently aligned with the ideology of a homophilic component in order to initiate a viral cascade.[18] Indeed, Aris Anagnostopoulos et al. find that the main factor in determining virality is "users' aggregation around shared beliefs."[19] This suggests that content could easily spread within a homophilic cluster if it is aligned with the ideology of the cluster regardless of whether the content is factual or not. Because misinformation could be crafted to align with the ideology of a cluster, one could surmise that misinformation actually spreads more effectively than the truth within homophilic network structures. Indeed, a study by Soroush Vosoughi et al. examined the structure of viral cascades that originate from content that has been verified as either true or false and found that false information spreads much more effectively than the truth:

> Whereas the truth rarely diffused to more than 1000 people, the top 1% of false-news cascades routinely diffused to between 1000 and 100,000 people. Falsehood reached more people at every depth of a cascade than the truth, meaning that many more people retweeted falsehood than they did the truth. The spread of falsehood was aided by its virality, meaning that falsehood did not simply spread through broadcast dynamics but rather through peer-to-peer diffusion characterized by a viral branching process. It took the truth about six times as long as falsehood to reach 1500 people and 20 times as long as falsehood to reach a cascade depth of 10. As the truth never diffused beyond a depth of 10, we saw that falsehood reached a depth of 19 nearly 10 times faster than the truth reached a depth of 10. Falsehood also diffused significantly more broadly and was retweeted by more unique users than the truth at every cascade depth.[20]

Current Moderation Strategies

The ill effects of homophily in online discourse are not just well-established in the academy. They have become central to a larger meta-discussion concerning the role of social media platforms in discourse and what the platforms' role should be in moderating discourse within their spheres. There is a general acknowledgment that the fully open and disintermediated discourse provided by social media platforms has had a host of unintended consequences including the emergence of strongly homophilic discourse. This discussion has led to a large and public slew of moderation efforts from the platforms, most visibly Facebook. Indeed, as Evelyn Douek notes:

> Facebook's update to the "values" that inform its Community Standards is perhaps the starkest example of the dominance of this new paradigm. Where once Facebook emphasized connecting people it now acknowledges that voice should be limited for reasons of authenticity, safety, privacy, and dignity.[21]

In my view, these interventions have not been successful, mainly evidenced by the continued existence of strong homophilic clusters and the spread of harmful discourse within them. In this section I discuss the current practice of content moderation on large social media platforms and some of its challenges.

We define *content moderation* generally as efforts to reduce the proliferation of content considered to be harmful, mainly through alteration of the content itself, limitation of its visibility, or ultimately its removal from the platform in question. A general definition is necessary because there is no wide consensus on what constitutes harmful content and the practice of moderating its spread and impact is currently solely in the hands of the social media platforms themselves (though Section 230 of the US Communications Act of 1934, enacted as part of the Communications Decency Act of 1996, guarantees that they have no legal obligation to do so).[22] The modern practice of content moderation is also in its infancy; most major platforms have always had rules forbidding certain classes of exceptional content (e.g., illegal activity, copyrighted material, or extremist content) but most consider its modern practice as beginning in earnest in 2016 when Facebook began working with third party fact checking organizations in response to the Cambridge Analytica scandal. By 2018, Facebook had more than 20,000 people working in content moderation.[23]

At "industrial" scale platforms (e.g., Facebook, Twitter, and YouTube) the general approach to content moderation follows the same model. A public facing set of "standards" or "guidelines" is published; the public goal of content moderation is to ostensibly ensure that all content on the platform is within these standards. However, the language in these standards is general and vague; it is not always possible to directly infer from them when a particular piece of content is in violation. The rules that are actually enforced in practice are written by a policy team that works under the umbrella of the larger apparatus responsible for content moderation. These rules are often opaque to the public. In the words of Robyn Caplan, "most platform companies keep their content moderation policies partially, if not mostly, hidden."[24] These policies are then operationalized into the actual practices of the content moderation teams. Content moderation teams are tasked with making the ultimate moderation decision on pieces of content served to them by both artificial intelligence (AI) detection capabilities and user reports. By all accounts the goal of this operationalization is to disambiguate complex concepts such as hate speech and disinformation into a discrete set of practices that could be applied consistently by moderation teams in order to deal with the huge scale of the task. In the words of one Facebook employee anonymously interviewed by Caplan, the goal is to create a "decision factory."[25] The end result is that it seems that much of content moderation is reduced to a mechanical process that is not considerate of the larger context from which content originates.[26] This can have disastrous outcomes, evidenced clearly, for example, by widespread political violence in Myanmar in 2018, which Facebook admits was driven by discourse on its platform.[27] Mark Zuckerberg further commented that its moderation teams lacked the proper cultural context to effectively identify the fomenting content.[28] In this sense, the lack of transparency into content moderation policies is particularly problematic because it is impossible for society to evaluate whether platforms are leveraging the best state of empirical knowledge to inform the goals of content moderation, the crafting of the related policies, and ultimately the operationalization of these policies into practice for the moderation teams.

There is a young but increasingly robust literature that generally proposes content moderation strategies based on foundational psychological literature and then measures their effectiveness in labs and survey settings. For example, in the arena of approaches for debunking misinformation, there is clear enough consensus on empirically-motivated best practices to allow for a distillation of the literature by Stephan Lewandowsky et al. into a *Debunking Handbook*.[29] (See "The emerging science of content labeling: Contextualizing social media content moderation" by Garrett Morrow et al. for a good review of similar work.[30]) One would hope that social media platforms are

leveraging this type of information to inform their content moderation policies, but in the current state of play, that is impossible to know. A main takeaway from this body of literature is that there are replicable, significant effects of content moderation that have been identified in lab settings but it is unclear how these effects will scale on the platforms. As an example, Gordon Pennycook et al. showed that Facebook's work with third party fact checking organizations to assign warning labels to news stories identified as false resulted in stories lacking a label to be perceived as more likely to be true, regardless of whether the unlabeled stories had been fact checked or not.[31] To investigate the effects that "local" phenomena like this so-called "implied truth effect" have on global features of the structure of discourse like homophily would require an investigation that leverages the full scale of platform data and necessarily more transparency from the platforms.

Indeed, scale seems to be the most problematic aspect of content moderation. In the words of the same anonymously interviewed Facebook employee, even in Facebook's earlier days when the user base was only around seventy million users, "you would have to hire everybody in India to look at all the content that was uploaded, and you still wouldn't be able to do it."[32] Even if content moderation policies and their operationalizations into the practices of content moderators are transparent and informed by empirical best practice this only serves to improve the moderation of content when the final moderation decision falls into the hands of a human moderator. The massive speed and scale of information on large social media platforms makes this completely intractable. Automation must be brought to bear on the problem. Facebook has acknowledged for years that AI systems play a major role in flagging content for review by its moderation teams and as early as 2018 acknowledged that these systems participate in the automatic removal of content in the previously mentioned exceptional classes.[33] The extent to which AI systems participate in automatic removal versus referral to the moderation queue remains unclear. In 2020 Facebook acknowledged that in response to pressures on its moderation workforce caused by the COVID-19 pandemic it would increasingly rely on AI to make content moderation decisions.[34] It seems that the work force was never fully replenished—users continue to receive messages that their reports will not be processed by content moderators due to a "high volume of reports."[35] Generally it seems that platforms originally used AI for automatic removal of unambiguously objectionable content while relegating more context dependent moderation decisions to humans and that this balance has increasingly shifted toward the use of AI systems as time has gone on. Recently, Facebook announced that it would investigate the use of large language models (LLMs) to allow for more use of context in automated moderation decisions.[36]

Regardless of where the current balance between human moderation and automated moderation stands, it has long been stated that the actions of human moderators are used as training data to improve the performance of the AI capabilities.[37] While on the surface this would seem to be in line with AI alignment best practices, it is concerning, given the current practice of content moderation, for several reasons. Fundamentally, the goal of such a system is to scale the decision-making power of the human moderators to the models. However, as noted, the goals of moderation policies are opaque outside of published community guidelines, the policies themselves similarly lack transparency, and the evidence we do have about how they are operationalized points to a system that prioritizes scalability of human decision-making over the nuances of context that content moderation requires. Additionally, human moderators are largely not Facebook employees and do not have any expertise in conflict reduction or psychology. There have been many whistleblowers' reports to the media and governments about working conditions and adverse mental effects from doing the job.[38] Said shortly, it seems that the decision-making process that is being scaled to automated AI capabilities, and used to intermediate the largest discourse platforms

ever created, is based on objectives and policy that society is ignorant of, is operationalized in a manner that prioritizes scalable process over contextual and cultural understanding, and is ultimately carried out by a workforce that has no expertise, is overworked, and has no hope of confronting the scale of the task. This does not bode well for the function of these systems at scale.

Scaling Expertise

So what is to be done? It is clear from the prevalence of homophily in the structure of online discourse, the frequency of large misinformation spreading events within homophilic clusters, and high levels of affective polarization between homophilic clusters, that current moderation policies used by large social media platforms are not effective at mitigating the worst effects of homophily. It is also clear from the sheer scale of discourse on these platforms that any attempt at effective moderation of this tendency must leverage automation to scale moderation strategies, likely in the form of AI-powered capabilities. A good initial step would be for platforms to make transparent the goals of content moderation and the specific moderation policies that attempt to accomplish these goals. This transparency would allow independent experts and researchers to have a voice in what best practices should look like and to publish suggestions as to how the policy should evolve and be tested by the platforms.

More importantly, the problem of access to platform data needs to be addressed. Currently, virtually all of the data needed to effectively measure online discourse, and necessarily the effectiveness of moderation strategies is wholly controlled by the platforms themselves and disseminated to independent researchers at their discretion and at great cost. Additionally, the methods by which the data is sampled for clients who purchase it are opaque. For example, Facebook recently announced that it would close CrowdTangle, its research data application programming interface (API), in August of 2024 and announced plans for a replacement but, at the time of writing, has provided little documentation as to the reasons for the change or what changes in functionality will occur. As previously noted, research into effective forms of content moderation is at the stage where there are plenty of theoretically motivated ideas, some of which have been tested in lab settings and have significant and replicable effects. Access to platform data would allow these strategies to be investigated at their intended scale. Moderation approaches should be evaluated on their effects on global platform phenomena such as homophily and viral misinformation spread. Greater access to platform data would also allow for more investigation into methods of measuring discourse and the dynamics of platform structure and content. Better understanding of the dynamics of global platform structure could be used to better inform the specific goals of moderation policies.

Finally, given a content policy that is grounded in empirical best practice, the job of content moderators should be viewed by the platforms as a role that requires expertise. Instead of relying on third party contractors, platforms should be thoughtful about crafting their moderation workforce and hire for skills relevant for the task such as conflict resolution and media literacy. Steps should be taken to ensure that moderators have a good cultural understanding of the content they interact with, ideally by hiring from a variety of geographies, and moderators should be deeply trained in the theory that grounds moderation policies. Content should be served to moderators with as much of its originating context as possible so that these skills can be effectively leveraged.

A content moderation framework that can confront structural problems in online discourse such as homophily should focus on a principle of “scaling expertise.” This requires two main elements. First it requires a moderation policy with goals and techniques informed by empirical best practice and a measurement capability that can capture the global structure of the platform

and its dynamics to assess the effectiveness of moderation and inform its further iteration. Second it requires a workforce of human moderators that have deep expertise in the moderation policy and have a cultural understanding of the content they are moderating. With these elements in place it is feasible to create a virtuous cycle for the creation of an automated moderation capability that can scale and augment the decision-making of the human moderators so that they can confront the massive scale of online discourse. Careful measurement of the online information environment (including current moderation activity) and testing of moderation strategies is translated into thoughtful and empirically informed moderation policies. These policies are deeply understood by moderators with cultural knowledge whose decisions are scaled to an automated capability which is then, of course, measured for effectiveness relative to the goals of the moderation policy and so on. Many of the problems with online discourse can be understood as a scaling of our own worst tendencies. Likewise, solutions to these problems demand similar scale but with greater consideration. In the words of Lewandowsky, "if technology can facilitate such epistemic fractionation in the first place, then it stands to reason that it might also contribute to the solution."[39]

Notes

[1] Howard Rheingold, *The Virtual Community, Revised Edition: Homesteading on the Electronic Frontier* (Cambridge, MA: MIT Press, 2000); Robert H. Anderson et al., "Universal Access to Email: Feasibility and Societal Implications," in *The Digital Divide: Facing a Crisis or Creating a Myth?*, ed. Benjamin M. Compaine (Cambridge, MA: MIT Press, 2001), 243–62; James S. Fishkin, "Virtual Democratic Possibilities: Prospects for Internet Democracy" (paper presented at Internet, Democracy, and Public Goods, Belo Horizonte, Brazil, November 2000); Vincent Price and Joseph N. Cappella, "Online Deliberation and Its Influence: The Electronic Dialogue Project in Campaign 2000," *IT & Society* 1, no. 1 (2002): 303–29.

[2] Jodi Dean, "Why the Net Is Not a Public Sphere," *Constellations* 10, no. 1 (2003): 95–112; Lincoln Dahlberg, "Cyberspace and the Public Sphere: Exploring the Democratic Potential of the Net," *Convergence* 4, no. 1 (1998): 70–84; Hubertus Buchstein, "Bytes That Bite: The Internet and Deliberative Democracy," *Constellations* 4, no. 2 (1997): 248–63; Eli Pariser, *The Filter Bubble: How the New Personalized Web Is Changing What We Read and How We Think* (New York: Penguin, 2011); John Kelly, Danyel Fisher, and Marc Smith, "Debate, Division, and Diversity: Political Discourse Networks in USENET Newsgroups" (paper presented at the Online Deliberation Conference 2005, Stanford University, May 2005), 3; Cass R. Sunstein, "Democracy and Filtering," *Communications of the ACM* 47, no. 12 (2004): 57–59; Cass R. Sunstein, *Republic.Com* (Princeton, NJ: Princeton University Press, 2001).

[3] Kiran Garimella et al., "The Effect of Collective Attention on Controversial Debates on Social Media," in *Proceedings of the 2017 ACM on Web Science Conference* (Troy, NY: ACM, 2017), 43–52, https://doi.org/10.1145/3091478.3091486; Matteo Cinelli et al., "The Echo Chamber Effect on Social Media," *Proceedings of the National Academy of Sciences* 118, no. 9 (March 2021), https://doi.org/10.1073/pnas.2023301118; Bjarke Mønsted and Sune Lehmann, "Characterizing Polarization in Online Vaccine Discourse—A Large-Scale Study," *PLOS ONE* 17, no. 2 (February 2022): e0263746, https://doi.org/10.1371/journal.pone.0263746; Ana Lucía Schmidt et al., "Polarization of the Vaccination Debate on Facebook," *Vaccine* 36, no. 25 (June 2018): 3606–12, https://doi.org/10.1016/j.vaccine.2018.05.040; Matteo Cinelli et al., "Selective Exposure Shapes the Facebook News Diet," *PLOS ONE* 15, no. 3 (March 2020): e0229129, https://doi.org/10.1371/journal.pone.0229129; Michela Del Vicario et al., "The Spreading of Misinformation Online," *Proceedings of the National Academy of Sciences* 113, no. 3 (January 2016): 554–59, https://doi.org/10.1073/pnas.1517441113; Soroush Vosoughi, Deb Roy, and Sinan Aral, "The Spread of True and False News Online," *Science* 359, no. 6380 (March 2018): 1146–51, https://doi.org/10.1126/science.aap9559.

[4] Meysam Alizadeh et al., "Content Moderation as a Political Issue: The Twitter Discourse around Trump's Ban," *Journal of Quantitative Description: Digital Media* 2 (October 2022), https://doi.org/10.51685/jqd.2022.023; Samuel Mayworm et al., "Content Moderation Folk Theories and Perceptions of Platform Spirit among Marginalized Social Media Users," *ACM Transactions on Social Computing* 7, no. 1 (March 2024): 1:1–1:27,

https://doi.org/10.1145/3632741; Ángel Díaz and Laura Hecht-Felella, "Double Standards in Social Media Content Moderation," Brennan Center for Justice, 2021, https://www.brennancenter.org/sites/default/files/2021-08/Double_Standards_Content_Moderation.pdf.

[5] Kelly, Fisher, and Smith, "Debate, Division, and Diversity"; Sunstein, "Democracy and Filtering"; Sunstein, *Republic.Com*.

[6] Pablo Barberá, "Birds of the Same Feather Tweet Together: Bayesian Ideal Point Estimation Using Twitter Data," *Political Analysis* 23, no. 1 (2015): 76–91, https://doi.org/10.1093/pan/mpu011; Pablo Barberá et al., "Tweeting From Left to Right: Is Online Political Communication More Than an Echo Chamber?," *Psychological Science* 26, no. 10 (October 2015): 1531–42, https://doi.org/10.1177/0956797615594620.

[7] Garimella et al., "The Effect of Collective Attention"; Kiran Garimella et al., "Political Discourse on Social Media: Echo Chambers, Gatekeepers, and the Price of Bipartisanship," in *Proceedings of the 2018 World Wide Web Conference, WWW '18* (Geneva: International World Wide Web Conferences Steering Committee, April 2018), 913–22, https://doi.org/10.1145/3178876.3186139; Mønsted and Lehmann, "Characterizing Polarization in Online Vaccine Discourse"; Schmidt et al., "Polarization of the Vaccination Debate on Facebook."

[8] Cinelli et al., "The Echo Chamber Effect on Social Media."

[9] Robert Bond and Solomon Messing, "Quantifying Social Media's Political Space: Estimating Ideology from Publicly Revealed Preferences on Facebook," *American Political Science Review* 109, no. 1 (2015): 62–78, https://www.doi.org/10.1017/S0003055414000525.

[10] Alessandro Bessi et al., "Science vs Conspiracy: Collective Narratives in the Age of Misinformation," *PLOS ONE* 10, no. 2 (2015): e0118093, https://doi.org/10.1371/journal.pone.0118093; Fabiana Zollo et al., "Debunking in a World of Tribes," *PLOS ONE* 12, no. 7 (2017): e0181821, https://doi.org/10.1371/journal.pone.0181821.

[11] Ana Lucía Schmidt et al., "Anatomy of News Consumption on Facebook," *Proceedings of the National Academy of Sciences* 114, no. 12 (March 2017): 3035–39, https://doi.org/10.1073/pnas.1617052114; Schmidt et al., "Polarization of the Vaccination Debate on Facebook."

[12] Dominic Spohr, "Fake News and Ideological Polarization: Filter Bubbles and Selective Exposure on Social Media," *Business Information Review* 34, no. 3 (September 2017): 150–60, https://doi.org/10.1177/0266382117722446; Pariser, *The Filter Bubble*.

[13] R. Kelly Garrett, "The 'Echo Chamber' Distraction: Disinformation Campaigns Are the Problem, Not Audience Fragmentation," *Journal of Applied Research in Memory and Cognition* 6, no. 4 (December 2017): 370–76, https://doi.org/10.1016/j.jarmac.2017.09.011.

[14] Eytan Bakshy, Solomon Messing, and Lada A. Adamic, "Exposure to Ideologically Diverse News and Opinion on Facebook," *Science* 348, no. 6239 (June 2015): 1130–32, https://doi.org/10.1126/science.aaa1160.

[15] Schmidt et al., "Anatomy of News Consumption on Facebook"; Schmidt et al., "Polarization of the Vaccination Debate on Facebook."

[16] Cinelli et al., "Selective Exposure Shapes the Facebook News Diet."

[17] Dimitar Nikolov et al., "Measuring Online Social Bubbles," *PeerJ Computer Science* 1 (December 2015): e38, https://doi.org/10.7717/peerj-cs.38.

[18] Del Vicario et al., "The Spreading of Misinformation Online."

[19] Aris Anagnostopoulos et al., "Viral Misinformation: The Role of Homophily and Polarization," preprint, submitted November 2014, https://doi.org/10.48550/arXiv.1411.2893.

[20] Vosoughi, Roy, and Aral, "The Spread of True and False News Online."

[21] Evelyn Douek, "Governing Online Speech: From 'Posts-As-Trumps' to Proportionality and Probability," *Columbia Law Review* 121, no. 3 (2021), https://doi.org/10.2139/ssrn.3679607.

[22] Tanner Mirrlees, "GAFAM and Hate Content Moderation: Deplatforming and Deleting the Alt-Right," in *Media and Law: Between Free Speech and Censorship*, ed. Mathieu Deflem and Derek M. D. Silva, Sociology of Crime, Law and Deviance, vol. 26 (Leeds, UK: Emerald Publishing Limited, 2021), 81–97, https://doi.org/10.1108/S1521-613620210000026006.

[23] Robyn Caplan, "Content or Context Moderation? Artisanal, Community-Reliant, and Industrial Approaches," Data & Society Research Institute, November 2018, https://apo.org.au/node/203666.

[24] Caplan.

[25] Caplan.

[26] Tomas Apodaca and Natasha Uzcátegui-Liggett, "How Automated Content Moderation Works (Even When It Doesn't)," *The Markup*, March 1, 2024, https://themarkup.org/automated-censorship/2024/03/01/how-automated-content-moderation-works-even-when-it-doesnt-work.

[27] Alexandra Stevenson, "Facebook Admits It Was Used to Incite Violence in Myanmar," *New York Times*, November 6, 2018, https://www.nytimes.com/2018/11/06/technology/myanmar-facebook.html.

[28] Steve Stecklow, "Why Facebook Is Losing the War on Hate Speech in Myanmar," *The Wire*, August 16, 2018, https://thewire.in/tech/why-facebook-is-losing-the-war-on-hate-speech-in-myanmar.
[29] Stephan Lewandowsky, John Cook, and Doug Lombardi, *Debunking Handbook 2020*, 2020, https://doi.org/10.17910/B7.1182.
[30] Garrett Morrow et al., "The Emerging Science of Content Labeling: Contextualizing Social Media Content Moderation," *Journal of the Association for Information Science and Technology* 73, no. 10 (2022): 1365–86, https://doi.org/10.1002/asi.24637.
[31] Gordon Pennycook et al., "The Implied Truth Effect: Attaching Warnings to a Subset of Fake News Headlines Increases Perceived Accuracy of Headlines Without Warnings," *Management Science* 66, no. 11 (November 2020): 4944–57, https://doi.org/10.1287/mnsc.2019.3478.
[32] Caplan.
[33] "How Does Facebook Use Artificial Intelligence to Moderate Content?," Facebook Help Center, accessed May 20, 2024, https://www.facebook.com/help/1584908458516247; Caplan, "Content or Context Moderation?"
[34] Kang-Xing Jin, "Keeping People Safe and Informed about the Coronavirus," Meta Newsroom, December 18, 2020, https://about.fb.com/news/2020/12/coronavirus/.
[35] Ina Fried, "Facebook's Content-Review Black Hole: The Flagged Posts That Never Get Read by a Human," Axios, June 22, 2023, https://www.axios.com/2023/06/22/facebook-content-moderation-black-hole-human-review.
[36] Nick Clegg, "Labeling AI-Generated Images on Facebook, Instagram and Threads," Meta Newsroom, February 6, 2024, https://about.fb.com/news/2024/02/labeling-ai-generated-images-on-facebook-instagram-and-threads/.
[37] "How Does Facebook Use Artificial Intelligence to Moderate Content?"
[38] Cristina Criddle, "Facebook Moderator: 'Every Day Was a Nightmare,'" BBC, May 12, 2021, https://www.bbc.com/news/technology-57088382; David Pilling and Madhumita Murgia, "'You Can't Unsee It': The Content Moderators Taking on Facebook," *Financial Times*, May 2023, https://www.ft.com/content/afeb56f2-9ba5-4103-890d-91291aea4caa; Casey Newton, "The Trauma Floor: The Secret Lives of Facebook Moderators in America," *The Verge*, February 25, 2019, https://www.theverge.com/2019/2/25/18229714/cognizant-facebook-content-moderator-interviews-trauma-working-conditions-arizona.
[39] Stephan Lewandowsky, Ullrich K. H. Ecker, and John Cook, "Beyond Misinformation: Understanding and Coping with the 'Post-Truth' Era," *Journal of Applied Research in Memory and Cognition* 6, no. 4 (December 2017): 353–69, https://doi.org/10.1016/j.jarmac.2017.07.008.