Twitter

Twitter is one of the world's most popular social media platforms,[113] with over 330 million monthly active users around the globe.[114] Like other social media platforms, Twitter has been heavily scrutinized for its role in facilitating the spread of misinformation and disinformation, particularly related to COVID-19. In response, Twitter launched a range of initiatives and efforts, which are documented, alongside regular updates, in an online repository hosted on the company's website.[115]

In January 2020, the company shared that it expanded its dedicated search prompt feature to ensure content from authoritative sources appears at the top of search results related to COVID-19. According to the company, this feature is now available in approximately 70 countries, and the company has partnered with national public health agencies, the WHO, and local partners to ensure users have access to verified information.[116] Twitter has also prevented its auto-suggest feature from directing users to misleading sources when they enter COVID-19 related search terms.[117] This change was part of an expansion of the company's "Know the Facts" prompt,[118] which was established in 2019 to provide users with access to clear, legitimate information related to immunizations and vaccinations.[119]

Like Facebook, Twitter has stated that it is increasingly relying on automated and machine learning tools to moderate content during the pandemic.[120] In particular, these automated tools will be used to identify reports related to pieces of content that are likely to cause harm and surface them for priority review and proactively identify violating content before it is reported.[121] As a result of this increased reliance on automated tools, users have been told to expect more moderation errors.[122] Consequently, the company has said it will not permanently suspend any accounts as a result of automated enforcement decisions. The company has also said it will maintain its appeals process, although there may be delays, and it will implement human review procedures where possible.[123] The company shared that, during this pandemic, it will prioritize the review of content that poses a direct risk to the health and well-being of individuals[124] and that content that requires contextual analysis, including misleading content related to the pandemic, will undergo human review.[125] The platform has warned, however, that it will not be able to review every tweet that contains misleading or disputed information about the virus during this time.[126]

In addition, Twitter has responded to the rapid spread of misleading information related to the virus on the platform by expanding its definition of harmful content to include content that goes against guidance provided by "authoritative sources of global and local public health information."[127] Twitter enforces these policies

with the support of its trusted partners, which include public health authorities and governments, and reviews content flagged for violating these policies against information provided by these trusted partners.[128] In addition, the platform has shared that it will continue to enforce its policies on platform manipulation during this time, which prohibits the use of the Twitter platform "in a manner intended to artificially amplify or suppress information or engage in behavior that manipulates or disrupts people's experience on Twitter."[129] The company has said that it has not yet seen any major coordinated platform manipulation efforts related to the virus.[130]

On April 22, the company also announced that it will prioritize the removal of content that could lead to the destruction or damage of critical 5G infrastructure.[131] This policy shift is in response to the spread of an internet conspiracy theory that claims radio waves emitted by 5G technology are eliciting changes in people's bodies that make them more susceptible to the coronavirus. The spread of such misinformation has resulted in dozens of acts of arson against wireless towers and telecom equipment, as well as the harassment of countless telecom employees in many countries.[132]

Under Twitter's COVID-19 content policies, the company does not permit tweets that:

- Deny global or local health authority recommendations (e.g. related to social distancing)
- Deny established scientific facts about transmission of the virus and the difference between the virus and other diseases
- Promote unproven or harmful treatments, protection measures, diagnostic criteria, and cures for the virus
- Share claims that intend to manipulate behavior to support a third-party (e.g. the virus is not real, leave your house and support business X)
- Propagate information that creates panic, unrest, and disorder
- Share claims made by an individual who is impersonating a government or health official or organization (e.g. parody accounts)
- Promote the notion that certain nationalities or groups are more or less susceptible to the virus[133]

The company has outlined, however, that it may apply its public interest exception policy to cases in which world leaders and elected and other government officials have violated these COVID-19 content guidelines. In these

cases, the company will determine that there is public interest value in keeping the content on the service, such as, the public will be able to know that these leaders are publishing misinformation. Therefore, instead of removing the content, Twitter will place the content behind a notice that provides context about the violation and allow people to view the content only if they wish to see it.[134] Users have a right to access information, including from world leaders and elected and other government officials, as well as a right to know what their leaders are saying, especially during a crisis period such as this one. Online platforms are a major outlet for information and as a result companies should institute such a public interest exception and notice policy. However, companies should institute this policy responsibly. If a leader's content violates the platform's content policies, in most cases, it should be left up with a clear notice that explains why the content has been left up. In addition, this content should be fact-checked and platforms should provide additional context to users in the notice detailing whether the post contains misleading information. However, if content posted by these leaders poses imminent harm, platforms should remove this content just as they would for content from anyone else, as it can have significant offline consequences.

In an update on April 1, the company shared that since the expansion of its content policies during the pandemic, the company has removed over 1,100 Tweets with misleading and harmful content, and its automated tools have challenged over 1.5 million accounts for spam or manipulative behavior in COVID-19 discussions.[135] These periodic updates are valuable for providing transparency and accountability around the platform's efforts to combat misleading content during the pandemic. However, the company's existing *Twitter Rules Enforcement* report does not currently cover moderation of misleading content. Going forward, Twitter should continue to provide periodic updates on its moderation efforts during the pandemic. Following the pandemic, the company should publish a comprehensive COVID-19-specific transparency report outlining the scope and scale of these moderation efforts during the pandemic as a whole. Further, the company should begin regularly reporting on the moderation of misleading content in its regular transparency report.

On May 11, Twitter also announced that it will begin appending labels to tweets that feature potentially misleading or harmful information related to the virus. These labels will direct users to a page, curated by the company or by an external trusted source, that contains additional information on the content of the tweet. The company has also stated that it will append a warning to certain tweets depending on their "propensity for harm and type of misleading information." These warnings will notify users, before they view the tweet, that the information in the tweet goes against public health guidance. The company shared an infographic,[136] included below, that outlines the scenarios in which the company would add a label, add a warning, remove, or take no action against a tweet that potentially contains misinformation during this time. Labels may be visible on

tweets even when they are embedded or when they are being viewed by individuals who are not logged in.[137] This new policy will also apply to posts shared by public officials, and may be applied retroactively.

Misleading Information	Label	Removal
Disputed Claim	Label	Warning
Unverified Claim	No action	No action*
	Moderate	Severe
Propensity for Harm		

Source: Twitter

According to Twitter, the company will use its internal tools to proactively monitor content on the platform and to make sure that the company is not amplifying content by appending labels to them. Twitter is also working with its trusted partners to flag content that could yield harmful offline consequences, and will be prioritizing the review and labeling of content that could result in increased exposure to or transmission of the virus.[139]

As previously highlighted, advertising can also be a source of COVID-19-related misinformation and disinformation. As a result, Twitter has introduced new rules that only permit advertisers to explicitly or implicitly mention the virus in their ads if they are discussing "adjustments to business practices and/or models in response to COVID-19" and "support for customers and employees related to COVID-19."[140] Twitter does not permit advertisers to run ads that feature sensational content and inflated product prices, or which are for products that are in high demand as a result of the pandemic (e.g. face masks, alcohol hand sanitizers, etc.).[141] Under Twitter's political ads content policy,[142] news publishers receive an exemption to these advertising rules and are able to promote content that discusses vaccines, treatments, and test kits.[143] The company is also permitting government entities to disseminate public health information through advertising on the platform. Further, Twitter is using its "Ads for Good" program to provide advertising credit to nonprofit organizations so that they can run advertising campaigns for fact-checking services and promote reputable health information.[144]

During the pandemic, Twitter should publish periodic updates on its ad enforcement efforts, including the number of ads the company has removed for violating its COVID-19-specific ad policies and the number of ads that were erroneously permitted. Following the pandemic, Twitter should publish more comprehensive data outlining the scope and scale of its ad enforcement efforts during this time, especially as they relate to COVID-19-specific advertising policies. In addition, where appropriate, the FTC should enforce Section (5)(a) of the FTC Act, and hold businesses and sellers accountable when they engage in unfair and deceptive trade practices through online ad campaigns.