Synthetic and manipulated media policy

1. Help Center (https://help.twitter.com/)
2. Platform integrity and authenticity (https://help.twitter.com/en/rules-and-policies#platform-integrity-and-authenticity)

Synthetic and manipulated media policy

Overview

You may not share synthetic, manipulated, or out-of-context media that may deceive or confuse people and lead to harm ("misleading media"). In addition, we may label Tweets containing misleading media to help people understand their authenticity and to provide additional context.

What is in violation of this policy

In order for content with **misleading media** (including images, videos, audios, gifs, and URLs hosting relevant content) to be labeled or removed under this policy, it must:

- Include media that is significantly and deceptively altered, manipulated, or fabricated, or
- Include media that is shared in a deceptive manner or with false context, and
- Include media likely to result in widespread confusion on public issues, impact public safety, or cause serious harm

We use the following criteria as we consider Tweets and media for labeling or removal under this policy as part of our ongoing work to enforce our rules and

ensure healthy and safe conversations on Twitter:

1. Is the content significantly and deceptively altered, manipulated, or fabricated?

In order for content to be labeled or removed under this policy, we must have reason to believe that media are significantly and deceptively altered, manipulated, or fabricated. Synthetic and manipulated media take many different forms and people can employ a wide range of technologies to produce these media. Some of the factors we consider include:

- whether media have been substantially edited or post-processed in a manner that fundamentally alters their composition, sequence, timing, or framing and distorts their meaning;
- whether there are any visual or auditory information (such as new video frames, overdubbed audio, or modified subtitles) that has been added, edited, or removed that fundamentally changes the understanding, meaning, or context of the media;
- whether media have been created, edited, or post-processed with enhancements or use of filters that fundamentally changes the understanding, meaning, or context of the content; and
- whether media depicting a real person have been fabricated or simulated, especially through use of artificial intelligence algorithms

We will not take action to label or remove media that have been edited in ways that do not fundamentally alter their meaning, such as retouched photos or color-corrected videos.

In order to determine if media have been significantly and deceptively altered or fabricated, we may use our own technology or receive reports through partnerships with third parties. In situations where we are unable to reliably determine if media have been altered or fabricated, we may not take action to label or remove them.

2. Is the content shared in a deceptive manner or with false context?

We also consider whether the context in which media are shared could result in confusion or suggests a deliberate intent to deceive people about the nature or origin of the content, for example, by falsely claiming that it depicts reality. We

assess the context provided alongside media to see whether it provides true and factual information. Some of the types of context we assess in order to make this determination include:

- whether inauthentic, fictional, or produced media are presented or being endorsed as fact or reality, including produced or staged works, reenactments, or exhibitions portrayed as actual events;
- whether media are presented with false or misleading context surrounding the source, location, time, or authenticity of the media;
- whether media are presented with false or misleading context surrounding the identity of the individuals or entities visually depicted in the media;
- whether media are presented with misstatements or misquotations of what is being said or presented with fabricated claims of fact of what is being depicted

We will not take action to label or remove media that have been shared with commentary or opinions that do not advance or present a misleading claim on the context of the media such as those listed above.

In order to determine if media have been shared in a deceptive manner or with false context, we may use our own technology or receive reports through partnerships with third parties. In situations where we are unable to reliably determine if media have been shared with false context, we will not label or remove the content.

3. Is the content likely to result in widespread confusion on public issues, impact public safety, or cause serious harm?

Tweets that share misleading media are subject to removal under this policy if they are likely to cause serious harm. Some specific harms we consider include:

- Threats to physical safety of a person or group
- Incitement of abusive behavior to a person or group
- Risk of mass violence or widespread civil unrest
- Risk of impeding or complicating provision of public services, protection efforts, or emergency response
- Threats to the privacy or to the ability of a person or group to freely express themselves or participate in civic events, such as:
    - Stalking or unwanted and obsessive attention
    - Targeted content that aims to harass, intimidate, or silence someone else's voice
    - Voter suppression or intimidation

We also consider the time frame within which the content may be likely to impact public safety or cause serious harm, and are more likely to remove content under this policy if immediate harm is likely to result.

Tweets with misleading media that are not likely to result in immediate harm but still have a potential to impact public safety, result in harm, or cause widespread confusion towards a public issue (health, environment, safety, human rights and equality, immigration, and social and political stability) may be labeled to reduce their spread and to provide additional context.

While we have other rules also intended to address these forms of harm, including our policies on violent threats, civic integrity, COVID-19 misleading information, and hateful conduct, we will err toward removal in borderline cases that might otherwise not violate existing rules for Tweets that include misleading media.

What is not a violation of this policy

We seek to protect public conversation surrounding various issues. Media often accompany these conversations and encourage further discourse. In the absence of other policy violations, the following are generally not in violation of this policy:

- **Memes or satire,** provided these do not cause significant confusion about the authenticity of the media;
- **Animations, illustrations, and cartoons,** provided these do not cause significant confusion about the authenticity of the media.
- **Commentary, reviews, opinions, and/or reactions.** Sharing media with edits that only add commentary, reviews, opinions, or reactions allows for further debate and discourse relating to various issues and are not in violation of this policy.
- **Counterspeech.** We allow for direct responses to misleading information which seek to undermine its impact by correcting the record, amplifying credible information, and educating the wider community about the prevalence and dynamics of misleading information.
- **Doctored or fake Tweets, social media posts, or chat messages.** Due to the challenges associated with conclusively verifying whether an alleged Tweet, post, or message existed, we generally do not enforce on doctored or fake Tweets, social media posts, or chat messages under this policy.

Who can report violations of this policy?

We enforce this policy in close coordination with trusted partners, including our partnership with AP and Reuters (https://blog.twitter.com/en_us/topics/company/2021/bringing-more-reliable-context-to-conversations-on-twitter), other news agencies, public health authorities, and governments. Our team has open lines of communication with various partners to consult and get various media and claims reviewed.

In Australia, South Korea, and the US, Twitter has begun testing (https://twitter.com/TwitterSafety/status/1427706890113495046?s=20) a new reporting feature that will allow users to report Tweets that seem misleading. As part of the experiment, the phrase “It’s misleading” will appear as an option when you select **Report an issue**.

What happens if you violate this policy?

The consequences for violating our synthetic and manipulated media policy depends on the severity of the violation.

Tweet Deletion
For high-severity violations of the policy, including misleading media that have a serious risk of harm to individuals or communities, we will require you to remove this content.

Labeling
In circumstances where we do not remove content which violates this policy, we may provide additional context on Tweets sharing the misleading media where they appear on Twitter. This means we may:

- Apply a label and/or warning message to the Tweet
- Show a warning to people before they share or like the Tweet;
- Reduce the visibility of the Tweet on Twitter and/or prevent it from being recommended;
- Turn off likes, replies, and Retweets; and/or
- Provide a link to additional explanations or clarifications, such as in a curated landing page (Twitter Moments) or relevant Twitter policies.

In most cases, we will take a combination of the above actions on Tweets we label. We prioritize producing Twitter Moments in cases where misleading content on Twitter is gaining significant attention and has caused public confusion on our service.

Account locks
If we determine that an account has advanced or continuously shares harmful misleading narratives that violate the synthetic and manipulated media policy, we may temporarily reduce the visibility of the account or lock or suspend the account.

If you believe that your account was locked or suspended in error, you can submit an appeal (https://help.twitter.com/forms/general?subtopic=suspended).

Additional resources

Learn more about our work and how we build rules to fight misleading media here (https://blog.twitter.com/en_us/topics/company/2020/new-approach-to-synthetic-and-manipulated-media). Learn more about our range of enforcement options and our approach to policy development and enforcement.