On Social Media, Only Some Lies Are Against the Rules

Your guide to every major social media company's misinformation policies on vital topics from COVID-19 to voting

By Kaveh Waddell. Visualizations by Andy Bergmann.

An outrageous conspiracy theory vilifying a political candidate. A meme encouraging shots of bleach to fend off COVID-19. An official-looking post falsely announcing that your polling place has moved.

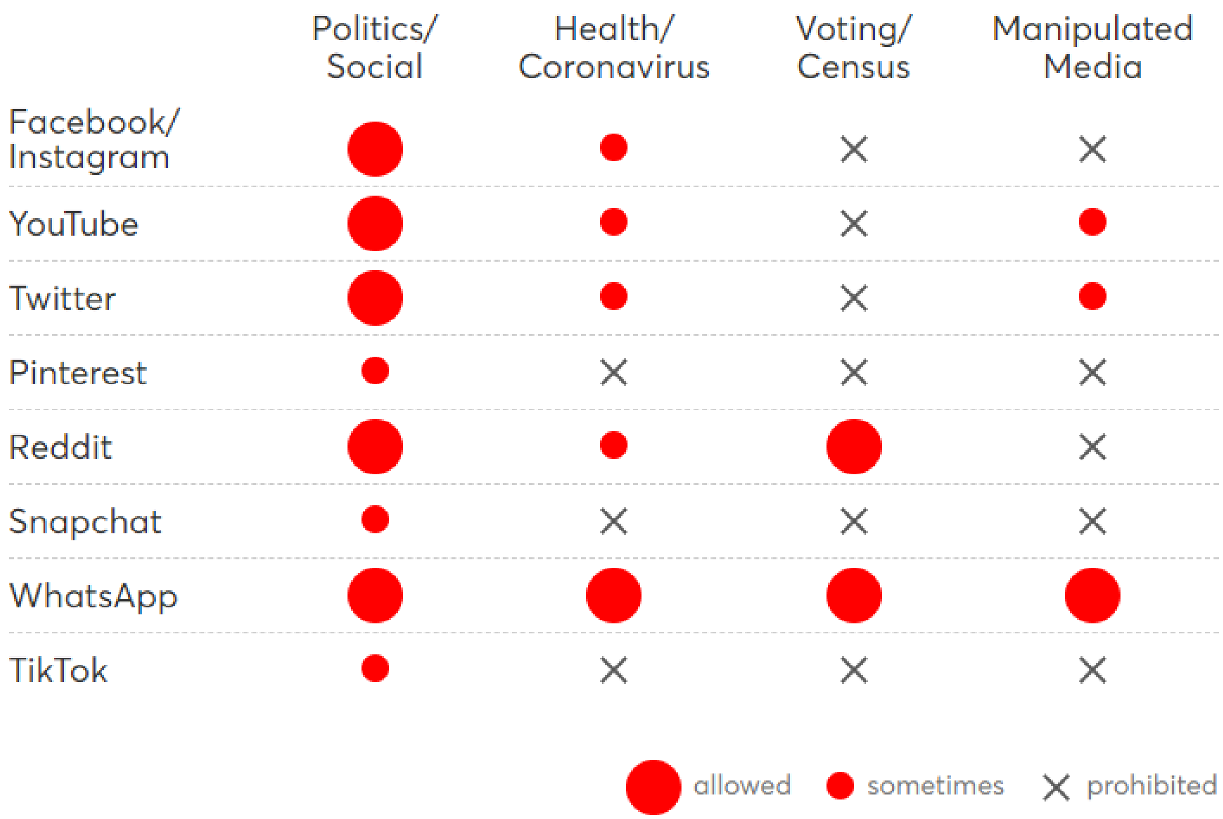
Lies like these seethe on the social media platforms, from Facebook to YouTube to Twitter, where Americans increasingly get their news and information.

These companies say they want to limit dangerous falsehoods while also protecting free speech. But the platforms' rules on misinformation vary widely. And their policies are often "confusing, unclear, or contradictory," according to Bill Fitzgerald, a privacy and technology researcher in CR's Digital Lab.

That makes it hard to know what to expect on each platform, and to choose where to go for important information. It's also difficult to demand that companies enforce their rules fairly when you can't even figure out what those rules are supposed to be.

To help sort this all out, Consumer Reports analyzed misinformation policies from the country's biggest social media platforms. (We also considered public statements by executives.) We focused on the most dangerous types of falsehoods, including misinformation on the coronavirus and on how to vote. Our findings are summarized here.

Which platforms allow false information



	Politics/ Social	Health/ Coronavirus	Voting/ Census	Manipulated Media
Facebook/ Instagram	●	•	×	×
YouTube	●	•	×	•
Twitter	●	•	×	•
Pinterest	•	×	×	×
Reddit	●	•	●	×
Snapchat	•	×	×	×
WhatsApp	●	●	●	●
TikTok	•	×	×	×

● allowed • sometimes × prohibited

As you can see, most of the platforms don’t have a blanket rule against posting false material, but they do ban certain kinds of misinformation. Some allow specific types of false claims. Others only *sometimes* allow a type of false post, depending on details such as the level of danger it may pose. And when it comes to a few issues, a platform may ban all misinformation. (You can click on any icon in the chart for more information.)

We provide more detail below on how platforms handle common categories of misinformation, from vague conspiracy theories to outright lies about the coronavirus or when polling places are open. We’ve grouped Facebook and Instagram because they share a rulebook.

In some cases, companies treat various types of users differently. Facebook and Twitter give world leaders more leeway than ordinary users to post misleading information, at least on some topics, arguing that a leader’s opinions are inherently newsworthy.

Advertisers face the strictest rules on almost every platform. Twitter says it doesn’t allow ads that are “false, deceptive, misleading, defamatory or libelous.” Snapchat says that ads can’t be “false or misleading," and Reddit ads need to be "truthful, non-deceptive, and defensible," according to the company. (As CR found earlier this year, it can be easy to get even dangerous health ads approved.)

We shared our results with the social media platforms, and most agreed with what we found. Pinterest disputed our determination, reflected in the charts, that it knowingly allows some misinformation. The company tells CR it bans misinformation of any kind, but its published policies state that it removes only posts that are determined to have the potential to cause harm to users or the public. A Reddit spokesperson said that misinformation on voting is banned, but that's not included in a published policy.

Snapchat's written policies say the company allows some false information in cases where the content is newsworthy. But in a statement to CR, Jennifer Stout, Snapchat's vice president for global public policy, stated that the newsworthiness carve-out does not apply to misinformation about health issues and voting, or to manipulated media. Those would always be against the rules. Snap disputes CR's determination that certain types of political and social misinformation would be allowed on the platform.

Separately, we've prepared a guide to using the companies' settings to fine-tune what appears in your social feeds, avoiding hate speech, violent images, and hoaxes—or simply tuning out political posts, if you prefer to get your news elsewhere.

We hope that laying all this out will strengthen the consumer's voice in shaping social media policies on misinformation, and insisting on better, fairer enforcement.

Politics and Social Issues

Which platforms allow false information about politics and social issues

Facebook/Instagram	●	Reddit	●
YouTube	●	Snapchat	•
Twitter	●	WhatsApp	●
Pinterest	•	TikTok	•

● allowed • sometimes ✕ prohibited

Companies are largely hands-off when it comes to falsehoods about a broad range of political and social issues. Examples range from a photo of a young Donald Trump flanked by his parents in KKK garb (it's Photoshopped); to allegations that a hacked Democratic National Committee server is hidden in Ukraine (it's a hoax); to claims that George Floyd, whose killing by Minneapolis police sparked nationwide Black Lives Matter protests, is a "crisis actor" who's still alive (that's untrue, of course).

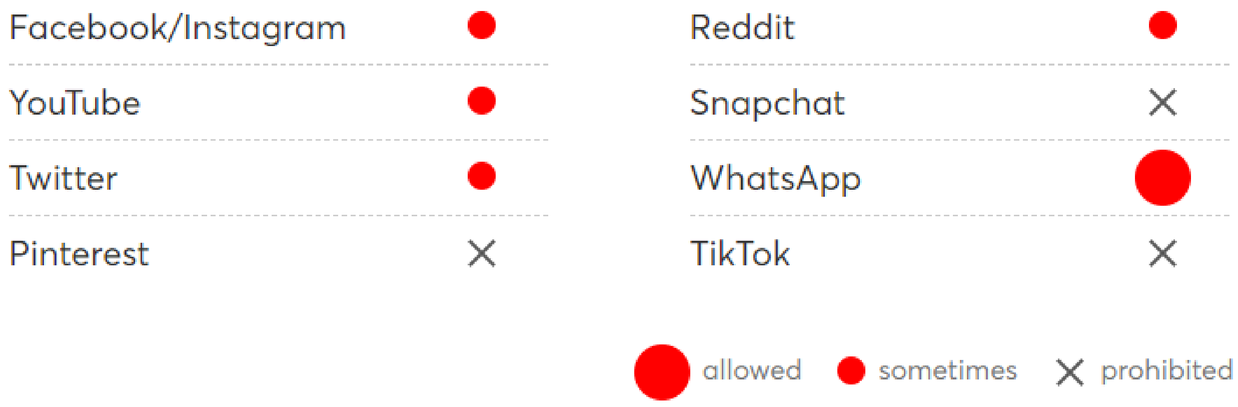
Social networks generally let people post this sort of thing, although since the 2016 election they say they've cracked down on coordinated disinformation campaigns.

But platforms may take actions short of deleting a misleading post. Both Facebook and Twitter label some posts as suspect, with links to more reliable information. They may also bury misleading posts deeper in users' search results and news feeds. In those cases, you can still find the false information if you go looking for it.

This halfway approach—flag and downrank, but don't delete—has advantages, says Bridget Barrett, a policy researcher at the University of North Carolina who led a recent study of social media misinformation policies that helped inform CR's own research. "It feels more honest about the fact that this is complicated," she says.

Health and Medicine

Which platforms allow false information about health issues



Facebook/Instagram	•	Reddit	•
YouTube	•	Snapchat	✕
Twitter	•	WhatsApp	●
Pinterest	✕	TikTok	✕

● allowed • sometimes ✕ prohibited

Misinformation on the coronavirus has been tightly restricted since dangerous falsehoods about the virus swept through social networks starting in early 2020.

Facebook said in May that it had already removed hundreds of thousands of posts and put “warning labels” on about 50 million posts that violated its coronavirus policies. Click on a coronavirus post that Facebook later takes down, and later you might see a big link to a mythbusting page appear at the top of your newsfeed.

“From my perspective, the [policies] that have been most effective . . . are the ones that are hyper-focused,” says Jevin West, director of the Center for an Informed Public at the University of Washington.

Twitter and Facebook recently took down posts showing a news clip where President Donald Trump said that children are “almost immune” to COVID-19 (not true), but plenty of coronavirus-related misinformation still slips through the cracks.

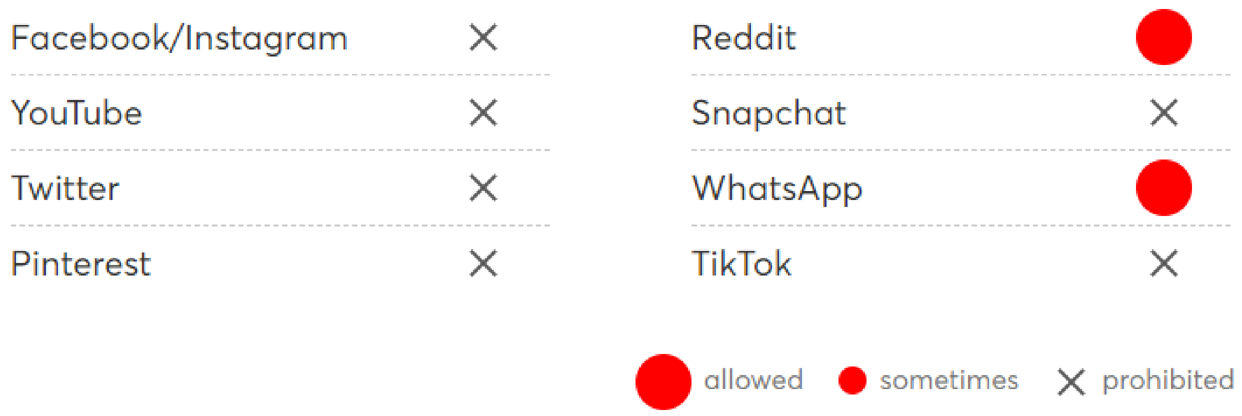
The nonprofit advocacy group Avaaz found dozens of egregious examples of coronavirus misinformation on Facebook that lacked a warning label—including some posts pushing information that had previously been flagged as misleading.

Other health-related falsehoods are allowed to proliferate on YouTube, Twitter, and Reddit, though the companies sometimes bury them behind trustworthy sources. Some platforms ban this content, which can include endorsements of fake miracle cures or anti-vaccination conspiracy theories.

All the platforms we studied have also taken up a new trick to promote safe, authoritative information on COVID-19: Try searching for “coronavirus” on any major network, and you’ll see a special box at the top of your results with links to the Centers for Disease Control, the World Health Organization, or credible news organizations. These results can take up several pages, and similar links have been rolled out for other hot-button issues like voting and suicide prevention.

Voting and the Census

Which platforms allow false information about voting and the census



Facebook/Instagram	✕	Reddit	●
YouTube	✕	Snapchat	✕
Twitter	✕	WhatsApp	●
Pinterest	✕	TikTok	✕

● allowed • sometimes ✕ prohibited

Most social media companies say they have little tolerance for baldly misleading posts about how to vote or to participate in the U.S. Census. Examples would include giving out the wrong hours for a polling place or claiming that people can cast a vote for president online, which they can’t.

However, this is an area where Twitter has diverging rules for world leaders and the rest of us. In the spring, Trump tweeted, falsely, that California was sending mail-in ballots to “anyone living in the state, no matter who they are or how they got there” and that mail-in ballots would inevitably lead to widespread abuses. Twitter added a link explaining that only registered voters received the ballots and that there’s no evidence linking voting-by-mail to widespread voter fraud.

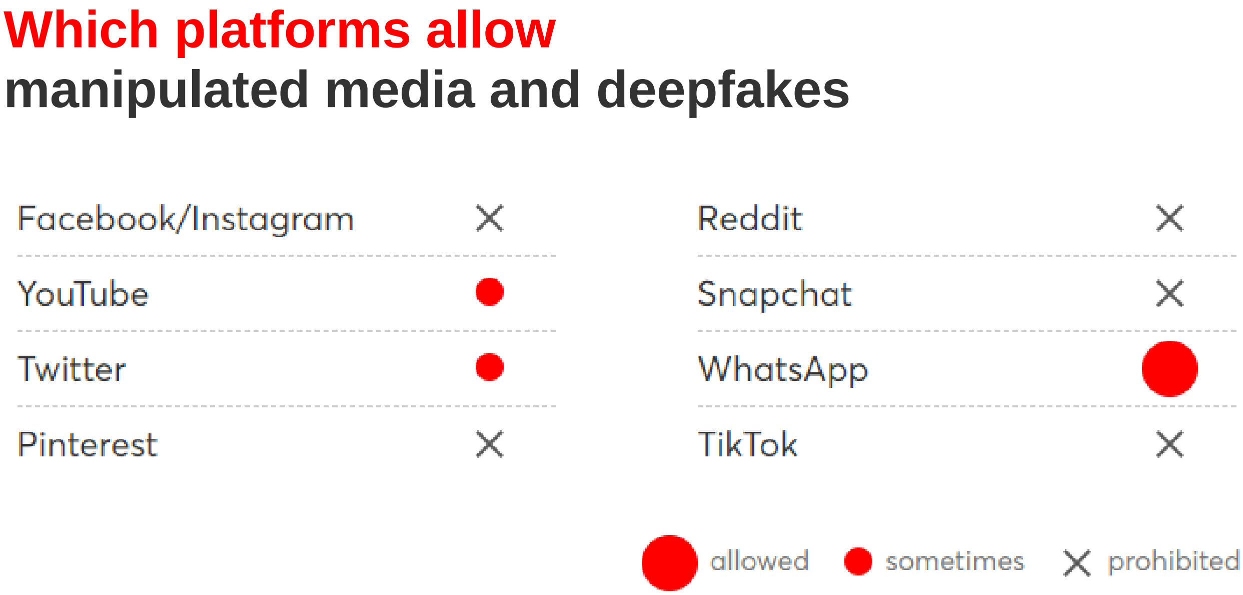
An average user tweeting the same message may have found their tweet deleted instead. On the other hand, the company took no action at all on more recent tweets by Trump making similar claims—including one post in which he floated the idea that the 2020 election would be so flawed it should be postponed.

Twitter has used similar labels for politicians overseas, including in Brazil.

In June, Facebook began adding links to an official government page on voting to every post from a presidential candidate or federal official that mentions voting—whether or not it's false. CEO Mark Zuckerberg wrote in a Facebook post that the decision stemmed from the "difficulty of judging [the veracity of posts] at scale."

Critics say the policy erodes an important distinction between lies and legitimate political speech. The same message from Facebook has been added both to unfounded claims that the upcoming election would be rigged and to straightforward posts from candidates asking for votes.

Deepfakes and Manipulated Video



Which platforms allow manipulated media and deepfakes

Facebook/Instagram	✕	Reddit	✕
YouTube	•	Snapchat	✕
Twitter	•	WhatsApp	●
Pinterest	✕	TikTok	✕

● allowed • sometimes ✕ prohibited

Social media companies are still figuring out how to deal with deepfakes—a new breed of hyper-convincing fake images, audio, and video made possible by fast-developing AI technology. With deepfakes, it's becoming possible to create a synthetic clip of anyone saying whatever you want them to.

YouTube and Twitter say they will take down particularly dangerous deepfakes—imagine a convincing video of a world leader declaring war, or a powerful CEO reporting false sales numbers—but will leave others alone.

Those platforms also ban videos edited with conventional techniques to mislead people. These videos are far more widespread than deepfakes. This month, YouTube and Twitter both removed a video of House Speaker Nancy Pelosi that was slowed down and edited to make it seem that she'd been slurring her words, as though intoxicated. (Similar manipulated videos of Pelosi have spread widely in the past.)

By contrast, Facebook placed a "partly false" label on the video.

In another example of companies' different approaches, Twitter labeled a video Trump tweeted as "manipulated" because it included a fake CNN banner and headline. Facebook left the video untouched. (Both companies ended up taking the video down for copyright infringement.)

Misinformation experts say we should expect more of these simple video edits, sometimes called "cheapfakes," to test the policies put in place by the platforms.

Making the Call on Misinformation

Social media companies use a combination of human teams and computer algorithms to hunt for misinformation. But most companies say little about how they balance the human touch with automated moderation.

The social networks say they have to rely on artificial intelligence because of the unending tidal wave of posts they deal with every day. Many thousands of people work as content moderators for Facebook—but their workload is enormous: In 2017, the most recent data point available, CEO Mark Zuckerberg said Facebook was reviewing more than 100 million posts a month.

Facebook says that 99 percent of the violent and graphic content that it removes is flagged by algorithms—ditto for 97.7 percent of the posts related to suicide and self-injury, and 88.8 percent of hate speech, according to the most recent data. Of course, these numbers don't include the posts that Facebook doesn't catch.

Automated moderation is much more likely to make mistakes than people are, computing experts say. "Machine learning can't do it, and it's not even close," says UW's West, who teaches courses on AI. Humor and sarcasm often slip through machines' grasp, for example.

That helps explain why social media companies are often accused of leaving up posts that should come down—and deleting posts that shouldn't. Journalists and researchers have turned up rampant examples of voting misinformation, falsehoods about Black Lives Matter protests, and conspiracy theories about vaccines.

Many social activists have complained about being put in "Facebook jail," having their accounts being suspended for discussing racism, even while some blatantly racist posts are left untouched.

"In some cases, what actually happens doesn't match up with the stated policy," says Nathalie Maréchal, senior policy analyst at Ranking Digital Rights, a nonprofit that grades tech companies on factors including their privacy and content moderation practices. "We urgently need more transparency from companies" about enforcement.

Experts in the field say that consumer pressure could lead the platforms to do more. In July, hundreds of advertisers pulled ads from Facebook, following popular outcry against lax policies toward hate speech and misinformation. In recent months, Twitter, Reddit, YouTube, and Facebook have all thrown dangerous conspiracy theorists off their platforms.

"When social media is so intrinsic to how society works, citizens should have some say on what those policies are," West says. "We are consumers of these major platforms that we don't necessarily pay for—but we could still make choices with where we spend our time."

Kaveh Waddell

I investigate privacy and discrimination online for Consumer Reports' Digital Lab. In the past, I've reported on technology at Axios, The Atlantic, and National Journal. I'm a Seattle native now based in the Bay Area, where I try to spend as much time as possible outside—and away from my screens. Find me on Twitter at @kavehwaddell.