COALITION TO FIGHT
DIGITAL
DECEPTION

Trained for Deception: How Artificial Intelligence Fuels Online Disinformation

A report from the Coalition to Fight Digital Deception

Authored by: The Anti-Defamation League, Avaaz, Decode Democracy, Mozilla and New America's Open Technology Institute

September 2021

Table of Contents

Executive Summary

Social media platforms are increasingly relying on artificial intelligence (AI) and machine learning (ML)-based tools to moderate and curate organic content online, and target and deliver advertisements. Many of these tools are designed to maximize engagement, which means they also have the potential to amplify sensationalist and harmful content such as misinformation and disinformation. This memo explores how AI and ML-based tools used for ad-targeting and delivery, content moderation, and content ranking and recommendation systems are spreading and amplifying misinformation and disinformation online.

It also outlines existing legislative proposals in the United States and in the European Union that aim to tackle these issues. It concludes with recommendations for how internet platforms and policymakers can better address the algorithmic amplification of misleading information online. These include: encouraging platforms to provide greater transparency around their policies, processes, and impact; directing more resources towards improving fact-checking, moderation efforts, and the development of effective AI and ML-based tools; providing users with access to more robust controls; and providing researchers with access to meaningful data and robust tools. Although platforms have made some progress in implementing such measures, we as a coalition believe that they can do more to meaningfully and effectively combat the spread of misinformation and disinformation online. However, recognizing the financial incentives underlying platforms' advertisement-driven business models—and the impact they have on platform approaches to misinformation and disinformation—we encourage lawmakers to pursue appropriate legislation and policies in order to promote greater transparency and accountability around online efforts to combat misleading information.

Introduction

Over the past several years, internet platforms have begun to develop and deploy a range of tools fueled by artificial intelligence (AI) and machine learning (ML) to shape and curate the content we see online. AI can be understood as machines that predict, automate, and optimize tasks in a manner that mimics human intelligence, while ML algorithms, a subset of AI, use statistics to identify patterns in data. Today, internet platforms use AI and ML tools to moderate and rank the content we see online, determine the items that are recommended to us, and the advertisements we see online. In order to deliver precise and highly personalized results to users, these tools rely on the vast collection of user data, including behavioral and location data.

Many of **these algorithmic tools are designed to maximize signals such as "engagement" and "relevance,"** and platforms often assert that by delivering "relevant" and personalized content to users, they are increasing the quality of the user experience. However, it is also important to recognize that by maximizing "relevant" and "engaging" content, companies are able to collect more user data, retain user attention, deliver more ads to users, and therefore earn more revenue. In addition, terms commonly used by platforms such as "relevance" and "quality" are often subjective and dependent on platforms' own definitions.

By focusing on "engagement" and delivering so-called "relevant" and "quality" content, platforms can also amplify online and offline harms. **Many types of harmful content, including hate speech and violent content, score higher engagement rates,** a fact Mark Zuckerberg himself highlighted. As a result, these algorithmic tools can amplify harmful content such as misinformation (verifiably false or misleading information with the potential to cause public harm—for example, by undermining democracy or public health, or encouraging discrimination or hate speech) and disinformation (verifiably false or misleading information that is spread with an intent to mislead or deceive). Conversations around how platforms tackle such falsehoods have particularly gained traction over the past year and a half, as misinformation and disinformation related to COVID-19 and the U.S. election have rapidly spread online.

This memo explores how AI and ML-based tools used for ad-targeting and delivery, content moderation, and content ranking and recommendation can spread and amplify misinformation and disinformation online. This memo also outlines existing legislative proposals in the United States and in the European Union that aim to tackle these issues. It concludes with recommendations for how internet platforms and policymakers can better address the algorithmic amplification of misleading information online.

Ad-Targeting and Delivery

Over the past two decades, the rapid rise and adoption of targeted advertising has radically transformed the internet ecosystem. Targeted advertising practices rely on the vast collection and monetization of internet users' personal data. Using this data, online advertisers are able to narrowly select and segment audiences based on their interests, behavioral information, demographic categories, personally identifiable information (PII), and more. As a result, advertisers are able to reach their target audiences with precision and scale, and garner more user attention. These incentives generate a vicious cycle of personal data collection: The more personal user data companies can collect, the more "relevant" advertisements they are able to deliver to users, therefore driving revenue on the platform. Due to the vast amount of data collection and subsequent microtargeting of ads that occurs online, the digital targeted advertising industry has recently been termed "surveillance advertising."

The targeted advertising industry has become a key component of the internet ecosystem, and ad-targeting and delivery practices are widely used by a number of internet platforms. According to recent statistics, approximately $356 billion was spent on digital advertising in 2020. This figure is projected to increase to $460 billion by 2024. Despite the fact that many internet companies use digital advertising tools, three companies—Google, Facebook, and Amazon—dominate the online ad market today. Estimates from late 2020 indicated that these three companies would make up approximately two-thirds of total U.S. digital ad spending that year, and that their market share of the online ad industry would continue to grow. Despite the impacts of the COVID-19 pandemic, the companies still have a triopoly over the digital advertising market.

Today, targeted advertising practices have become a critical element of the business models of most technology companies, influencing the way social media companies

operate their platforms. As the targeted-advertising ecosystem has become a lucrative option for generating revenue, many companies have introduced targeting and delivery tools that rely on AI and ML to enhance and scale their ad operations. These automated tools are interwoven throughout the ad-targeting and delivery process, and the exact role that these tools play varies from platform to platform. Generally, however, internet platforms rely on automated tools to make recommendations on targeting user 'categories' to advertisers. These tools can make such recommendations based on a range of data points, including what activities or items users have explicitly or implicitly demonstrated interest in. Research has indicated, however, that at times, **these tools can suggest categories of users to advertisers in ways that reflect societal biases and exacerbate discriminatory and harmful practices**.

Internet platforms also rely on automated tools to shape which ads are delivered to a user and when. Generally, advertisers need to place a bid and participate in an ad auction before their ad will be delivered to a user. The methodology that determines whether an ad is eventually delivered to a user or not is platform dependent, and, in many cases, AI and ML tools play a role in determining the outcome of an auction. For example, on Facebook, a ML model is used to predict the "quality" of an ad, which is one factor that is considered during the ad auction and delivery process. The quality of an ad is based on numerous data points, including feedback from users who view or hide ads, and Facebook's assessments of low-quality features in an ad, such as too much text in an image.

As research has outlined, the algorithms that are used to power the ad-delivery process can generate insights that result in an ad being delivered to an audience segment that is different from the target audience outlined by the advertiser. This is because the automated tool predicts that an ad will be more relevant for certain audience categories. However, **this has resulted in discriminatory outcomes for protected groups when delivering ads related to housing, employment, and credit**. For example, an ad-delivery algorithm may only deliver ads for traditionally-male dominated careers, such as doctors or engineers, to male job seekers. As a result, women could be excluded from seeing these opportunities without regard to their qualifications. This is because the ad-delivery algorithm bases its optimization strategy on data about a given user in conjunction with current and historical data on job seekers, which may reflect gender discrimination in these career fields. This is an area where policymakers can help address the harms generated by AI and ML-based tools, as they can clarify via legislation or other methods that offline anti-discrimination statutes, such as the Civil Rights Act of 1964 and the Fair Housing Act, apply in the digital environment.

Internet platforms also use automated tools for a range of other purposes during the ad-targeting and delivery process. These include identifying which subset of users are

more likely to react to an advertiser's ads, tailoring the creative elements of an ad for distinct audiences (dynamic creative optimization), and tracking engagement metrics during the ad delivery process.

Since the 2016 U.S. presidential election, a significant amount of research has been conducted on the role online political advertisements play in spreading misleading information, and how AI and ML-based ad-targeting and delivery tools can amplify such messages.

For example, in October 2019, the Trump campaign shared an ad on Facebook attacking President Biden's record on Ukraine using debunked conspiracy theory claims. The ad was viewed millions of times, and despite repeated requests from the Biden presidential campaign, Facebook refused to remove the ad, arguing the company should not be an arbiter of truth.

The company allows false claims in ads directly from politicians, though it does appear to fact check content from interest groups. **This policy has allowed false information to circulate and to be precisely targeted and delivered via political advertising on the service.** In addition, because political advertisers have access to robust algorithmic targeting tools, they can precisely target users based on their interests and behaviors. As a result, these advertisers can easily engage with users who are more susceptible to believing certain false narratives.

Platforms have varied in how they have addressed misinformation and disinformation in political advertising. Many companies have broadened or changed their political advertising rules over the past several years. However, these rules do not always go far enough and are often difficult to understand and access. Some companies, such as Twitter, LinkedIn, Pinterest, and TikTok, have opted to ban all political advertising in order to address the spread of misleading content through ads. Additionally, in the run up to—and in the weeks and months following—the 2020 U.S. presidential election, Facebook and Google both imposed temporary bans on political advertising. However, it can take time for automated and human systems to adapt to new parameters and rules. Further, some companies, such as Facebook require advertisers to self-categorize their ads as political ads—a process that can easily be evaded. As a result, political ads, including ads containing misleading information, can still slip through the cracks.

Additionally, while many have lauded platforms' decisions to temporarily or permanently ban political advertising, it is important to recognize that the definition of political advertising is not fixed, and paid political content can still appear on these services. For example, politicians and political groups have partnered with TikTok influencers in order to

promote their ideas and gain traction with certain audience segments, thus side-stepping overt political advertising.

Misleading information has also spread in other categories of advertising and is particularly apparent in ads related to the COVID-19 pandemic. Numerous online ads claiming to sell verified prevention tools and cures for the coronavirus have circulated during the pandemic. Research has indicated that communities of color and other marginalized groups may be especially susceptible to such campaigns.

Many internet platforms have changed or expanded their advertising policies in order to address the rise of COVID-19 misinformation and disinformation on their services. In the early days of the pandemic, Facebook prohibited advertisements for products claiming to prevent or treat the coronavirus. The company also temporarily banned ads and commerce listings for medical face masks, hand sanitizer, surface disinfecting wipes, and COVID-19 testing kits. Many other social media and commerce platforms took a similar approach. However, as previously noted, it can take time for automated and human systems to adapt to these parameters, so ads with misleading information could still circulate online.

Although companies have introduced numerous changes to their advertising policies and practices over the past several years, they still do not make meaningful disclosures around how these systems operate and what impact they have. Companies such as Facebook, Google, and Reddit have responded to calls for greater transparency by publishing ad transparency libraries or ad transparency reports. **However, platforms themselves decide how these ad libraries are structured and what ads are included in them**. For example, Google publishes a political ad transparency report which provides data on impressions, targeting criteria, and other factors about political ads in the United States, Australia, and a handful of other countries and regions. However, the report only includes ads that feature a "current officeholder or candidate for an elected federal or state office, federal or state political party, or state ballot measure, initiative, or proposition that qualifies for the ballot in a state." As a result, the report does not provide a comprehensive overview of all political ads that are run on the platform, and there are a plethora of ads—including ones that could contain misleading information—that are not available for public scrutiny. Facebook and Reddit's ad libraries also have similar flaws, (as illustrated by NYU's Ad Observer project, which was blocked by Facebook) therefore limiting their value as transparency and accountability mechanisms.

In addition, if a platform or advertiser fails to accurately categorize an ad as a political ad, then it may not be entered into the platform's ad library and will not be visible to researchers. Currently, researchers have no way of verifying if a platform has made mistakes when reviewing or categorizing ads. This is because **internet platforms publish little to no comprehensive data around how they enforce their advertising content and**

targeting policies, how many ads they have removed for violating these policies, and how many enforcement mistakes they have made, including by erroneously allowing ads in violation of their policies to run. Reddit is the only company that publishes any data in this regard, as it shares details on ads it approved in error in its political ads transparency subreddit. However, there is still little information around how platforms enforce their ad policies related to misleading information and what impact this has on the state of misinformation and disinformation on their services. Further, many platforms do not provide researchers with access to useful ad APIs, which makes conducting meaningful research unnecessarily difficult.

As the harms caused by the ad-targeting and delivery systems have become more apparent, numerous transparency advocates have set forth proposals to address the underlying problematic targeted-advertising business model. For example, Accountable Tech, a policy-focused nonprofit organization, recently launched a campaign calling for a ban on surveillance advertising. The campaign has garnered the support of 42 organizations. However, some grassroots and political organizations have not voiced support for this effort because targeted advertising also serves as a lifeblood for many such groups that are seeking to engage with certain constituencies—a fact that demonstrates the extent to which these systems are so deeply entrenched. In addition to calls from advocacy groups, some legislative efforts in the United States and the European Union have also sought to address the harms caused by algorithmic ad-targeting and delivery systems through algorithmic audits and impact assessments (discussed in the *Relevant Legislation* section).

Content Moderation

Today, most social media companies engage in content moderation to enforce their content policies, which determine what content, individuals, and groups are permitted on their services. While overbroad content moderation raises freedom of expression concerns, content moderation is important for addressing misinformation, disinformation, harassment, and racist or hateful content online. Generally, companies rely on a combination of human moderators and AI and ML-based tools to carry out their content moderation efforts, which include flagging, reviewing, and making determinations about content.

Companies typically deploy AI and ML-based tools during two stages of the content moderation process: pre-moderation and post-moderation. During the pre-moderation stage, content is reviewed before it is published on a platform. In this situation, if a user drafted a post containing misinformation or disinformation, a company's content moderation tools could flag the content as a violation and prevent the user from publishing the post. However, this approach is best deployed when categories of content are clearly defined. This is because automated tools are often unable to accurately determine context and make subjective decisions, and so the risk of error and overbroad content removals increases when it comes to categories of content that have fluid definitions. As a result, pre-moderation of misleading content is less common.

During the post-moderation stage, companies moderate content that has already been published on a platform. For example, if a user shared a post containing misinformation, and another user or the company's automated tools flagged the content as potentially violating the company's content policies, the post could then be routed to a human moderator for review—or it could be automatically removed, depending on how automated content moderation tools are used.

Using AI and ML-based tools to identify and remove misinformation and disinformation online has certain benefits for both social media companies and consumers of these companies' products and services. The first and most obvious benefit is efficiency. With billions of pieces of content surfacing on platforms, technology companies can use automated tools to scan content and make determinations on the nature of the content and whether it should be permitted on a service at scale. Another benefit of using AI and ML-based content moderation tools is that doing so allows companies to offload content moderation work from human moderators to automated systems. This can allow human content moderators to navigate a more manageable volume of content and protect them from gratuitous exposure to harmful content. As research and recent reporting have indicated, human moderators often experience adverse consequences from reviewing vast amounts of harmful content online, including mental health concerns and themselves becoming more susceptible to conspiracy theories.

However, if bias is incorporated into the design AI and ML-based content moderation tools, they can amplify harmful content and generate discriminatory outcomes. Automated systems learn to make decisions based on training data. However, training data is often representative of societal and institutional inequities, and datasets can be influenced by human prejudices. This can result in bias against certain communities or forms of content within an algorithmic system. There are numerous examples of internet platforms relying on biased AI and ML-based tools to make content moderation decisions that have resulted in harmful and discriminatory outcomes offline. For example, throughout 2020, Instagram's content moderation system flagged and removed the hashtag #Sikh from its platform for an extended period of time. While Instagram alleged that the block was a mistake due to a report "inaccurately reviewed" by Instagram's teams, large-scale implementation of this decision was ultimately fueled by AI. This is discriminatory, problematic, and perpetuates systemic racism by erasing voices of religious minorities, particularly as the incident occurred in the midst of protests in support of farmers in India, many of whom are Sikh. Instances such as these outline why it is so critical for companies to provide adequate notice to users who have had their content or accounts impacted by moderation, and to give these users access to a timely, scalable, and robust appeals process.

Another example of biased AI and ML-based tools producing harmful outcomes occurred in 2017, when Microsoft released a "teenage" chat bot that used AI, which was inadvertently programmed to shut down any conversation about religious identity or the Middle East. For example, if a human told the bot "I get bullied sometimes for being Muslim," the bot would respond "so I really have no interest in chatting about religion," or "For the last time, pls stop talking politics..it's getting super old." The bot gave a similar

response when the words “Jew,” “Middle East,” “Hijabs,” and “Torah” were used, although it would not provide a similar response when a user discussed Christianity.

Additionally, social media companies’ content moderation efforts are also limited in that platforms have consistently had unequal content moderation support and capabilities for non-English content. Facebook, for example, claims to be available in over 100 languages; however, its content moderators speak only about 50 languages, and Facebook’s automatic moderation tools are only able to flag hate speech in about 30 languages. Fewer human moderators (and for some languages no human moderators) means there is less accurate training data for ML systems used to detect disinformation in non-English languages.

This is a self-perpetuating problem: non-English violative content is less likely to be seen by human moderators for review; therefore, **someone who uses a social media platform in a language other than English may be more likely to be exposed to harmful disinformation.**

This issue was clearly illustrated during the recent U.S. presidential election when APIAVote, a nonpartisan organization that mobilizes AAPI individuals in electoral and civic participation, expressed serious concern about how voters with limited English proficiency were vulnerable to voting disinformation spread on social media. In one article, APIAVote told Vox, “It appeared that certain communities were more vulnerable and targeted, with the information translated into their language and posted onto WeChat or Facebook.” Similarly, in 2021 the *Ya Basta! Facebook* coalition was formed after several nonprofit organizations discovered rampant disinformation campaigns targeting Latinx communities in the U.S. There are approximately 41 million Spanish speakers in the U.S. and millions consume Spanish language content on social media; however, several reports have found that social media platforms are falling short in addressing online disinformation campaigns that target Latinx communities. This is a gap that those companies must work to address.

Many automated content moderation tools are also limited in their ability to effectively moderate certain categories of content. Theoretically, if an automated content moderation tool is operating using clear definitions for a category of content and it is trained on a diverse and robust enough dataset, it should be able to flag violating content easily and more effectively than a human moderator. This is because these automated systems are easily scalable, meaning that once an AI system is trained, it can be duplicated. In comparison, training new human content moderators can be difficult, costly, and time intensive. However, the vast majority of automated content moderation tools are deployed against categories of content that have fluid definitions, such as misinformation, disinformation, hate, and extremism. These categories of content often require context and subjective understanding in order to determine the meaning of a word, image, or

video, without relying on specific terms (for moderation of words or phrases) or hashes (for moderation of images and videos). Additionally, companies often change the parameters for these categories of content in response to real-world events. As a result, the effectiveness and accuracy of these systems is limited, as they can fail to flag and remove violating content or erroneously take action against content or accounts that do not violate a platform's policies.

The consequences of these limitations have been profound, often further marginalizing already vulnerable populations. On numerous platforms, disinformation about the November 2020 U.S. elections circulated widely, with bad-faith actors posting inaccurate information about voting logistics, hindering individuals' ability to vote. Disinformation campaigns also cast doubt on election security in the United States, with claims about tampered votes and a stolen election rapidly gaining steam online. When companies fail to moderate misleading information in this context, it can prevent people from participating or trusting in the U.S. election process. This has a grave effect on democracy and reinforces systems of oppression. In order to help combat this, policymakers should clarify that offline anti-discrimination statutes, such as the Voting Rights Act, apply in the digital environment.

The limitations of automated content moderation tools were also especially visible in the wake of the COVID-19 pandemic. In March 2020, many companies, including Facebook, were unable to initially make use of their large human content moderator workforces due to the COVID-19 work-from-home requirements. Instead, many platforms increased their reliance on AI and ML-based tools for content moderation purposes. As a result, more content was flagged and removed than before, including posts featuring legitimate news articles about the pandemic, which were incorrectly flagged as spam. Simultaneously, numerous posts containing misleading information slipped through the cracks and continued to circulate online, including conspiracy theories about COVID-19's existence, COVID-19 vaccines, testing, and symptoms.

When users share misleading information, it can establish echo chambers online, which can amplify and fuel extremism and hate. Additionally, since headlines around COVID-19, including disinformation, have dominated mainstream and social media for over a year, there has been a significant spike in hate and racism directed at the Asian American and Pacific Islander (AAPI) community—both offline and online. For example, in the days immediately following then-President Trump's COVID-19 diagnosis, there was a significant spike in anti-Asian sentiment and conspiracy theories about COVID-19 on Twitter. Asian-Americans have experienced the largest single rise in severe online hate and harassment year-over-year in comparison to other groups. In this way, limited and

flawed content moderation systems can exacerbate hate by amplifying misinformation and disinformation.

Although some platforms partner with third-party fact-checkers to identify potentially misleading content, the scale of these fact-checking efforts often varies. In addition, some platforms remove content that has been fact-checked and deemed to be misleading, while others opt to algorithmically reduce or label it. While these efforts could be helpful in combating the spread of misleading information **there is a great deal of inconsistency in how fact-checking and alternative moderation techniques are applied, and there is a fundamental lack of transparency around what policies guide the implementation of these practices.** This makes it difficult to monitor how platforms are combating misleading information online and hold them accountable for these efforts.

As content moderation tactics become more complex, some bad actors aiming to spread disinformation have also identified ways to circumvent both pre-moderation and post-moderation practices. For example, some individuals attempt to evade automated content moderation systems and obfuscate their messages by typing "C0v1D" instead of "Covid," In response to such evasion efforts, many companies have invested significant resources in training their AI and ML-based systems to identify altered and duplicated versions of text, images, and other forms of communication, with the aim of augmenting their misinformation and disinformation moderation efforts—although as these bad actors adapt their evasion efforts, content may still slip through the cracks. While some platforms currently publish transparency reports outlining the scope and scale of their content policy enforcement efforts, very few platforms publish data around their efforts to moderate misleading content. Some platforms share this data, but in a disparate manner, which renders it hard to track and find, making it difficult to hold these platforms accountable.

Given the limitations of AI and ML-based content moderation tools, the best form of online content moderation is a combination of AI and human review. Decreased human oversight increases the risk of errors from automated systems, which can result in the amplification of hate, extremism, systemic biases, discrimination, and misleading information. For these reasons, it is essential that social media companies invest in their content moderation systems, increase resources for both human and AI content moderation, and work to decrease the harmful impact of biased AI systems to meet the goal of reducing disinformation on social media platforms. In addition, given that there is often a spike in misleading information surrounding major world events, companies should invest more resources in preparing for important world events could result in the spread of more misleading information.

Ranking and Recommendation Systems

As previously noted, today, the majority of social media platforms rely on the use of algorithmic systems which personalize user experiences and can predict and optimize for user engagement. Companies can curate the content users see in many ways, including by algorithmically ranking content (such as in a News Feed) or by algorithmically recommending content to users (such as suggesting which video to watch next). In many cases, ranking and recommendation algorithms work in tandem to provide users with a curated and personalized experience. Both ranking and recommendation algorithms are designed to consider a plethora of signals, many of which are informed by implicit and explicit user behaviors. Companies assert that these tools promote “relevant” and “useful” content to users, but they also allow platforms to maximize user attention and ad revenue.

Typically, social media platforms will rank content in one of two ways. Either the company will rank content on a News Feed or similar feature based on user behavior, or the platform will rank search results generated by a search engine that receives direct user input, such as the Google search engine or the YouTube search feature. Twitter, for example, uses a model that "predicts how interesting and engaging a Tweet would be” in order to determine how posts are ranked on the Twitter timeline. Similarly, Facebook uses a system that "determines which posts show up in your News Feed, and in what order, by predicting what you’re most likely to be interested in or engage with.” YouTube has also used systems that rely on clicks, watch time, and surveys in order to curate and present content.

Internet platforms also deploy different types of recommendation systems. These include content-based systems (which suggest items to a user that are similar to items they have

previously shown interest in), collaborative-filtering systems (which suggest items to a user by assessing the interests and behaviors of users who have similar interests), and knowledge-based systems (which suggest items to a user by evaluating a user's interests and characteristics, in addition to the characteristics of an item). Most companies use a combination of these systems to drive their recommendations.

While ranking and recommendation systems may deliver "relevant" content to users, the fact that many of these systems are designed to optimize for engagement means that harmful content—such as misinformation, disinformation, hate speech, and graphic and violent content, which are often more engaging—are also amplified. In this way, both ranking and recommendation algorithms can profoundly shape a user's online experience and determine what kind of information they engage with.

For example, if a ranking algorithm was designed to weigh engagement heavily in its decision-making, it could algorithmically amplify potentially harmful content by ranking it higher in a user's News Feed. Similarly, the algorithm could downrank content that is less engaging in nature, even if the content contains reliable information. As a result, users would be less likely to view this legitimate content.

Although ranking algorithms can amplify harmful and misleading content, many platforms have begun using algorithmic ranking techniques to reduce the spread of misleading content, especially during the COVID-19 pandemic and in the months leading to the 2020 U.S. presidential election. For example, when a piece of content on Facebook is fact-checked and debunked by one of the company's fact-checking partners, the company generally appends a warning label to the content and reduces its distribution by algorithmically downranking the post in users' News Feeds. However, methods such as downranking still enable misleading content to remain online, which means users can still access and share this content. According to 2018 findings published in the leading academic journal *Science*, false information was 70% more likely to be retweeted than the truth, so **algorithmic downranking alone may not be sufficient to stop the spread of harmful and misleading information.**

Companies such as Facebook have explored other methods for combating misleading information online, such as labeling posts that have been debunked or that contain information on topics that are commonly the focus of misinformation and disinformation campaigns, such as COVID-19. These practices are often deployed in tandem with algorithmic downranking methods. Last year, Facebook shared that it had labeled 167 million user posts for featuring information about the coronavirus that had been debunked by its fact-checking partners. However, earlier this year, the company announced it would append labels that link to additional information to all posts related to COVID-19 vaccines. While some lauded this move, others noted the broad scale application of such labels

would render them meaningless. Therefore, some of the additional methods that platforms have deployed alongside algorithmic downranking are also limited. Further, there is currently little transparency around how downranking and other similar practices are applied, which makes it difficult to monitor whether they are consistently enforced. Most internet platforms have also not provided meaningful data indicating that these efforts are effective.

Like ranking algorithms, recommendation algorithms designed to emphasize engagement can also amplify harmful and misleading content. This has been especially visible on YouTube. Over the past several years, numerous efforts—from *The New York Times* technology columnist Kevin Roose's Rabbit Hole series to Mozilla's YouTube Regrets project—have demonstrated how the platform's recommendation system can drive people towards extreme and harmful videos.

YouTube's shortcomings were especially pronounced in the weeks following the 2020 U.S. election, when former President Trump's false claim that he had won gained major traction on the platform. A 2021 study by Pendulum shows that "14,000 YouTube videos, which accounted for 820 million views, supported President Trump's false claims of widespread voter fraud." Although YouTube offers little transparency into its recommendation system, there's little doubt that hundreds of millions of viewers wouldn't have stumbled onto these videos on their own without some nudge from the algorithm—as YouTube itself has stated, 70% of watch time on the platform is driven by video suggestions made by its algorithm. In 2019, Mozilla called on YouTube to provide researchers with access to meaningful data, better simulation tools, and tools that empower research and analysis, so that researchers could better understand how recommendations impact the online experience and can amplify harmful content on the platform. The company has not implemented any of these recommendations, despite mounting pressure.

Algorithmic recommendations of "groups" for users to join have also raised numerous concerns over the past several years, as platforms such as Facebook have suggested groups that have been penalized on the platform for amplifying harmful content, including misleading information and conspiracy theories. Earlier this year, Facebook announced its plans to remove civic and political groups, as well as newly-created groups, from recommendations worldwide. It had already stopped recommending such groups in the United States in January. The company also shared it would restrict the reach of groups that violated the platform's Community Standards. This came more than five years after Facebook employees first raised the alarm about the dangers of group recommendations, and more than five months since organizations such as Mozilla and Accountable Tech, along with thousands of internet users, called on Facebook to stop amplifying election disinformation by pausing group recommendations in the United States through Inauguration Day in 2021. However, a recent study from The Markup, shows that despite

Facebook's alleged policy change, the company has continued to recommend political groups.

Like algorithmic ad-targeting and delivery systems, algorithmic ranking and recommendation systems rely on the collection and use of vast amounts of user data. Given the harms these systems can cause, **users should have access to more robust privacy protection and user controls.** In particular, users should be able to determine how their personal data is collected and used by algorithmic systems, and what kind of content they are served by these systems. For example, users should be able to control how their personal data is used to inform the video recommendations they receive, and be able to opt-out of being recommended certain categories of videos, such as political videos, if they so desire.

There are currently a number of research initiatives and policy proposals seeking to limit or contain the harmful impacts of algorithmic ranking systems. One approach centers on generating product friction, or anything that inhibits user action within a digital interface. On the front-end, a company could generate friction by encouraging users to be more thoughtful before sharing or posting content. On the back-end, a company could alter the algorithms that deliver content to users by, for example, requiring content that reaches a certain threshold of circulation to be reviewed before it can continue being ranked highly in News Feeds. If a platform is recommending too much content, the company could also require more specificity in search algorithms, so as to not predict user preferences. Other proposed approaches include pursuing personalization systems that optimize for both user engagement and consumption diversity, so that users are consuming a broader range of content on a particular platform, promoting more understandable and accessible user controls and information that allow users to understand how adjusting controls would alter what they see, developing survey-based measures to refine content selection, and recommending feeds to users rather than items. In the disinformation context, researchers have especially recommended designing algorithms to deprioritize content known to be inaccurate and untrustworthy (as was done with news sources during the 2020 U.S. election transition period).

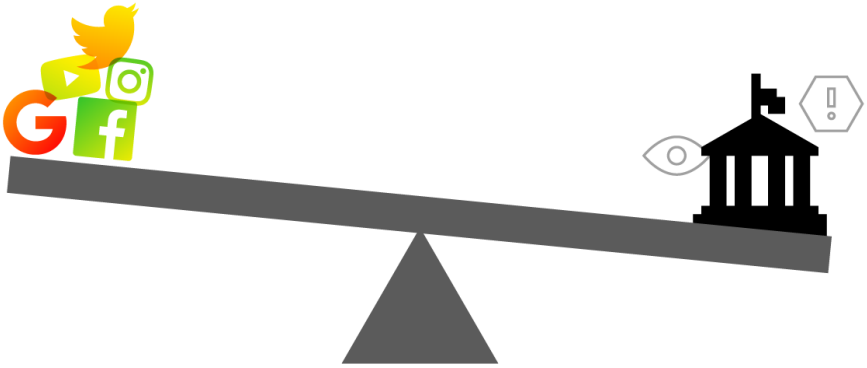
In addition, NewsQuality Initiative has worked with journalists and technologists to define appropriate goals for content selection and ranking (e.g. prioritize original local reporting, differentiate opinion pieces from news, give more weight to fact-checked articles), in order to tackle harms caused by algorithmic ranking systems. The initiative, a program at the Tow-Knight Center for Entrepreneurial Journalism at the Craig Newmark Graduate School of Journalism, seeks to elevate quality journalism in a time when algorithms are pervasively used to rank and recommend news articles online.

Some online platforms have also experimented with altering recommendation systems in order to combat the spread of misleading information. For example, in the months leading up to the 2020 U.S. presidential election, Facebook altered its ranking and recommendation algorithms to deprioritize new content known to be inaccurate and untrustworthy. This change, however, was reversed shortly after the election. Additionally, Twitter has proposed building a decentralized algorithm marketplace that allows users to "choose" their own recommendation algorithms.

While platform efforts to address the harms associated with algorithmic ranking and recommendation systems are important, regulation could also play a role in this regard. For example, regulation could require companies to provide more transparency around how these algorithmic systems are developed and how they operate through independent audits or impact assessments. Both audits and impact assessments require platforms or external third parties to evaluate platform operations using clear metrics. Regulators could then enforce penalties if a company fails to comply with these processes. Although there is currently no consensus around how audits and impact assessments should be structured, some researchers and civil society organizations have suggested that these mechanisms could consider factors such as algorithmic design, transparency, impact, public harm, user choice, and whether a company enforces its existing policies, such as those related to downranking, consistently. In addition, in order for externally conducted audits and impact assessments to succeed, companies must provide researchers with reliable and robust access to necessary data.

The initial draft of the Digital Services Act (DSA) in the European Union provides a good framework for considering this kind of algorithmic scrutiny, as it imposes special transparency and auditing obligations for very large digital platforms with more than 45 million monthly active users. The DSA proposes requiring these platforms to submit to mandatory audits both by third-party independent auditors as well as regulators, with risk assessments of "significant systemic risks" and transparency of recommender systems being subject to audit. Some analysts also favor a co-regulation approach, in which large digital platforms voluntarily adhere to minimum standards for regulating their own recommender systems.

Relevant Legislation



Despite the growing importance of AI and ML-based tools in curating and determining the content we see online, their use is still largely unregulated. **In the United States, the lack of legally binding mechanisms reflects ongoing difficulties in designing policy solutions that are compatible with the First Amendment.** Some experts suggest that under the Supreme Court's jurisprudence many algorithm-based decisions are considered speech protected by the First Amendment. The debate on whether the First Amendment and its interpretation by the Supreme Court are fit to address the challenges posed by automated tools and misleading information online is an ongoing and complex one. However, aside from this debate, there are areas where legislative action in the United States could be valuable.

One primary area of concern in the United States is the absence of comprehensive federal data privacy legislation that protects users from biased or otherwise harmful algorithmic decisions. Personal data is central to the development and deployment of AI and ML-based tools by social media platforms, which use them for training and refining their algorithms and making inferences about a specific individual. Over the past several years, members of Congress from both sides of the aisle have introduced at least ten privacy bills to regulate personal data collection and processing by technology companies. While all of these bills contain provisions relevant to the deployment of AI tools by social media platforms, and while many draw on key privacy principles—including data minimization, retention periods for personal information, and whether users can access, challenge, or correct decisions made by an algorithmic system—only a few introduce mechanisms to protect users from biased or otherwise harmful algorithmic decisions:

- The Online Privacy Act, introduced by Reps. Anna Eshoo (D-Calif.) and Zoe Lofgren (D-Calif.): The bill would give users the right to request a "human review of

impactful automated decisions" and require platforms to obtain opt-in permission from an individual before processing their personal information using a personalization algorithm.
- The Mind Your Own Business Act, introduced by Sen. Ron Wyden (D-Ore.): This bill would ask platforms to assess the impact that algorithms that process personal data have on accuracy, fairness, bias, discrimination, privacy, and security, and submit periodic reports to the Federal Trade Commission (FTC). The bill would also require the FTC to create a national 'Do Not Track' system that allows consumers to opt out of pervasive tracking, data selling or sharing, and the use of their personal information for ad-targeting purposes.
- The Consumer Online Privacy Rights Act, introduced by Sen. Maria Cantwell (D-Wash.): This bill would require platforms that use advertising algorithms to conduct an annual impact assessment that must address, "among other things, whether the system produces discriminatory results."

While these bills consistently focus on the impact of automated decision-making on bias and discrimination, they do not directly address algorithms' role in disseminating hate, promoting disinformation, and perpetuating systemic oppression for profit. To date, the only attempt to deal with the role of automated systems in promoting harmful material has been Sen. Edward J. Markey's (D-Mass.) KIDS Act, which aims to regulate how online content is presented to children, including through the use of AI tools.

A second legislative gap is the lack of transparency requirements for online platforms. Although transparency is a contested concept, and the design of transparency efforts depends on the target audience (e.g., users, public authorities, academia, civil society), transparency can be a useful way to help hold platforms accountable for the impact their algorithmic systems' have on the flow of information. Some bills that address this issue are:

- The Filter Bubble Transparency Act, introduced by Sens. Mark Warner (D-Va.) and John Thune (R-S.D.): Despite its name, this bill would not force platforms to disclose how their ranking algorithms work, but it would require them to notify users that the content they see—or do not see—online is filtered using an algorithm that processes personal data. Users should also be allowed to opt-out of this "filter bubble."
- The PACT Act, introduced by Sens. Brian Schatz (D-Hawaii) and Thune: In contrast to the Filter Bubble Transparency Act, this Section 230-reform bill (see below) would require online platforms to disclose their content moderation practices and publish biannual reports with disaggregated statistics on content that has been removed, demonetized, or deprioritized—including by an "automated detection tool."

- The Algorithmic Fairness Act, introduced by Sen. Chris Coons (D-Del.): This legislation would introduce transparency requirements for internet platforms. Additionally, it would direct the Federal Trade Commission (FTC) to evaluate the fairness of algorithms used to deliver online ads and search results and use its Section 5 authority to prevent unfair algorithmic decision-making.
- The Algorithmic Justice and Online Platform Transparency Act of 2021, introduced by Sen. Markey and Rep. Doris Matsui (D-Calif): This bill would prohibit discriminatory algorithms, empower the FTC to review platforms' algorithmic processes, and require online platforms to explain to users how they use algorithms to moderate, recommend, or amplify content, and what data they collect to power these algorithms. This legislation would also create an inter-agency task force to investigate the use of discriminatory algorithms in a variety of sectors.
- The Social Media DATA Act, introduced by Reps. Lori Trahan (D-Mass.) and Kathy Castor (D-Fla.): This legislation would increase transparency about online advertising by requiring large social media platforms to maintain an ad library open to academic researchers and the FTC, and directing the agency to set up a stakeholder group tasked with identifying best practices for sharing social media data with researchers.
- The Honest Ads Act, introduced by Sen. Amy Klobuchar (D-Minn.) with Sens. Warner and the late John McCain (R-Ariz.) in 2017: The bill would increase transparency around how advertising algorithms deliver political ads by requiring large online platforms to maintain a public database of all online political ads shown to their users and provide information on who the ads targeted, the buyer, and the rates charged. The Honest Ads Act also clarifies that digital political ads should be subject to the same disclaimer requirements as offline communications. The bill language was incorporated in the For The People Act, which passed out of the U.S. House in March 2021.
- The Social Media Transparency and Accountability Act of 2021, introduced by California Assemblymember Jesse Gabriel: This bipartisan state-level bill would require social media companies to file quarterly reports disclosing their policies on hate speech, disinformation, extremism, harassment, and foreign political interference; their efforts to enforce those policies; and any changes to their policies or enforcement practices.

Finally, **there are no accountability requirements for platforms that use algorithmic systems.** One of the most relevant attempts to address this gap is the Algorithmic Accountability Act of 2019, introduced by Sens. Cory Booker (D-N.J.) and Wyden, with Rep. Yvette Clarke (D-N.Y.) sponsoring a companion bill in the House. This bill would direct the FTC to create regulations requiring "companies that use, store, or share personal information" to assess the impact of their automated decision systems—including training

data—on "accuracy, fairness, bias, discrimination, privacy, and security," and address any identified issues "in a timely manner."

Some legislative efforts seeking to hold platforms accountable for their use of AI and ML-based tools have sought to amend Section 230 of the Communications Decency Act. Section 230 rightfully protects freedom of speech on the internet by establishing that platforms are not liable for third-party content on their services. But, it also allows social media companies to algorithmically amplify or recommend dangerous and inflammatory content with impunity.

While most Section 230 reform proposals are grounded in the unsubstantiated claim that platforms censor conservative viewpoints, some bills currently being considered by Congress would weaken the legal shield if a platform actively amplifies harmful content. These include:

- The Protecting Americans from Dangerous Algorithms Act, reintroduced by Reps. Tom Malinowski (D-N.J.) and Eshoo: The bill aims to keep platforms accountable for harms caused by their algorithms by removing liability protections when platforms' algorithms amplify or recommend content directly relevant to a civil rights case or cases involving acts of international terrorism. Notably, the bill would address harms caused by ranking and recommendation algorithms, but it would not remove the legal shield when algorithmic systems fail to remove harmful content or deliver harmful online ads.
- The SAFE TECH Act, introduced by Sen. Warner: The bill would remove platforms' legal immunity when they accept payment to make speech available or they have created or funded (both in whole or in part) the speech. Platforms would also lose their liability protections if a plaintiff seeks an injunction because the service failed to "remove, restrict access to or availability of, or prevent dissemination of material that is likely to cause irreparable harm," thus incentivizing platforms to calibrate their algorithms in favor of over-moderation.
- The Civil Rights Modernization Act, introduced by Rep. Clarke: The bill would amend Section 230 to ensure civil rights laws apply to the targeting and delivery of advertisements, including when ads are delivered or published using "any information technology, including an algorithm or a software application."

However, some advocates have noted that many existing proposals, which make injudicious changes to Section 230, do not adequately address the harms caused by the surveillance advertising business model.

In the absence of legislation, the FTC has stepped in and outlined principles and best practices surrounding algorithmic transparency, explainability, bias, and robust data models. It has also taken unprecedented enforcement actions to limit the use of algorithms

that have discriminatory effects on consumers. FTC Commissioner Rebecca Kelly Slaughter also created a "rulemaking group" within the office of the agency's general counsel tasked with drafting new rules to address anti-competitive corporate behavior, including rules focused on transparency around algorithms. More recently, the federal agency issued guidance to companies on how they should manage the consumer protection risks stemming from AI and algorithms: be transparent about data collection and processing practices; explain to consumers impacted by algorithmic decisions which factors were taken into account; ensure the fairness of algorithmic decision-making; ensure that data models are robust and sound; and abide by strict ethical standards. With these guidelines, the FTC also signaled that it stands ready to take law enforcement action against companies that use algorithmic systems that entrench racial and gender bias.

While U.S. lawmakers are struggling to reach a consensus on how to regulate internet platforms and the algorithmic systems they use, the EU has set forth a bold agenda based on three pillars: ensuring that digital technologies actually work for the people; promoting a fairer and more competitive digital economy; and creating a trustworthy digital environment that empowers citizens, enhances democratic values, and respects fundamental rights. So far, the European Commission has introduced three legislative proposals that address the use of AI and ML-based tools by online internet platforms:

- The Digital Services Act (DSA): The DSA seeks to promote transparency, accountability, and regulatory oversight over EU digital services. The DSA outlines obligations that online intermediary services must meet when they remove illegal and harmful content from their services and when they deploy content moderation and curation mechanisms. For example, the DSA would require platforms to provide users with meaningful information on digital ads, including why they have been targeted, and it would require very large online platforms to meet a higher standard of transparency and accountability around how they moderate content, deliver advertising, and use algorithmic processes.
- The Digital Markets Act (DMA): Although this legislative proposal doesn't directly address the problem of misleading content online or the use of AI and ML-based tools, it establishes new prohibitions and obligations for large online platforms (so-called "gatekeepers") to avoid unfair market practices that may harm competition. There is growing consensus among policymakers in the United States and the EU, as well as in academic circles, that internet platforms' unchecked monopoly power is a threat to democracy.
- The Artificial Intelligence Act: This legislative proposal sets out a risk-based approach to regulating the use of AI systems. Although the draft does not include provisions that specifically target internet platforms, it does prohibit the use of AI systems that deploy "subliminal techniques" to manipulate behavior in a manner that "causes or is likely to cause" physical or psychological harm to self or others.

While this provision could theoretically include recommendation and advertising systems used to curate online content, it will be up to an enforcing agency or the courts to determine whether they are exploitative or manipulative.

These three proposals are currently being negotiated by the European Parliament and EU member states.

Additionally, the European Commission recently formulated recommendations on how social media companies should govern their algorithms in its guidance for the upcoming revision of the EU Code of Practice on Disinformation. Created in 2018, the Code contains voluntary commitments to tackle misleading information online. Current signatories include Facebook, Google, Twitter, Mozilla, Microsoft, TikTok, trade associations representing online platforms, and other key players in the ad-tech industry. In its recent guidance, the Commission has emphasized that it wants social media companies to disclose the criteria used to prioritize or deprioritize content, give users the option to customize ranking algorithms, and remove "false and/or misleading information when it has been debunked by independent fact-checkers and [exclude] webpages, and actors that persistently spread disinformation." Current signatories have already started revising the Code, and a first draft is expected in late 2021.

The European Commission also conducted a public consultation on how to regulate sponsored political content, both online and offline, with the goal of introducing draft legislation later in 2021. The Commission already outlined the need for greater transparency obligations for digital ads in the DSA. In this regard, the European Data Protection Supervisor (EDPS) has called on EU legislators to consider a gradual phasing out of surveillance advertising, as well as restrictions on categories of data that can be processed to target users.

Recommendations

The following section outlines areas of work that internet platforms should prioritize in order to better address the role AI and ML-based tools play in fueling online misinformation and disinformation. Civil society groups and lawmakers should similarly prioritize advocacy around these efforts and encourage platforms to implement these recommendations.

1. **Publish accessible and comprehensible versions of their content moderation, ranking, recommendation, and advertising policies that are available to the general public.** These policies should outline what kinds of organic content and ads are permitted on the service, how the company enforces these policies, and how automated tools are used for detection and enforcement. These policies should incorporate specific provisions related to misinformation and disinformation, including policies that prevent users and entities from being able to advertise and monetize on a service if they repeatedly spread misleading information, and policies that would require a platform to remove, downrank, or prevent the recommendation of groups that spread misleading content as well as content that has been fact-checked and deemed misleading. Platforms should ensure that their organic content and advertising policies are easily accessible in one central location and should strive to ensure these policies are consistently enforced.

2. **Establish processes for fact-checking all advertisements, and fact-checking high-reach content.** These policies should be applied to organic and paid content, regardless of who posts them**,** and should be consistently enforced.

3. **Issue transparency reports which outline how the company has used AI and ML-based tools for content moderation and curation purposes and what impact**

these tools have had on online speech. For example, companies should publish data outlining how much content and how many accounts they have taken enforcement action against for violating their policies on misinformation and disinformation. This data should be easily accessible and available in one central location.

4. **Establish accessible and searchable ad transparency libraries which feature all of a platform's online ads, including all political and issue ads.** These ad libraries should be public and not based on private one-on-one agreements between the platform and individual researchers. The data should be directly accessible with an open and accessible public API, not behind some form of custom software that the platform controls, and can therefore rescind or deprecate. In addition, platforms should collaborate with civil society and researchers to understand how they can restructure their ad libraries to provide more standardization and more meaningful and comprehensive transparency.

5. **Share advertising enforcement data.** This data should outline how many ads the company has removed for violating its ad policies, broken down by category of ad, and how many violating ads the company mistakenly allowed to run before removing them.

6. **Provide users with adequate notice when their content or accounts have been flagged for violating one of the company's moderation or curation policies.** This notice should clearly explain what policy the user violated—and, where relevant, include information on how the user can appeal the moderation decision.

7. **Give users access to a timely appeals process.** Appeals should involve timely review by a person or panel of persons who were not involved in the original decision, and should allow users to provide additional information to be considered during the review. In addition, users who are regularly subject to hate, harassment, and misleading information should be able to report content at scale.

8. **Ensure humans are kept in the loop when deploying algorithmic systems that do not have a high degree of accuracy.** This is especially important for algorithmic content moderation purposes, as overbroad content moderation can chill free speech.

9. **Invest in processes that tackle the spread of misleading information in different languages.** In particular, companies should allocate more resources towards hiring and training human content moderator workforces that cover a range of languages

and regions. Companies should similarly invest in developing technological tools that can more effectively moderate content across different linguistic and regional contexts.

10. **Conduct regular proactive audits and/or submit to external third-party audits on ad-targeting and delivery, content moderation, ranking, and recommendation systems in order to identify potentially harmful outcomes, such as bias and discrimination.** Companies should take concrete steps to eliminate or address any identified harms—for example, by making adjustments to the algorithm or training data. Companies should also publish a public summary of audit findings and any mitigation efforts they made.

11. **Prepare adequately for an increase in problematic content surrounding important events and accounts.** Companies should invest in developing resources and providing training for skilled and experienced moderators who are focused on content moderation and curation around these events.

12. **Introduce robust privacy protections and user controls.** These controls should allow users to determine how their personal data is collected and used by algorithmic systems, and what kind of content they see. For example, users should be able to control how their personal data is used to inform the recommendations or ads they receive, and users should be able to opt out of seeing certain categories of recommendations or ads.

13. **Create robust tools and mechanisms that enable researchers to conduct thorough research and analysis on algorithmic content curation systems.** In particular, companies should provide researchers with access to better simulation tools and other tools that empower, rather than limit, large-scale research and analysis. Companies should also provide researchers with access to social media data in a privacy-preserving fashion. Further, companies should support efforts by researchers and think tanks to monitor and evaluate the impact of online misinformation and disinformation, especially on communities of color. Lastly, platforms should include exceptions for public-interest research in their terms of service, for example with regard to scraping public information or creating temporary research accounts.

In addition, although platforms' ongoing efforts to ensure that AI and ML tools work in the public interest through self-regulation should be continued, it is important to recognize that self-regulation will often be insufficient due to the financial incentives underlying the platforms' advertising-driven business models. As a result, when feasible, lawmakers

should pursue policy and legislation necessary to change platforms' incentives and ensure their commitment to tackling online misinformation and disinformation. In particular, lawmakers should:

1. **Pass comprehensive federal privacy legislation.** This legislation should draw on key privacy principles, including data minimization, retention periods for personal information, and whether users can access, challenge, or correct decisions made by an algorithmic system.

2. **Enact rules to require greater and meaningful transparency from online platforms.** This could include rules that require platforms to issue regular reports on their content moderation, curation, and ad targeting and delivery efforts.

3. **Clarify that offline anti-discrimination statutes apply in the digital environment and ensure adequate enforcement mechanisms.** These include the Voting Rights Act, the Civil Rights Act of 1964, and the Fair Housing Act.

4. **Ensure that any legislative efforts seeking to hold platforms accountable for their use of AI and ML-based tools directly address the harms of these systems.** Lawmakers should especially avoid using Section 230 as a mechanism for tackling algorithmic harms, unless doing so would clearly resolve the harms.