TECHNOLOGY PRIMERS FOR POLICYMAKERS

Social Media Recommendation Algorithms

HARVARD Kennedy School
BELFER CENTER
for Science and International Affairs
TECHNOLOGY AND PUBLIC PURPOSE PROJECT

ASH CARTER, TAPP FACULTY DIRECTOR

AMRITHA JAYANTI, TAPP ASSOCIATE DIRECTOR

August 2022

AUTHOR

Constanza M. Vidal Bustamante (Harvard)

REVIEWERS

Joaquin Quiñonero-Candela (LinkedIn)

Lucas Wright (Cornell)

Leisel Bogan (Harvard)

Marc Faddoul (Tracking Exposed)

EDITORS

Ariel Higuchi (Harvard)

Amritha Jayanti (Harvard)

The Technology Primer Series was designed to provide a brief overview of each technology and related policy considerations. These papers are not meant to be exhaustive.

Technology and Public Purpose Project

Belfer Center for Science and International Affairs
Harvard Kennedy School
79 John F. Kennedy Street, Cambridge, MA 02138

www.belfercenter.org/TAPP

The Technology and Public Purpose (TAPP) Project would like to thank Zachary Lemisch who provided preliminary research assistance in creating this primer. Additionally, the TAPP Project would like to thank Sofía Jarrín for constructive copy editing of the primer.

Printed in the United States of America

Contents

Executive Summary

The use of social media platforms like Facebook, Twitter, YouTube, and TikTok is increasingly widespread, currently amounting to billions of users worldwide. In 2021, a majority of adults in the U.S. reported using at least one social media site, with the most popular ones being YouTube (81 percent) and Facebook (69 percent), and TikTok being especially common among 18–24-year-olds (55 percent).

Billions of pieces of content are published daily by users including individuals, public figures, interest groups, and organizations. Content ranges from personal updates to entertainment, tutorials, and news stories. In 2021, about one in two Americans reported getting their news from social media at least sometimes.

Social media companies deploy **proprietary recommendation algorithms to automate the selection, ranking, and presentation of content** on the platform's "feed" or recommended content section, every time a user opens or refreshes the site or app. YouTube estimates that over 70 percent of views on the platform come from the company's recommended section, as opposed to self-directed searches or shared links.

These algorithms leverage **complex, distributed machine-learning models,** such as deep neural networks, to identify, rank, and serve the subset of all available posts that are predicted to be "relevant" to each user based on **how likely the user is to engage with it** via views, clicks, likes, shares, and others.

To make these engagement predictions accurate and personalized to each user and at each point in time, these algorithmic recommendation systems are **trained on billions of data points drawn from the user's prior activity history and inferred interests** (as well as those of other "similar" users) and adjusted to a particular context (e.g., time of day, device being used, etc.)

Recommendation systems are a crucial tool to drive and retain user engagement. Companies insist that their recommendation systems seek to connect users with people and content that matter to them, but critics argue that their business model prioritizes content that lures users to stay on the site longer and come back often, even if the content is controversial, harmful, or otherwise low-quality.

Numerous bills recently introduced in Congress reveal interest in regulating social media platforms due to their **large influence over users' online and offline experiences and mounting evidence of their downstream harms,** including the amplification of misinformation and harmful content, worsening mental health, and the perpetuation of bias and inequality.

Regulatory approaches to social media recommendation algorithms include legislation, self-regulation, and external oversight. Legal scholars and commentators have suggested privacy and antitrust legislation (e.g., for algorithmic interoperability) as more feasible regulatory avenues than those related to content hosting and amplification liability, since they avoid imposing content preferences (which might face

constitutional challenges) and would instead focus on increasing users' agency by having greater control over their own data and a greater variety of recommendation algorithms to choose from.

Technology-driven solutions include the design of alternative recommendation systems that optimize for human values like fairness, well-being, and factual accuracy. In practice, **aligning algorithms with complex human values is challenging**, and usually involves trade-offs and unforeseen consequences. Sustainable solutions will likely require a better understanding of how these algorithms operate and how their benefits and harms manifest, underscoring the need to **provide external researchers and regulatory agencies greater access to data** on social media platforms' algorithmic practices and outcomes. Ultimately, a successful approach to the regulation of social media recommendation algorithms will require a combination of government regulation, self governance, and external oversight to facilitate value alignment across these diverse actors and tackle the various challenges associated with this technology.

Introduction: What Are Social Media Recommendation Algorithms and Why Are They Important?

The use of social media platforms like Facebook, Twitter, YouTube, and TikTok is increasingly widespread, currently amounting to billions of users worldwide. In 2021, 72 percent of adults in the U.S. reported using at least one social media site, with the most popular ones being YouTube (81 percent) and Facebook (69 percent), and TikTok being especially common among young adults (48 percent).[1]

Social media serve as digital platforms for user-generated content. Users can range from regular individuals to celebrities, and from small communities and organizations to established companies and news outlets. The content circulated in these platforms is similarly varied and can include updates from friends and family, entertainment, tutorials and training, and increasingly, news stories. In fact, in 2021, 48 percent of adults in the U.S. got their news at least sometimes from social media, with 31 percent of people in the U.S. regularly getting news from Facebook, followed by YouTube (22 percent) and Twitter (13 percent).[2] While the content on YouTube and TikTok is primarily in video format (with TikTok specializing in short-form videos), the content on Facebook and Twitter can take various formats, including text, links to external websites, photos, videos of various lengths, or a combination of the above.

Social media companies deploy recommendation algorithms to automate the selection, ranking, and presentation of content seen by the user. Informed by users' personal characteristics and the type of content they have interacted with in the past, among many other signals, these algorithms use machine learning to select and rank content that is predicted to be most "relevant" to the user (how exactly companies define "relevant" is a thorny issue and will be discussed under "How Do Companies Define 'Relevant' and 'Valuable' Content?"). The recommended content is usually presented in the form of a highly personalized "feed" that users can scroll down at their own pace, such as Facebook's "Feed," Twitter's "Timeline," and TikTok's "For You" page, or via sidebars displayed next to the main piece of content being viewed, like YouTube's "Up Next" side panel. These content recommendation algorithms are turned on by default and have a massive influence over the content users end up consuming on these sites. For example, it is estimated that over 70 percent of views on YouTube come from the company's recommendation algorithm,[3] as opposed to self-directed searches or shared links.

1 Brooke Auxier and Monica Anderson, "Social Media Use in 2021," Pew Research Center, April 07, 2021, https://www.pewresearch.org/internet/2021/04/07/social-media-use-in-2021/.

2 Mason Walker and Katerina Eva Matsa, "News Consumption across Social Media in 2021," Pew Research Center, September 20, 2021, https://www.pewresearch.org/journalism/2021/09/20/news-consumption-across-social-media-in-2021/.

3 Joan E. Solsman, "YouTube's AI Is the Puppet Master over Most of What You Watch," CNET, January 10, 2018, https://www.cnet.com/news/youtube-ces-2018-neal-mohan/.

Although their immediate use is curating content for the users, ultimately these algorithms are a crucial tool for companies to drive and retain user engagement. Companies are invested in serving content that will attract users and keep them coming back, especially considering the strong "network effects" of social media: since individuals gravitate to platforms where their friends are, attracting a user often means attracting and retaining their wider social circles as well. Critics argue that this business model prioritizes attention-grabbing content that lures users even if the content is controversial, harmful, or otherwise low-quality. But companies insist that their content recommendation algorithms seek to improve the user experience and are guided by a set of values whose ultimate mission is to "connect users with the people and stories that matter to them most."[4]

Social media content recommendation algorithms are receiving increased scrutiny due to their large influence over users' online and offline experiences and mounting evidence of these algorithms' downstream harms. These harms include the amplification of misinformation and harmful content, mental health concerns, and the furthering of bias and inequality (see "PART 2: Public Purpose Considerations"). Although at least some of these issues are not exclusive to the use of recommendation algorithms per se (e.g., misinformation also spreads via services like WhatsApp and Telegram, which do not provide any recommendations), this algorithmic boost is exacerbating the speed and reach of these harms.

About This Document

The present document provides a policy-oriented overview of social media recommendation algorithms, including how these algorithms work, relevant public purpose considerations, and the current regulatory landscape, with a focus on the United States. This field is complex and rapidly evolving, and the reader should note that the content of this document is current as of May 2022.

In this document, when we say *social media*, we refer to popular, multi-purpose digital platforms that host content generated by all sorts of users, such as Facebook, Twitter, YouTube, and TikTok. Of note, recommendation algorithms are used by a much wider range of digital services, including entertainment companies like Spotify and Netflix for their music and movie recommendations, Amazon and the Apple App Store for purchase recommendations, and more broadly by search engines like Google. This primer focuses specifically on large, multi-purpose social media due to their broad influence over individuals' information consumption and understanding of the world, as well as social media's increasingly documented negative societal effects.

4 Adam Mosseri, "Building a Better News Feed for You," Newsroom, Meta, June 29, 2016, https://about.fb.com/news/2016/06/building-a-better-news-feed-for-you/.

When we say *content recommendation algorithms*, we refer to two kinds of algorithms: (1) algorithms that determine suggestions for *new content* that the user has not yet subscribed to (e.g., a news commentary channel on YouTube); and (2) content-ranking algorithms that filter and rank *all* the content the user sees on their feed, which combines content the user has already subscribed to ads, and new content recommendations (e.g., YouTube's news commentary mentioned above, or "People You May Know" suggestions on Facebook).

Of note, this document will not focus on content *moderation* algorithms, which are used to flag and *remove* (as opposed to amplify or recommend) content that is in violation of the platform's rules or are otherwise problematic. Limited information on the measures taken by social media companies to handle problematic content is provided under "What Is Done About Problematic Content?" Content recommendation and content removal both influence what users see on their feeds and are closely related; for example, recommender systems will implicitly set social norms for what gets amplified and what does not and, therefore, can either heighten or reduce the burden of content moderation. But they operate as two sets of algorithms that follow separate rules and have similar but not perfectly overlapping public purpose considerations. **Critically, many of the public-interest issues associated with recommendation algorithms covered in this document apply even if all the individual pieces of illegal or otherwise problematic content were instantly removed from the sites.**

PART 1: Technology

How Do Content Recommendation Algorithms Work?

Each company's content recommender systems are based on proprietary, distributed machine-learning models[5] that select and curate the content each user sees on the platform, each time they open or refresh the site or app. Generally, recommender systems make predictions of what content the user will find interesting and engage with (via views, clicks, "likes," etc.) in two main stages: 1) *candidate generation*, where the full inventory of millions of available posts are narrowed down to a few hundred that are broadly relevant to the user based on their activity history (and that of "similar" users) and context (e.g., current time and location),[6] and 2) *ranking*, where features of the candidate posts and user activity history are further analyzed to score the posts based on predicted user engagement and, finally, to select and sort the few dozen posts that are ultimately presented to the user when they open or refresh the app.

Recommender systems have historically been supported by several kinds of machine-learning algorithms, including clustering, singular value decomposition, linear and logistic regression, decision trees, and neural networks.[7] In recent years, big social media companies are increasingly incorporating advances in deep learning.[8] This machine-learning method offers several advantages, including more efficiently accommodating the vast amounts of training data, dynamically incorporating the most recent content and user actions, and making robust predictions with heterogeneous and unstructured data.

The main elements of content recommendation systems are described in more detail below.[9]

5 For an overview of machine learning, see Amy Robinson and Ariel Herbert-Voss, *Technology Factsheet: Machine Learning* (Cambridge, MA: Belfer Center for Science and International Affairs, Harvard Kennedy School, 2019), https://www.belfercenter.org/publication/technology-factsheet-machine-learning.

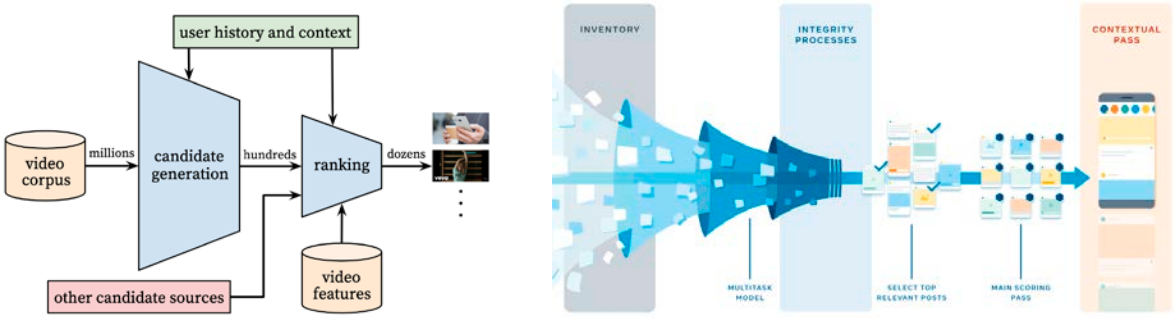
6 Paul Covington, Jay Adams, and Emre Sargin, "Deep Neural Networks for YouTube Recommendations," in *Proceedings of the 10th ACM Conference on Recommender Systems* (New York: Association for Computing Machinery, 2016), 191–8.

7 Gediminas Adomavicius and Alexander Tuzhilin, "Toward the Next Generation of Recommender Systems: A Survey of the State-Of-The-Art and Possible Extensions," *IEEE Transactions on Knowledge and Data Engineering* 17, no. 6 (2005): 734–49, https://doi.org/10.1109/TKDE.2005.99.

8 Covington, Adams, and Sarging, "Deep Neural Networks;" and Maxim Naumov et al., "Deep Learning Recommendation Model for Personalization and Recommendation Systems," *arXiv:1906.00091* (2019), https://doi.org/10.48550/arXiv.1906.00091.

9 "Our Approach to Ranking," Transparency Center, Meta, April 21, 2022, https://transparency.fb.com/features/ranking-and-content/; Cristos Goodrow, "On YouTube's Recommendation System," YouTube Official Blog, September 15, 2021, https://blog.youtube/inside-youtube/on-youtubes-recommendation-system/; and "How TikTok Recommends Videos #ForYou," Product, TikTok, June 18, 2020, https://newsroom.tiktok.com/en-us/how-tiktok-recommends-videos-for-you.

Figure 1. Reprinted schematics of YouTube's (left) and Facebook's (right) recommendation systems' architectures.[10]



Content inventory. This is all the available content (or updates to previous content) that the user has not yet seen. This includes content posted by the user's "friends," other publishers they follow, or the groups they have joined, as well as recommended ads and content the user has not yet subscribed to. The algorithm picks from an inventory of all new posts since the user last used or refreshed the platform, as well as some posts that the user already saw that have since gotten new comments or reactions, or any posts that had been selected in a previous session but were not actually seen by the user.

Data points and features. These are the data that recommender systems use to make decisions. Recommendation algorithms make predictions about whether the user will engage (and how) with each new piece of content. To do this, they are trained on millions of daily data points consisting of previous posts hosted on the platform and their associated engagement outcome (e.g., the user liked/commented/shared, the user did not engage at all, etc.). Each of these previous data points is, in turn, described by up to thousands of features related to the data point (described below), including information about the content of the piece, the device used while seeing the piece, and the user's engagement history.

- **Connectivity, device, and account settings:** These include language settings, current speed of the user's connection, whether they are viewing the content from a browser or using a phone application, etc. These types of information tend to receive lower weight relative to the information about the post and about the user.
- **Information about the content itself:** This includes what the post is about, who posted it, how old it is, what format it is (e.g., photo, video, link, etc.), how popular the post is, etc. Information about the content of a post is automatically inferred using powerful computer vision and natural language processing algorithms that can identify thousands of categories from images and videos (e.g., this video contains children playing soccer in the street) and summarize and categorize text (e.g., the comments to this post are generally positive and show compassion, encouragement, and parental advice).

10 Paul Covington, Jay Adams, and Emre Sargin, "Deep Neural Networks for YouTube Recommendations," in *Proceedings of the 10th ACM Conference on Recommender Systems* (New York: Association for Computing Machinery, 2016), 191–8, https://doi.org/10.1145/2959100.2959190; and Akos Lada, Meihong Wang, and Tak Yan, "How Does News Feed Predict What You Want to See? Personalized ranking with machine learning," Tech at Meta blog, January 16, 2021, https://tech.fb.com/news-feed-ranking/

- **User demographics, expressed preferences, and engagement history:** Metrics related to the user's previous engagement with content, including how often the user comments or likes posts from a particular source, whether the user tends to engage with a specific content format (e.g., pictures, videos, or text), how long they spend looking at or reading specific types of content, what kind of content the user has been engaging with the most lately, etc. Many of these signals, which can be explicitly measured (e.g., frequencies, time spent, etc.), are used to build high-dimensional abstract "embeddings" of a user's *inferred* taste and preferences, represented as vectors of thousands of features that are constantly being updated based on the user's engagement. Both explicit and inferred signals (as well as the signals of "similar" users, in an approach known as "collaborative filtering") are used by the algorithms.

KEY INSIGHT

Recommendation algorithms are trained on billions of data points corresponding to the user's explicit engagement history (e.g., likes and shares of certain types of content) as well as to ongoing representations of the user's *implicit* taste and interests, which the platforms infer dynamically from the user's engagement history.

Prediction and candidate generation. Thousands of signals are used to make predictions about how likely users are to engage with any given text, photo, or video, including whether they stop to look at it and for how long, and how likely they are to comment on it, share it, report it, or hide it. Sometimes the engagement likelihood score will vary by engagement type, e.g., a user might be more likely to *like* picture 1 than picture 2, but they might be more likely to *comment* on picture 2 than on picture 1. To calculate these scores for thousands of posts and billions of users every day (and every time they refresh their feeds, amounting to tens of trillions of daily predictions), companies run all models for all possible posts in parallel and on multiple computers. Following an "integrity" filter to assess if the post requires any moderation (see "What Is Done about Problematic Content?" below for more detail), the posts are narrowed down to a subset (usually a few hundred) of the most relevant so that more powerful models can be applied in the last stage.

Relevance scores and final ranking. All engagement prediction scores get weighted and combined into a "relevance" score for each post, which reflects the company's best estimate for how engaging the user will find it. Of note, there is no single definition of "relevance;" this term is often defined by short-term proxies based on measurable engagement and is highly personalized to each user (for a longer discussion, see "How Do Companies Define 'Relevant' and 'Valuable' Content?"). For example, for users that mostly interact with posts via "likes" rather than comments, the likelihood of liking a given post will be assigned a larger weight in deciding what content to show them first. Or the algorithm might penalize posts in categories that the user already interacted with earlier that day, to avoid "boredom" (as described by TikTok). Companies will sometimes also calibrate the weights of different engagement prediction scores based on user feedback on what they find valuable via surveys, interviews, and focus groups.

Relevance scores are ultimately used to sort the order in which posts appear in the user's feed at any given time, with the most relevant ones at the top. A few other sources of information are sometimes used to calibrate relevance scores. For example, Facebook's algorithms apply a final "contextual pass" over this ranking to ensure that users see a diversity of content formats when they scroll down their Feed, including a mix of videos, pictures, and text-based posts, rather than seeing just one type of content (e.g., several videos one after the other). Similarly, TikTok claims its algorithms seek to diversify the content users see on their "For You" page beyond their expressed preferences, in an effort to "break repetitive patterns" and to "help promote exposure to a range of ideas and perspectives."[11]

How Do Companies Define "Relevant" and "Valuable" Content?

As they design their products, companies must translate their big-picture mission and values (e.g., "help people connect with content and experiences that are most valuable to them")[12] into a computational and algorithmic implementation. Algorithms must optimize for something quantifiable, so company employees must determine the specific metrics the algorithms will optimize for and the relative weight assigned to each of these metrics. Generally, social media companies use engagement metrics as a proxy for content that users find interesting and broadly valuable. The logic is that the more users engage with certain content (e.g., via clicks and shares), the more appealing they find it; or conversely, a lack of engagement reflects that the users find little interest and value in that type of content.

However, selecting specific engagement metrics to optimize for often involves trade-offs and unintended consequences, notably as illustrated through the evolution of Facebook's News Feed.[13] In its origins, Facebook's News Feed algorithms optimized for basic engagement metrics like clicks and likes, and publishers (from *BuzzFeed* to individual users) quickly learned to generate attention-grabbing content that would entice the users to click, resulting in an abundance of low-quality content. As users expressed discontent, in the mid-2010s the company started penalizing clickbait and putting more weight on time spent on content as a better proxy of meaningful engagement. By 2016, the company realized that these adjustments had come at the cost of increased passive use and reduced content creation and interaction. So, in a big 2018 update,[14] Facebook announced that their algorithms would move to prioritizing "meaningful interactions between people" by showing more content from friends and family. They prioritized engagement metrics such as "back-and-forth discussion in the comments and posts that [users] share and react

11 "An Update On Our Work to Safeguard and Diversify Recommendations," Safety, TikTok, December 16, 2021, https://newsroom.tiktok.com/en-us/an-update-on-our-work-to-safeguard-and-diversify-recommendations.

12 Akos Lada, Meihong Wang, and Tak Yan, "How Does News Feed Predict What You Want to See? Personalized Ranking with Machine Learning," Tech at Meta blog, January 16, 2021, https://tech.fb.com/news-feed-ranking/.

13 Will Oremus, Chris Alcantara, Jeremy B. Merrill, and Artur Galocha, "How Facebook Shapes Your Feed," *Washington Post*, October 26, 2021, https://www.washingtonpost.com/technology/interactive/2021/how-facebook-algorithm-works/.

14 Adam Mosseri, "Bringing People Closer Together," Newsroom, Meta, January 11, 2018, https://about.fb.com/news/2018/01/news-feed-fyi-bringing-people-closer-together/.

to." The downside was that negative feelings like anger tend to be more successful at eliciting reactions and comments from users compared to positive feelings. Problematic and outrage-inducing posts thus gained a privileged position in users' News Feeds.

KEY INSIGHT

Translating companies' goals into specific, quantifiable metrics to be optimized by recommendation algorithms is a challenging task that often involves trade-offs and unintended consequences.

Leaked company documents, such as the 2021 "Facebook Papers," revealed that social media companies are generally aware of the downsides of their algorithmic choices.[15] Companies routinely assess the performance of various versions of the algorithm via user research and A/B experiments on subsets of users before they do an official rollout on the entire user base (see "How and How Often Are These Algorithms Updated?" for more detail), and internal research teams of social scientists and ethicists evaluate the short- and long-term impacts of their products. Companies claim that users' reports on the types of experiences and content they find enjoyable and meaningful (collected via surveys, interviews, and user-initiated reports) are used to inform their values and to tweak their recommendations and content moderation algorithms, e.g., by elevating authentic stories and reducing the distribution of misleading, sensational, and spammy ones (see "What Is Done about Problematic Content?").

However, it is unclear whether these measures can appropriately mitigate the effects of recommendation algorithms that are primarily optimized for observed, quantitative engagement metrics rather than for what users qualitatively report to find truly valuable.

How and How Often Are These Algorithms Updated?

The machine-learning algorithms used for content recommendation systems are dynamic and continuously fine-tuned to provide more accurate predictions of what users will find interesting, and to adjust for changes in usage trends over time. These updates reflect the notion that, even within a company, we cannot refer to recommendation algorithms as a singular entity.

The currently deployed "production" recommendation systems are retrained with new incoming data as often as every few hours.[16] Additionally, when engagement metrics are down, engineers get automatically

15 Georgia Wells, Jeff Horwitz, and Deepa Seetharaman, "Facebook Knows Instagram Is Toxic for Teen Girls, Company Documents Show," *Wall Street Journal*, September 14, 2021, https://www.wsj.com/articles/facebook-knows-instagram-is-toxic-for-teen-girls-company-documents-show-11631620739.

16 Krishna Gade (@krishnagade), "I was an eng leader on Facebook's NewsFeed and my team was responsible for the feed ranking platform," Twitter, February 11, 2021, 11:55 a.m., https://twitter.com/krishnagade/status/1359908897998315521.

notified so they can further diagnose the problem using a real-time analytics system, fix bugs, and retrain the models if needed.[17]

Companies are also continuously experimenting with new models.[18] For example, engineers at Facebook use a platform called FBLearner Flow[19] to train multiple complex machine-learning models in just a few days, adjusting things like the input features used by the model to describe each data point and the ranking model architecture. These new models are compared in parallel with A/B testing, through which different pockets of users receive slightly different versions of the algorithm to see how each performs on engagement metrics, including daily active users and their rate of likes, comments, and shares. The best performing ones are kept and further refined, and they are eventually rolled out to the full user base.

What Is Done about Problematic Content?

Social media companies employ content moderation practices to identify problematic content that should be removed, flagged, or reduced. Although content moderation is done in parallel to their recommendation systems, the two are related insofar as content judged to be problematic is either removed from the platform or flagged and demoted by the content-ranking recommendation algorithms.

Companies use their public policies and community standards to present their core values (e.g., safety, privacy, diversity, dignity, and authenticity), define what is and is not allowed on their platforms based on those values, and outline their content moderation enforcement procedures. Content moderation comes in a few different forms:[20]

- **Removing.** Content that is judged to violate standards is removed and does not appear on users' feeds after that. This can include violent and graphic content, violent extremism, hateful behavior, illegal activities, suicide, self-harm and dangerous acts, child abuse, harassment and bullying, adult nudity and sexual activities, and misinformation that has the potential to cause physical harm or interfere with voting.
- **Reducing distribution.** Content that is problematic but that does not necessarily violate community standards (i.e., "borderline" content, which ranges from misinformation to clickbait, spam, sensationalistic content, and low-quality videos or comments) gets down-ranked by the content-ranking algorithm, effectively reducing the visibility of that content. In many cases, borderline content also gets an added label to warn users.

17 Karen Hao, "How Facebook Got Addicted to Spreading Misinformation," MIT Technology Review, March 11, 2021, https://www.technologyreview.com/2021/03/11/1020600/facebook-responsible-ai-misinformation/.

18 "How TikTok Recommends Videos #ForYou," Newsroom, TikTok; and Goodrow, "On YouTube's Recommendation System."

19 Jeffrey Dunn, "Introducing FBLearner Flow: Facebook's AI Backbone," Engineering at Meta blog, May 09, 2016, https://engineering.fb.com/2016/05/09/core-data/introducing-fblearner-flow-facebook-s-ai-backbone/.

20 "Taking Action," Transparency Center, Meta, accessed December 28, 2021, https://transparency.fb.com/enforcement/taking-action/.

- **Labeling.** Companies add labels to posts that might be sensitive or misleading, even if they do not violate their community standards, to help users decide what to view, trust, and share. For example, Facebook shows warning screens over content that might be explicit and a pop-up screen warning a post might be outdated before users share it.
- **Demonetizing.** Platforms like YouTube sometimes withhold the ad revenue of specific videos or an entire account when they are in violation of the company's "advertiser-friendly" content guidelines.[21] Although demonetized videos are allowed to remain on the platform, blocking ad revenue is meant to deter the creation of this type of content and reduce its visibility.

The flagging and removal of content is done by a combination of automated algorithms, human review, and user reports. Algorithms help flag and label content at scale, and some posts get passed on to humans for review. AI-driven estimates of severity of potential harm, virality, and likelihood of being in violation of policies are used to prioritize human review. For example, to determine borderline content, YouTube reviewers evaluate whether the content of a video is inaccurate, misleading or deceptive, insensitive or intolerant, and harmful or with the potential to cause harm.[22] These are combined into a composite score reflecting likelihood to contain harmful misinformation or be borderline. Human decisions are used to continuously retrain the automated systems.

Companies claim that content that either violates standards or is otherwise considered borderline is a very small fraction of all the content shared on their sites. For example, YouTube claims that consumption of borderline content that comes from the company's recommendations is well below 1 percent of total views on the site.[23] However, some have contested that these numbers might be misleading, as human raters cannot readily review the vast amount of content posted to these sites, and therefore problematic content can remain on the site for a while without getting a borderline label and without getting demoted. Moreover, given how easy it is to share content in social media, even a very small portion of users and content can still have an outsized reach across the user base. For example, a recent investigation found that two-thirds of all anti-vaccine content circulated on Facebook and Twitter were created by only twelve individual accounts (referred to as "the disinformation dozen").[24]

21 Robyn Caplan and Tarleton Gillespie, "Tiered Governance and Demonetization: The Shifting Terms of Labor and Compensation in the Platform Economy," Social Media+Society 6, no. 2 (2020): 2056305120936636, https://journals.sagepub.com/doi/full/10.1177/2056305120936636.

22 Goodrow, "On YouTube's Recommendation System."

23 Ibid.

24 *The Disinformation Dozen: Why Platforms Must Act on Twelve Leading Online Anti-Vaxxers* (Washington, D.C.: Center for Countering Digital Hate, March 24, 2021), https://www.counterhate.com/disinformationdozen.

What Information and Control Does the User Currently Have?

Given the heightened scrutiny over the impact of their content recommendation algorithms, companies have taken a series of measures to improve transparency over how their algorithms work and to give users some control over the curation of their own feeds.

Explainability and transparency. A general criticism of recommendation algorithms, and of complex artificial intelligence techniques more broadly, is that they are not explainable.[25] They are likened to a "black box," where even its designers cannot fully explain why machines arrive at a specific decision, beyond general notions of the data used and the overarching architecture of the model (like those described under "How Do Content Recommendation Algorithms Work?"). Seeking to increase transparency and trust in their recommendation systems, social media companies publish general information on their algorithms and the kinds of data they use, though the amount of information provided pales in comparison to the complexity of their machine-learning models and the thousands of data features fed into the algorithms. Additionally, companies like Facebook have added features like a "Why am I seeing this?"[26] button to each post on their Feed, which provides a brief explanation for how the users' previous interactions on the site influenced the content selected by the algorithm and allows the user to change their Feed settings. Curated algorithm performance metrics are also released in periodic public reports accompanied by blog posts summarizing their content, but the data used to generate the reports are not shared, thus restricting external assessment.

Feedback and user-controlled feed curation. Platforms covered in this document include features for users to provide feedback on the individual pieces of content they see on their feeds. This includes hiding posts they are not interested in, unfollowing an account that posted the content, and reporting the content as violating community standards. YouTube users can also indicate whether a video recommendation was helpful or not, which is incorporated into the algorithm so that less or more of that kind of content is shown on their feeds. YouTube users can view, pause, and edit their search history, which will also influence recommendations.

Additionally, some companies offer features for alternative content curation. Facebook and Twitter allow users to display the feed or timeline content in reverse chronological order instead of the default recommended content. Facebook also offers a "Favorites" feature that only shows the posts of up to thirty

25 Jessica Fjeld et al., "Principled Artificial Intelligence: Mapping Consensus in Ethical and Rights-Based Approaches to Principles for AI," Berkman Klein Center Research Publication No. 2020-1, January 15, 2020, http://dx.doi.org/10.2139/ssrn.3518482; Paul Voosen, "How AI Detectives Are Cracking Open the Black Box of Deep Learning," Science, July 06, 2017, https://www.science.org/content/article/how-ai-detectives-are-cracking-open-black-box-deep-learning; and Jessica Newman, "Explainability Won't Save AI," Tech Stream, Brookings Institution, May 19, 2021, https://www.brookings.edu/techstream/explainability-wont-save-ai/.

26 Ramya Sethuraman, "Why Am I Seeing This? We Have an Answer for You," Newsroom, Meta, March 31, 2019, https://about.fb.com/news/2019/03/why-am-i-seeing-this/.

pages and accounts of the user's choice. However, these settings cannot be adopted permanently; the feed reverts to the default recommendation algorithm each time the user leaves or refreshes the site or app.

What Control *Could* the User Have?

Various technologists, scholars, and activists have proposed to let users access a wider range of options for how to display the content they see on traditional social media. Computer scientist Stephen Wolfram has proposed third-party "final ranking providers,"[27] which would take pre-digested feature vectors from the underlying content platform, then use these to do the final ranking of items using different optimization criteria. A similar proposal is that of "middleware" companies,[28] which would sit on top of traditional social media platforms to provide third-party alternatives for content ranking, recommendation, and moderation algorithms that users could choose from. Others have proposed cross-platform interoperability that would allow users to aggregate content across multiple social media accounts, for example, into a third-party "social media browser"[29] where users could experiment with the content-curating algorithms (e.g., users could select to display content only posted by females, or filter posts by "rudeness" score).

Such proposals argue that a wider range of recommendation algorithms that users can choose from (or even tweak themselves) could lead to better options and dampen the network effects that tend to keep a user locked into a specific social media service. Users would be able to select the flavor of content recommendation that fits their needs and explore alternative social media services altogether (e.g., ones designed to support local communities[30] and civic engagement[31]) without losing the information and connections they have formed in mainstream social media, or without needing all their contacts to move to the alternative social media service with them. Although some fear that letting users have more granular control on how to display their content might heighten echo chambers, the proponents of these alternatives argue that such echo chambers already exist and that, at least this way, users would be conscious of their choices, much like deciding to watch one news channel over another.[32]

27 Stephen Wolfram, "Testifying at the Senate about AI-Selected Content on the Internet," Stephen Wolfram Writings blog, June 25, 2019, https://writings.stephenwolfram.com/2019/06/testifying-at-the-senate-about-a-i-selected-content-on-the-internet/

28 Francis Fukuyama, "Making the Internet Safe for Democracy," *Journal of Democracy* 32, no. 2 (2021): 37–44.

29 Rahul Bhargava et al., "Gobo: A System for Exploring User Control of Invisible Algorithms in Social Media," in *Conference Companion Publication of the 2019 on Computer Supported Cooperative Work and Social Computing* (Austin: Association for Computing Machinery, 2019), 151–5, https://doi.org/10.1145/3311957.3359452.

30 Chand Rajendra-Nicolucci and Ethan Zuckerman, "Local Logic: It's Not Always a Beautiful Day in the Neighborhood," in *An Illustrated Field Guide to Social Media,* eds. Chand Rajendra-Nicolucci and Ethan Zuckerman (New York: Knight First Amendment Institute, 2021), 17–22. https://knightcolumbia.org/blog/an-illustrated-field-guide-to-social-media.

31 Chand Rajendra-Nicolucci and Ethan Zuckerman, "Civic Logic: Social Media with Opinion and Purpose," in *An Illustrated Field Guide to Social Media,* eds. Rajendra-Nicolucci and Zuckerman (New York: Knight First Amendment Institute, 2021), 9–15.

32 Wolfram, "Testifying at the Senate;" and Fukuyama, "Making the Internet Safe for Democracy."

The idea of optimizing third-party content-curation algorithms for pro-social outcomes has been informed by the field of mechanism design for social good.[33] For example, proponents have suggested that alternative recommendation algorithms could be built to optimize for human values like diversity, fairness, well-being, time well spent, and factual accuracy.[34] However, a big challenge in this area is how to translate such complex human values into quantitative metrics that can be optimized by algorithms, also referred to as the "AI alignment" problem.[35] All such decisions involve trade-offs and potentially unforeseen downstream consequences (see "How Do Companies Define 'Interesting' and 'Valuable' Content?" for a discussion of this topic).

The creation of alternative approaches that ameliorate some of the harms of Big Tech's recommendation algorithms necessitates a thorough understanding of how these algorithms operate and how their benefits and harms manifest. Thus, in addition to interoperability, a key step toward building public-interest recommendation algorithms is providing qualified researchers greater access to data on social media platforms' algorithmic practices and outcomes (see "Third-Party Research" for more information). Proposed regulation related to third-party interoperability and research are included in the European Commission's 2020 Digital Services Act proposal and the Platform Accountability and Transparency Act introduced in the U.S. Congress in December 2021 (see "Proposed Legislation" for more details).

33 Rediet Abebe and Kira Goldner, "Mechanism Design for Social Good," *AI Matters* 4, no. 3 (2018): 27–34.

34 Jonathan Stray, Ivan Vendrov, Jeremy Nixon, Steven Adler, and Dylan Hadfield-Menell, "What Are You Optimizing For? Aligning Recommender Systems with Human Values," *arXiv:2107.10939* (2021), https://arxiv.org/abs/2107.10939.

35 Shira Mitchell, Eric Potash, Solon Barocas, Alexander D'Amour, and Kristian Lum, "Algorithmic Fairness: Choices, Assumptions, and Definitions," *Annual Review of Statistics and Its Application* 8 (2021): 141–63, https://doi.org/10.1146/annurev-statistics-042720-125902; and Brian Christian, *The Alignment Problem: Machine Learning and Human Values* (WW Norton & Company, 2020).

PART 2: Public Purpose Considerations

Especially as social media platforms permeate the daily lives of billions of individuals, the recommended content found on these sites can have a large impact on their understanding of the world and the decisions they make both online and offline. New evidence continues to mount regarding social media's contribution to the erosion of information ecosystems, the exacerbation of mental health issues, and the perpetuation of biases and discrimination.

KEY INSIGHT

As social media platforms become ingrained in modern daily life, the recommended content found on these sites can have a large impact on people's thoughts, feelings, and behaviors, with potentially long-lasting consequences for individuals and society at large.

Influencing how we make sense of the world. Social media's algorithmic filtration and curation of content influences what type of information individuals pay attention to, leading to a closed loop of preference amplification where the algorithm shows individuals content that grabs their attention, and individuals pay more attention to what the algorithm shows.[36] The reinforcement of related content through recommendation algorithms makes it such that a few clicks on controversial content can result in extremist content recommendations that are difficult to stop receiving and that might influence their beliefs, feelings, and preferences.[37] For example, a recent *Wall Street Journal* investigation on TikTok deployed bot users to watch recommended sad videos for longer than other videos, and after only about thirty minutes of total watch time, 93 percent of recommended videos were depression-related.[38]

Influencing decision-making offline, from elections to public health. The influence of social media recommendation algorithms is not limited to our thoughts. Reports reveal that social media platforms can affect real-world behavior on a large scale, from purchasing trends and voter and protest turnout, to treatment of minorities and pandemic response. A 2012 study found that a recommended message on Facebook showing pictures of the user's "friends" who had already voted increased turnout in the 2010 US congressional elections directly by about 60,000 voters and indirectly through social contagion by another 280,000 voters, for a total of 340,000 additional votes.[39] Moreover, according to a 2018 United

36 Dimitris Kalimeris, Smriti Bhagat, Shankar Kalyanaraman, and Udi Weinsberg, "Preference Amplification in Recommender Systems," in *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining* (New York: Association for Computing Machinery, 2021): 805–15, https://doi.org/10.1145/3447548.3467298.

37 *Algorithms and Amplification: How Social Media Platforms' Design Choices Shape Our Discourse and Our Minds,* Subcommittee on Privacy, Technology, and the Law (April 27, 2021) (statement by Joan Donovan, Research Director, Shorenstein Center on Media, Politics and Public Policy, Harvard Kennedy School), https://www.judiciary.senate.gov/imo/media/doc/Donovan%20Testimony%20(updated).pdf.

38 WSJ Staff, "Inside TikTok's Algorithm: A WSJ Video Investigation," *Wall Street Journal*, July 21, 2021, https://www.wsj.com/articles/tiktok-algorithm-video-investigation-11626877477.

39 Robert M. Bond et al., "A 61-Million-Person Experiment in Social Influence and Political Mobilization," *Nature* 489, no. 7415 (2012): 295–8, https://www.nature.com/articles/nature11421.

Nations investigation, social media (especially Facebook) had a "determining role" in amplifying hate speech and the escalation of violence against the Rohingya ethnic minority in Myanmar.[40] Social media content has also played a role in influencing the public's attitudes and adoption of public health measures during the COVID-19 pandemic. A recent study found that among participants with the lowest baseline knowledge about the virus, social media were the preferred source of pandemic-related information, and this was in turn associated with greater engagement with protective behaviors.[41] Another study found that misinformation on Twitter was associated with early COVID-19 vaccination hesitancy and refusal, even after accounting for political, demographic, and socioeconomic factors.[42]

Rewarding harmful, misleading, and extreme content. Posts containing offensive, disturbing, false, extreme, or otherwise harmful content are more likely to receive engagement via clicks, likes, comments, and shares, thus revealing that sensational and problematic content is alluring to the average social media user. A study found that each word of moral outrage added to a tweet increases the rate of retweets by 20 percent.[43] Although these platforms have worked to label and reduce the spread of low-quality and harmful content via both automated and human reviewers, they are insufficient in an engagement-driven content-curation system that tends to favor controversial and emotional content (not to mention the limited content moderation resources outside of the United States; see "Bias and Inequality" point below). For example, Facebook's most widely viewed page in the last quarter of 2021, with over 120 million content viewers, was a page that the company later removed for violating Facebook's Community Standards.[44]

Eroding integrity of information ecosystems. Media scholars have found that mis/disinformation often performs better on social media platforms, spreading farther, faster, deeper, and more broadly than true news in all categories of information and most strongly for political news.[45] While the problem of mis/disinformation is not exclusive to digital platforms that use recommendation algorithms, this algorithmic amplification can help extend their reach at high speeds,[46] including through "coordinated inauthentic behavior" that seeks to game the algorithms to boost misleading content via coordinated fake accounts. Although some research shows that, generally, people who get their news online are exposed

40 Tom Miles, "U.N. Investigators Cite Facebook Role in Myanmar Crisis," Reuters, March 12, 2018, https://www.reuters.com/article/us-myanmar-rohingya-facebook/u-n-investigators-cite-facebook-role-in-myanmar-crisis-idUSKCN1GO2PN.

41 Sooyoung Kim et al., "Impact of COVID-19-Related Knowledge on Protective Behaviors: The Moderating Role of Primary Sources of Information," *PloS one* 16, no. 11 (2021): e0260643.

42 Francesco Pierri et al., "Online Misinformation Is Linked to Early COVID-19 Vaccination Hesitancy and Refusal," *Scientific Reports* 12, no. 1 (2022): 1–7.

43 William J. Brady et al., "Emotion Shapes the Diffusion of Moralized Content in Social Networks," *Proceedings of the National Academy of Sciences* 114, no. 28 (2017): 7313–8.

44 *Widely Viewed Content Report: What People See on Facebook,* Q4 2021 Report (Menlo Park: Transparency Center, Meta, 2022), https://transparency.fb.com/data/widely-viewed-content-report/.

45 Laura Edelson et al., "Understanding Engagement with US (Mis)Information News Sources on Facebook," in *Proceedings of the 21st ACM Internet Measurement Conference* (New York: Association for Computing Machinery, 2021), 444–63; and Soroush Vosoughi, Deb Roy, and Sinan Aral, "The Spread of True and False News Online," *Science* 359, no. 6380 (2018): 1146–51, https://www.science.org/doi/full/10.1126/science.aap9559.

46 Ferenc Huszár et al., "Algorithmic Amplification of Politics on Twitter," *Proceedings of the National Academy of Sciences* 119, no. 1 (2022), https://doi.org/10.1073/pnas.2025334119.

to more diverse views than people who use traditional sources[47] (mostly due to incidental exposure),[48] the hyper-personalization of information diets can lead to echo chambers and more extreme rejection of opposing views when they encounter them online. For example, a recent study found that Republican participants expressed substantially more conservative views after following a liberal Twitter bot,[49] and another study found that debunking posts were ignored by 98 percent of conspiracy theory information users, and when they did interact with the posts via comments, this was followed by greater number of likes and comments in conspiracy echo chambers.[50]

Social comparison and exacerbating mental health issues. Academic studies, as well as internal research by social media companies leaked to the press, have shown that social comparison on online platforms can make users feel worse about themselves and exacerbate mental health issues, especially among young females.[51] Internal research from Instagram showed that the use of the app makes body issues worse for one in three teenage girls, and among adolescents who reported suicidal thoughts, 13 percent of British users and 6 percent of U.S. users could trace the desire to kill themselves to Instagram. Moreover, a 2021 annual survey found that 77 percent of plastic surgeons report their patients' main motivation for undergoing facial cosmetic surgery is looking good on "selfies" posted to social media.[52]

Bias and inequality. Based on the underlying training data, recommendation algorithms may learn and amplify existing demographic biases. For example, algorithms might not recommend or display certain housing, education, employment, or credit opportunities based on the user's race and gender.[53] Algorithms might also reduce the recommendation (and visibility) of content created by women and ethnic minorities[54] or disproportionately penalize certain groups in algorithmic detection and demotion of harmful content.[55] Furthermore, the amplification of harmful content—including violence against ethnic minorities, employment ads that result in human trafficking, pornography-related content, and the

47 Seth Flaxman, Sharad Goel, and Justin M. Rao, "Filter Bubbles, Echo Chambers, and Online News Consumption," *Public Opinion Quarterly* 80, no. S1 (2016): 298–320, https://doi.org/10.1093/poq/nfw006; and Eytan Bakshy, Solomon Messing, and Lada A. Adamic, "Exposure to Ideologically Diverse News and Opinion on Facebook," *Science* 348, no. 6239 (2015): 1130–32, https://doi.org/10.1126/science.aaa1160.

48 Richard Fletcher and Rasmus Kleis Nielsen, "Are People Incidentally Exposed to News on Social Media? A Comparative Analysis," *New Media & Society* 20, no. 7 (2018): 2450–68, https://doi.org/10.1177/1461444817724170.

49 Christopher A. Bail et al. "Exposure to Opposing Views on Social Media Can Increase Political Polarization," *Proceedings of the National Academy of Sciences* 115, no. 37 (2018): 9216–21, https://doi.org/10.1073/pnas.1804840115.

50 Walter Quattrociocchi, Antonio Scala, and Cass R. Sunstein, *Echo Chambers on Facebook*, SSRN Scholarly Paper (Rochester, NY: Social Science Research Network, 2016), https://doi.org/10.2139/ssrn.2795110.

51 Wells, Horwitz, and Seetharaman, "Facebook Knows Instagram Is Toxic;" and Marika Tiggemann, Susannah Hayden, Zoe Brown, and Jolanda Veldhuis, "The Effect of Instagram 'Likes' on Women's Social Comparison and Body Dissatisfaction," *Body Image* 26 (2018): 90–7, https://www.sciencedirect.com/science/article/abs/pii/S1740144518301360.

52 "AAFPRS Announces Annual Survey Results: Demand for Plastic Surgery Skyrockets as Pandemic Drags On," American Academy of Facial Plastic and Reconstructive Surgery, February 10, 2022, https://www.aafprs.org/Media/Press_Releases/2021%20Survey%20Results.aspx.

53 Muhammad Ali et al., "Discrimination through Optimization: How Facebook's Ad Delivery Can Lead to Biased Outcomes," *Proceedings of the ACM on Human-Computer Interaction,* vol. 3 (New York: Association for Computing Machinery, 2019), 1–30, https://doi.org/10.1145/3359301.

54 Michael Ekstrand and Daniel Kluver, "Exploring Author Gender in Book Rating and Recommendation," *User Modeling and User-Adapted Interaction* 31, no. 3 (2021): 377–420, https://link.springer.com/article/10.1007/s11257-020-09284-2.

55 Maarten Sap et al., "The Risk of Racial Bias in Hate Speech Detection," in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics* (Florence: Association for Computational Linguistics, 2019), 1668–78, https://aclanthology.org/P19-1163.

repression of political dissent—are more pronounced in developing countries, at least partly due to the reduced allocation of content moderation resources in these regions. For example, according to a 2020 internal Facebook report, although the United States comprises less than 10 percent of Facebook's daily users, 84 percent of the company's budget to fight misinformation was devoted to the United States, with the remaining 16 percent allocated to the rest of the world and its thousands of languages.[56]

56 Cat Zakrzewski, Gerrit De Vynck, Niha Masih, and Shibani Mahtani, "How Facebook Neglected the Rest of the World, Fueling Hate Speech and Violence in India," *Washington Post*, October 24, 2021, https://www.washingtonpost.com/technology/2021/10/24/india-facebook-misinformation-hate-speech/; and Justin Scheck, Newley Purnell, and Jeff Horwitz, "Facebook Employees Flag Drug Cartels and Human Traffickers. The Company's Response Is Weak, Documents Show," *Wall Street Journal*, September 16, 2021, https://www.wsj.com/articles/facebook-drug-cartels-human-traffickers-response-is-weak-documents-11631812953.

PART 3: Regulation and Oversight

This section provides an overview of relevant legislation, proposed bills, self-regulatory approaches, and external oversight approaches to the regulation of social media algorithms. A big challenge for policymakers is to craft solutions that address the present harms caused by social media recommendation algorithms without amplifying other harms (e.g., strict regulations could pose insurmountable financial strain on smaller companies, thus further reducing competition) or severely downgrading the user experience on these platforms (e.g., the banning of algorithmic recommendations and mandating that content is shown in reverse chronological order could lead to frequent spam by content creators just to stay at the top of users' social media feeds).

KEY INSIGHT

Sustainable regulation of social media recommendation algorithms will require a combination of government regulation, self governance, and external oversight approaches to tackle the various challenges associated with this technology.

Legislation and Proposals

There are several legislative approaches relevant to the regulation of recommendation algorithms, including legislation related to speech, privacy, and antitrust and competition. Legal scholars and commentators have suggested privacy and antitrust regulation as more feasible avenues to regulate platforms' algorithmic content curation, since they avoid imposing government control over content preferences, which might run into constitutional dead-ends under the First Amendment. Instead, privacy and antitrust legislation would be grounded on increasing users' agency via greater data control and a wider range of recommendation algorithms to choose from. However, this type of regulations might face the fiercest opposition and lobbying from platforms, as it would likely lead to substantial changes in their business models.

Speech Legislation

Section 230 of the Communications Decency Act of 1996 (CDA 230). Enforced by the Federal Communications Commission, CDA 230 provides protection to "interactive computer service providers," including social media companies, from liability related to: (1) hosting information provided by another information content provider, and (2) voluntarily acting "in good faith" to restrict access to objectionable material. These two immunity provisions do not apply in federal criminal laws, intellectual property laws,

and for several sex trafficking offenses (a carveout known as SESTA-FOSTA passed in Congress in 2018).[57] Many commentators and legislators argue that CDA 230 should be revised, especially since the internet looks very different from what it did when this law was first enacted in 1996. More specifically, they propose to further narrow online platforms' immunity or even repeal immunity altogether. However, others have noted that these proposals might face constitutional challenges under the First Amendment.[58]

First Amendment. The creation and curation of content is considered speech by the Supreme Court, which means that content on social media platforms can be constitutionally protected as lawful speech under the First Amendment. Crucially, attempting to regulate the role of social media content recommendation algorithms in the *amplification* (or demotion) of content will run into the same strict First Amendment scrutiny and logistical issues as regulating the *removal* or banning of content altogether.

Correctly identifying illegal speech is hard given the nuances of human communication and of complex legal systems, and this is even harder to incorporate into automated AI algorithms. Laws that mandate the swift removal of illegal content might incentivize platforms to err on the side of over-enforcing and suppressing lawful speech just to be safe, therefore running into First Amendment challenges.[59] However, some argue that certain categories of speech may be more readily regulated (e.g., via the creation of additional carve outs to CDA 230 immunity like SESTA-FOSTA). This includes content promoting terrorism or violence, given that incitement and true threats are speech categories that have received more limited First Amendment protection in the past.[60]

Privacy Legislation

Recent data protection legislation in Europe and in the United States requires greater transparency regarding how individuals' personal data are used, including for automated decision-making like social media content recommendation algorithms. Privacy protection provides users more control over recommendation algorithms, for example, by selecting to exclude certain data streams from being used by these algorithms, which would in turn influence the type of recommendations received.

European Union's General Data Protection Regulation (GDPR). The GDPR provides comprehensive regulation of personal data processing. Most relevant to social media recommendation algorithms, the GDPR protects individuals' right to be informed about the collection and use of their personal data in a

57 SESTA, the Stop Enabling Sex Traffickers Act, and FOSTA, the Fight Online Sex Trafficking Act, were Senate and House-sponsored bills, respectively.

58 Daphne Keller, "Amplification and Its Discontents: Why Regulating the Reach of Online Content Is Hard," *Journal of Free Speech Law* 1 (2021): 227–68.

59 Ibid.

60 Valerie Brannon, "Regulating Big Tech: Legal Implications," Congressional Research Service, September 11, 2019, https://crsreports.congress.gov/product/pdf/LSB/LSB10309.

manner that is concise, transparent, easily accessible, and presented in plain language, as well as individuals' right to object to data processing that involves automated decision-making without human involvement. Social media companies with a broad international market like the ones covered in this document must comply with the GDPR, since this legislation's reach also extends to non-EU-established entities that offer goods or services to individuals in the EU. This has generally resulted in these companies adopting stricter data protection measures for all their customers regardless of geographical location.

California Consumer Privacy Act (CCPA). The United States does not have a comprehensive federal data protection law. At the state level, the CCPA is the most sweeping personal data protection in the United States and was largely inspired by the GDPR. The CCPA provides consumers with three main rights: the right to know what data businesses collect about them and for what purpose, the right to opt out of the selling of personal data, and the right to delete personal data collected on them.[61] In the context of social media, companies may need to disclose to consumers what data they use for their algorithmic recommendations and allow consumers to delete that personal information.

Children's Online Privacy Protection Act of 1998 (COPPA). Implemented by the Federal Trade Commission, COPPA[62] provides data protection requirements for children's information collected by online operators. Specifically, COPPA prohibits collecting personal data from children under the age of thirteen without obtaining parental consent prior to data collection.

Antitrust and Competition Legislation

Relatively few social media companies—and their respective content-curation algorithms—dominate the market, and individuals have little choice but to accept the terms of these major platforms. Proponents of antitrust legislation contend that a greater number of competing social media platforms and recommendation algorithms would reduce the domination and reach of any given platform's content curation, therefore providing individuals with more options and the chance to adjust their preferences.[63] Two key pieces of antitrust law in the United States are the Clayton and Sherman Antitrust Acts.[64]

Clayton Antitrust Act of 1914. This act prohibits any merger that could substantially lessen competition or tend to create a monopoly, and it could be used to retroactively unwind past mergers (e.g., the Facebook-Instagram merger from 2012). However, the government would have to show that the merger

61 Eric Holmes, "California Dreamin' of Privacy Regulation: The California Consumer Privacy Act and Congress," Congressional Research Service, November 01, 2018, https://crsreports.congress.gov/product/pdf/LSB/LSB10213.

62 Stephen P. Mulligan and Chris D. Linebaugh, "Data Protection Law: An Overview," Congressional Research Service, March 25, 2019, https://crsreports.congress.gov/product/pdf/R/R45631.

63 Ryan Tracy and John McKinnon, "Antitrust Tech Bills Gain Bipartisan Momentum in Senate," *Wall Street Journal*, November 25, 2021, https://www.wsj.com/articles/antitrust-tech-bills-gain-bipartisan-momentum-in-senate-11637836202?mod=article_inline.

64 Jake Sykes, "Antitrust Law: An Introduction," Congressional Research Service, May 29, 2019, https://crsreports.congress.gov/product/pdf/IF/IF11234.

meaningfully decreased competition in a particularly defined marketplace, a challenging task in the case of dynamic technology industries.

Sherman Antitrust Act of 1890. Section 2 of this act makes it illegal to monopolize or to conspire to monopolize a marketplace, which could be used to break up companies like Meta for allegedly engaging in exclusionary conduct. However, this might require evidence for the companies' market power (defined as the ability to profitably raise prices above competitive levels), which is challenging as social media companies do not charge users for their services.

Other practical antitrust challenges include the multiple barriers to entry faced by new companies to successfully compete with Big Tech's massive resources and the network effects of social media. This might mean that even if Big Tech social media companies get broken up, they might be able to grow back up to their original size relatively quickly. However, some have proposed that these challenges might be surmountable by allowing new competitors (and even individual users) to borrow and customize some of the infrastructure built by Big Tech social media companies by providing open-access protocols[65]—like the ones that dominated the early internet—or "middleware" software[66] with various options for content ranking, recommendation, and moderation algorithms that would sit on top of the social media service and that users could choose from (see "What Control Could the User Have?" for more details).

KEY INSIGHT

Privacy and antitrust legislation focus on increasing users' agency by having greater control over their data and a variety of recommendation algorithms to choose from. These might be more feasible regulatory avenues than those related to content hosting and amplification liability, since they avoid imposing content preferences.

Proposed Legislation

United States

Members of the U.S. Congress have introduced bills (many of them bipartisan and in some cases bicameral) that attempt to curb Big Tech's power and their products and services' associated harms. Here we present only the bills that would impact social media recommendation algorithms and that have been introduced by Representatives and Senators as part of the 117th Congress cycle (2021–2022). Legislative

65 Mike Masnick, "Protocols, Not Platforms: A Technological Approach to Free Speech," Knight First Amendment Institute, August 21, 2019, https://knightcolumbia.org/content/protocols-not-platforms-a-technological-approach-to-free-speech.

66 Fukuyama, "Making the Internet Safe for Democracy."

efforts related to antitrust and children's privacy have received explicit support from President Joe Biden's administration, including through executive orders[67] and in public addresses by the president.[68]

Bills Related to Algorithmic Transparency and CDA 230 Reform

- **S. 2024. Filter Bubble Transparency Act.**[69] First proposed in the House in 2019 and reintroduced in the Senate in November 2021 by Reps. Ken Buck (R-CO), David Cicilline (D-RI), Lori Trahan (D-MA), and Burgess Owens (R-UT). This bill proposal harps on "opaque algorithm requirements" and would require that "internet platforms give users the option to engage with a platform without being manipulated by algorithms driven by user-specific data."[70]
- **H.R. 5596. Justice Against Malicious Algorithms Act.**[71] Introduced in October 2021 by Reps. Frank Pallone Jr. (D-NJ), Michael Doyle (D-PA), Janice Schakowsky (D-IL), and Anna Eshoo (D-CA). The bill proposal would "carve out Section 230 so that a digital service could face liability if they knowingly or recklessly make a personalized recommendation that materially contributed to a physical or severe emotional injury to any person." These measures would apply to content recommendations that use algorithms to boost certain content over others based on the users' personal information.
- **S. 1896/H.R. 3611. Algorithmic Justice and Online Platform Transparency Act.**[72] Introduced in May 2021 by Sen. Edward J. Markey (D-MA) and Rep. Doris Matsui (D-CA). This bill proposal is set "to prohibit harmful algorithms, increase transparency into websites' content amplification and moderation practices, and commission a cross-government investigation into discriminatory algorithmic processes throughout the economy."[73]
- **Algorithmic Accountability Act.**[74] Presented in February 2022 by Sens. Ron Wyden (D-OR), Cory Booker (D-NJ), and Rep. Yvette Clarke (D-NY). An update of their 2019 bill by the same name, this new bill proposal, to be introduced in the Senate and House, "requires companies to assess the impacts of the automated systems they use and sell, creates new transparency about when and how

67 "FACT SHEET: President Biden to Announce Strategy to Address Our National Mental Health Crisis, as Part of Unity Agenda in His First State of the Union," The White House Briefing Room, March 01, 2022, https://www.whitehouse.gov/briefing-room/statements-releases/2022/03/01/fact-sheet-president-biden-to-announce-strategy-to-address-our-national-mental-health-crisis-as-part-of-unity-agenda-in-his-first-state-of-the-union/.

68 "Executive Order on Promoting Competition in the American Economy," The White House Briefing Room, July 09, 2021, https://www.whitehouse.gov/briefing-room/presidential-actions/2021/07/09/executive-order-on-promoting-competition-in-the-american-economy/.

69 Filter Bubble Transparency Act, S. 2024, 117th Congress, https://www.congress.gov/bill/117th-congress/senate-bill/2024/text.

70 "Document: Filter Bubble Transparency Act Proposed in House," *Tech Policy Press*, November 9, 2021, https://techpolicy.press/document-filter-bubble-transparency-act-proposed-in-house/.

71 Justice Against Malicious Algorithms Act of 2021, H.R. 5596, 117th Congress, https://www.congress.gov/bill/117th-congress/house-bill/5596/text.

72 Algorithmic Justice and Online Platform Transparency Act, S. 1896, 117th Congress, https://www.congress.gov/bill/117th-congress/senate-bill/1896.

73 "Senator Markey, Rep. Matsui Introduce Legislation to Combat Harmful Algorithms and Create New Online Transparency Regime," Ed Markey, press release, May 27, 2021, https://www.markey.senate.gov/news/press-releases/senator-markey-rep-matsui-introduce-legislation-to-combat-harmful-algorithms-and-create-new-online-transparency-regime.

74 "Wyden, Booker and Clarke Introduce Algorithmic Accountability Act of 2022 to Require New Transparency and Accountability for Automated Decision Systems," Ron Wyden, press release, February 03, 2022, https://www.wyden.senate.gov/news/press-releases/wyden-booker-and-clarke-introduce-algorithmic-accountability-act-of-2022-to-require-new-transparency-and-accountability-for-automated-decision-systems.

automated systems are used, and empowers consumers to make informed choices about the automation of critical decisions."

- **Platform Accountability and Transparency Act.**[75] This working draft proposal was announced in December 2021 by Sens. Chris Coons (D-DE), Amy Klobuchar (D-MN), and Rob Portman (R-OH). The main focus of this bill proposal is mandating companies to share data with sanctioned third-party researchers (see "Third-Party Research and Auditing" below). Under this act, failing to comply with providing data access to qualified projects would result in losing the immunities provided by CDA 230.[76]

Bills Related to Privacy

- **American Data Privacy and Protection Act.**[77] This bipartisan discussion draft bill was released by leaders of the House Energy and Commerce Committee in June 2022. The bill seeks "to provide consumers with foundational data privacy rights, create strong oversight mechanisms, and establish meaningful enforcement," and it would establish a national standard for personal data collection and use.
- **S. 2134. Data Protection Act of 2021.**[78] Introduced in June 2021 by Sen. Kirsten Gillibrand (D-NY), this bill proposal seeks to create an independent Data Protection Agency for the regulation of high-risk data practices relating to personal data, including in automated, machine-learning decision systems, the profiling of individuals, and the processing of personally identifying biometric information.
- **S. 1494. Consumer Data Privacy and Security Act of 2021.**[79] Introduced in April 2021 by Sen. Jerry Moran (R-KS), this bill proposal seeks to set standards for the collection of personal data, including obtaining consent, publishing privacy policies, implementing data security programs, and providing data controls to the users.
- **H.R. 1816. Information Transparency & Personal Data Control Act.**[80] Introduced in March 2021 by Rep. Suzan DelBene (D-WA), this bill proposal requires the Federal Trade Commission to establish requirements for collection and handling of sensitive personal information, including obtaining affirmative consent from users, publishing a privacy and data use policy that is readily understandable, opt-out features, and periodic privacy audits.
- **S. 3663. The Kids Online Safety Act.**[81] Introduced in February 2022 by Sens. Richard Blumenthal (D-CT) and Marsha Blackburn (R-KY), this bipartisan bill proposal seeks to enhance parental control over children's online data and time spent online.

75 Platform Accountability and Transparency Act, S. 797, 117th Congress, December 9, 2021, https://www.coons.senate.gov/imo/media/doc/text_pata_117.pdf.

76 Corin Faife, "New Social Media Transparency Bill Would Force Facebook to Open Up to Researchers," *Verge*, December 10, 2021, https://www.theverge.com/2021/12/10/22827957/senators-bipartisan-pata-act-social-media-transparency-section-230.

77 American Data Privacy and Protection Act, 117th Congress, https://www.commerce.senate.gov/services/files/6CB3B500-3DB4-4FCC-BB15-9E6A52738B6C

78 Data Protection Act of 2021, S. 2134, 117th Congress, https://www.congress.gov/bill/117th-congress/senate-bill/2134/text.

79 Consumer Data Privacy and Security Act of 2021, S. 1494, 117th Congress, https://www.congress.gov/bill/117th-congress/senate-bill/1494.

80 Information Transparency & Personal Data Control Act, H.R. 1816, 117th Congress, https://www.congress.gov/bill/117th-congress/house-bill/1816.

81 Kids Online Safety Act, S. 3663, 117th Congress, https://www.congress.gov/bill/117th-congress/senate-bill/3663?s=1&r=10.

- **H.R. 4801. Protecting the Information of Our Vulnerable Children and Youth Act.**[82] Introduced in July 2021 by Rep. Kathy Castor (D-FL) and 26 cosponsors, this bill proposal would amend COPPA to increase the age threshold for the banning of personal data collection without parental consent to age 18, and it would extend the rules to all sites visited by minors. Additionally, it would ban targeted advertising on minor, and require companies to get consent from children aged 12–17 before collecting their data (in addition to parental consent for younger children).
- **S. 1628. Children and Teens' Online Privacy Protection Act.**[83] This bipartisan bill proposal, introduced in May 2021 by Sens. Edward Markey (D-MA) and Bill Cassidy (R-LA), would amend COPPA to ban targeted advertising directed at children.

Bills Related to Antitrust and Competition

- **S. 2992/H.R. 3816. American Innovation and Choice Online Act.**[84] Introduced in October 2021 by Sen. Amy Klobuchar (D-MN) and supported by a group of six Republicans and six Democrats. A similar bill proposal was introduced in the House in June 2021 by Rep. David Cicilline (D-CT). This bill treats Big Tech companies, such as Amazon, Google, and Facebook, like dominant, gatekeeping platforms with special responsibilities.[85] Similar to the EU's Digital Markets Act, this bill would make it illegal for large companies to advantage their own products and services at the expense of other businesses that rely on these platforms, to impede third-party business users to access and interoperate with the same technology tools available to the covered platform's own products and services, and to impede users from uninstalling pre-installed software and default features.
- **H.R. 3849. Augmenting Compatibility and Competition by Enabling Service Switching (ACCESS) Act of 2021.**[86] This bipartisan bill proposal would require large online platforms to facilitate consumers and businesses switching from one platform to another through secure data portability and cross-platform interoperability. The ACCESS Act is part of a six-bill antitrust package (which also includes a House version of the American Innovation and Choice Online Act described above) introduced by the House Antitrust Subcommittee and approved by the House Judiciary committee in June 2021.

European Union

The European Union's comprehensive proposals for the governance and regulation of digital services published in 2020 and 2021 are the most advanced in this area and have served as inspiration for similar

82 Protecting the Information of Our Vulnerable Children and Youth Act, H.R. 4801, 117th Congress, https://www.congress.gov/bill/117th-congress/house-bill/4801?s=1&r=7.

83 Children and Teens' Online Privacy Protection Act, S. 1628, 117th Congress, https://www.congress.gov/bill/117th-congress/senate-bill/1628.

84 American Choice and Innovation Online Act, H.R. 3816, 117th Congress, https://www.congress.gov/bill/117th-congress/senate-bill/2992.

85 Ryan Tracy and John McKinnon, "Antitrust Tech Bills Gain Bipartisan Momentum in Senate," *Wall Street Journal*, November 25, 2021, https://www.wsj.com/articles/antitrust-tech-bills-gain-bipartisan-momentum-in-senate-11637836202?mod=article_inline.

86 ACCESS Act of 2021, H.R. 3849, 117th Congress, https://www.congress.gov/bill/117th-congress/house-bill/3849.

regulatory action in the United States and worldwide. Especially as Big Tech's social media companies operate internationally, their compliance with the EU directives will likely extend to all users of these platforms regardless of geographic location.

EU's Digital Services Act (DSA) and Digital Markets Act (DMA). Released by the European Commission in December of 2020 and final versions announced in early 2022, the DSA and DMA[87] are parallel legal packages seeking to upgrade the governance of digital services.[88] While the DSA is mostly targeted at online platforms and seeks to protect the fundamental rights of all users of digital services, the DMA is targeted at a smaller subset of "gatekeeper platforms" that act as "bottlenecks" between businesses and users (including Big Tech companies like Google, Facebook, and Amazon) and seeks to establish a fair digital market to foster innovation, growth, and competitiveness.

The DSA requires online platforms to transparently inform users of their online advertising and algorithmic content recommendation practices, to let users turn off personalized recommendations, and show them any options to adjust the parameters used in ranking. The DSA also requires enhanced avenues for monitoring platforms' practices, including granting access to key platform data to qualified researchers.

The DMA requires proactive measures to ensure interoperability with third-party software without losing proper function of their services and prohibits unfair practices such as giving priority in visibility ranking to their own services over similar services offered by third-parties.

EU's Artificial Intelligence Act (AIA). Released by the European Commission in April of 2021, the AIA leverages a risk-based approach to regulate AI, categorizing AI systems from low or minimal risk—which can be used with no legal obligations—to unacceptable risk—whose use is prohibited.[89] Under its unacceptable risk category, the AIA includes "AI systems that deploy harmful manipulative 'subliminal techniques'." This language has been criticized as ambiguous, which might lead to overly strict restrictions.[90] For example, it is unclear whether social media content recommendation algorithms fall under this category, and if so, whether all flavors of these algorithmic practices can be considered as "subliminal techniques," or, for instance, only recommendations related to advertising.

87 Proposal for a Regulation of the European Parliament and of the Council on a Single Market for Digital Services (Digital Services Act) and Amending Directive 2000/31/EC, European Commission, December 15, 2020, https://digital-strategy.ec.europa.eu/en/library/proposal-regulation-european-parliament-and-council-single-market-digital-services-digital-services.

88 "Europe Fit for the Digital Age: Commission Proposes New Rules for Digital Platforms," European Commission Press Corner, December 15, 2020, https://ec.europa.eu/commission/presscorner/detail/en/ip_20_2347.

89 Tambiama Madiega, "EU Legislation in Progress: Artificial Intelligence Act," European Parliament Research Service briefing, January 14, 2022, https://www.europarl.europa.eu/thinktank/en/document/EPRS_BRI(2021)698792.

90 Josephine Wolff, "The EU's New Proposed Rules on A.I. Are Missing Something," *Slate*, April 30, 2021, https://slate.com/technology/2021/04/eu-proposed-rules-artificial-intelligence.html.

Self-Regulation and Private Oversight

Internal auditing and companies' "responsible algorithms" initiatives. In response to increased evidence of the downstream harms related to their platforms, companies have formed dedicated initiatives and teams for the study of the societal impact of their content recommendation algorithms. These include Facebook's Responsible AI[91] team and Twitter's Responsible Machine Learning initiative.[92] These teams are usually made up of individuals from across the company, including engineers, data scientists, and in some cases social scientists and ethicists. Their research is claimed to be used by the companies to combat downstream or unintentional harms and ultimately improve users' experience on their platforms.

However, some have expressed skepticism around such initiatives, arguing that these teams tend to be formed mostly to placate criticism in response to public outcries, and that their scope and impact is rather limited. For example, Facebook's Responsible AI team has been criticized[93] for focusing primarily on combating algorithmic bias (e.g., presenting certain job ads to white users but not to racial minorities, or disproportionately censoring conservative content) and not doing enough on other urgent and potentially much more prevalent problems like the spread of misinformation and hate speech. Further, although social media companies release periodic curated reports on algorithm performance and content moderation, the data underlying these reports are not shared, thus preventing independent assessments of their accuracy or complementary analyses.

Independent oversight models like Meta's Oversight Board. The Oversight Board was established in 2020 as an independent body to make content moderation decisions for Facebook and Instagram. It is composed of around forty individuals from a diverse set of disciplines and backgrounds, and its goal is to promote free expression "by making principled, independent decisions regarding content on Facebook and Instagram and by issuing recommendations on the relevant Facebook company content policy."[94] The Oversight Board currently only has power over appealed content removal decisions, and while its decisions regarding the removal of content are binding, its wider policy recommendations are not.

Third-Party Research

Independent research by academics and journalists has been key in exposing problematic data and algorithmic practices by digital technology companies, including topics related to personalized advertising

91 "Facebook's Five Pillars of Responsible AI," Meta AI blog, June 22, 2021, https://ai.facebook.com/blog/facebooks-five-pillars-of-responsible-ai/.

92 Jutta Williams and Rumman Chowdhury, "Introducing our Responsible Machine Learning Initiative," Twitter blog, April 14, 2021, https://blog.twitter.com/en_us/topics/company/2021/introducing-responsible-machine-learning-initiative.

93 Hao, "How Facebook got addicted."

94 Meta Oversight Board, https://oversightboard.com/.

and misinformation.[95] Companies share some data through their APIs (application programming interface) or via direct agreements with researcher consortia, but these projects and the data shared by companies are usually very limited. Facebook's data tool, CrowdTangle, stands out in how much data access they provide to researchers compared to much more limited transparency efforts from companies like YouTube and TikTok. However, CrowdTangle has also been criticized for giving incomplete information on what content is popular on the platform,[96] and the company recently admitted[97] that it had shared a flawed dataset with a consortium of social scientists that had already used the data in dozens of published papers on social media's effect on elections.

Journalists and academic researchers have also devised ways to obtain data related to content without the companies' explicit collaboration, including so-called data-donation approaches. For example, the Ad Observatory project at New York University and various platform research initiatives by the Mozilla Foundation deploy browser plug-in extensions to scrape platform data from consenting users. These approaches expose researchers to civil and criminal liability via the 1984 Computer Fraud and Abuse Act and the companies' Terms of Service. For example, Facebook sent Cease and Desist letters and eventually suspended the accounts of researchers behind the Ad Observatory.[98]

Enhancing external researchers' access to hard data might enrich our understanding of social media companies' algorithmic practices and impact, and it would consequently inform all legislative efforts related to the regulation of social media. Recent efforts to give qualified independent researchers secure access to platform data include the EU's DSA and, in the United States, the Platform Accountability and Transparency Act[99] (currently in draft form).

KEY INSIGHT

Providing external researchers greater access to social media platform data will be important to widen our understanding of these platforms' algorithmic practices and impacts and inform appropriate regulatory approaches.

95 Elizabeth Hansen Shapiro, Michael Sugarman, Fernando Bermejo, and Ethan Zuckerman, "New Approaches to Platform Data Research," NetGain Partnership, February 2021, https://drive.google.com/file/d/1bPsMbaBXAROUYVesaN3dCtfaZpXZgI0x/view.

96 Kevin Roose, "Here's a Look Inside Facebook's Data Wars," *New York Times*, July 14, 2021, https://www.nytimes.com/2021/07/14/technology/facebook-data.html.

97 Craig Timberg, "Facebook Made Big Mistake in Data It Provided to Researchers, Undermining Academic Work," *Washington Post,* September 10, 2021, https://www.washingtonpost.com/technology/2021/09/10/facebook-error-data-social-scientists/.

98 Meghan Bobrowsky, "Facebook Disables Access for NYU Research into Political-Ad Targeting," *Wall Street Journal*, August 04, 2021. https://www.wsj.com/articles/facebook-cuts-off-access-for-nyu-research-into-political-ad-targeting-11628052204.

99 Platform Accountability and Transparency Act.

Selected Readings

The list below highlights some of the citations in this document and is not meant to be exhaustive.

On the Technology

- Covington, Paul, Jay Adams, and Emre Sargin. "Deep Neural Networks for YouTube Recommendations." In *Proceedings of the 10th ACM Conference on Recommender Systems*, 191–8. New York: Association for Computing Machinery, 2016. https://doi.org/10.1145/2959100.2959190
  - Although it goes into some technical detail, this paper provides an accessible overview of the main components and data types involved in YouTube's recommendation system (which is similar to those used by other large social media companies).
- Lada, Akos, Meihong Wang, and Tak Yan. "How Does News Feed Predict What You Want to See? Personalized Ranking with Machine Learning." Tech at Meta blog, January 16, 2021. https://tech.fb.com/news-feed-ranking/
  - This public blog post by Meta describes in lay terms the algorithmic processes and data used by Facebook's News Feed to select and curate content for users.

On Public Purpose Considerations

- Stray, Jonathan, Ivan Vendrov, Jeremy Nixon, Steven Adler, and Dylan Hadfield-Menell. "What Are You Optimizing For? Aligning Recommender Systems with Human Values." *arXiv:2107.10939* (2021). https://arxiv.org/abs/2107.10939
  - This paper presents case studies of recommender systems that have been modified to align with various human values such as fairness, well-being, and factual accuracy. The authors discuss the difficulties of "value engineering" and offer recommendations for future work in this area, including crafting useful measures of value alignment, adopting participatory design and operation, and designing for informed and deliberate judgments.

On Regulatory Approaches

- Brannon, Valerie. "Regulating Big Tech: Legal Implications." Legal Sidebar, Congressional Research Service, September 11, 2019. https://crsreports.congress.gov/product/pdf/LSB/LSB10309
  - This document provides an overview of the current regulatory landscape governing Big Tech companies, most of which applies to social media companies more broadly. It reviews and discusses relevant existing legislation and the main proposals for new legislation that are being considered in Congress related to speech, privacy, antitrust, and other policy issues.
- Keller, Daphne. "Amplification and Its Discontents: Why Regulating the Reach of Online Content Is Hard." *Journal of Free Speech Law* 1 (2021): 227–268.
  - Keller, a lawyer with expertise in platform regulation and internet users' rights, and who has testified in Congress, discusses the legal hurdles associated with policy proposals that seek to hold social media companies liable for the content they amplify through their algorithms. She suggests that content-neutral approaches to platform regulation might be more viable, including those grounded in privacy or competition law.

About the Technology and Public Purpose (TAPP) Project

The arc of innovative progress has reached an inflection point. It is our responsibility to ensure it bends toward public good.

Technological change has brought immeasurable benefits to billions through improved health, productivity, and convenience. Yet as recent events have shown, unless we actively manage their risks to society, new technologies may also bring unforeseen destructive consequences.

Making technological change positive for all is the critical challenge of our time. We ourselves—not only the logic of discovery and market forces—must manage it. To create a future where technology serves humanity as a whole and where public purpose drives innovation, we need a new approach.

Founded by Belfer Center Director, MIT Innovation Fellow, and former U.S. Secretary of Defense Ash Carter, the TAPP Project works to ensure that emerging technologies are developed and managed in ways that serve the overall public good.

TAPP Project Principles:

1. Technology's advance is inevitable, and it often brings with it much progress for some. Yet, progress for all is not guaranteed. We have an obligation to foresee the dilemmas presented by emerging technology and to generate solutions to them.
2. There is no silver bullet; effective solutions to technology-induced public dilemmas require a mix of government regulation and tech-sector self-governance. The right mix can only result from strong and trusted linkages between the tech sector and government.
3. Ensuring a future where public purpose drives innovation requires the next generation of tech leaders to act; we must train and inspire them to implement sustainable solutions and carry the torch.

For more information, visit: www.belfercenter.org/TAPP