Coordinated harmful activity

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

Coordinated harmful activity

Overview

January 2021

The Twitter Rules (https://help.twitter.com/rules-and-policies/twitter-rules) exist to ensure that people can participate in the public conversation freely and safely. In some cases, we identify groups, movements, or campaigns that are engaged in coordinated activity resulting in harm on and off of Twitter. We evaluate these groups, movements, or campaigns against an analytical framework, with specific on-Twitter consequences if we determine that they are harmful. This article explains how we perform these assessments, and what happens when we've identified that a group is engaged in coordinated harmful activity.

Coordinated harmful activity is an actor-level framework, meaning we assess groups, movements, and campaigns and then take enforcement action on any accounts which we identify as associated with those entities. In order to take action under this framework, we must find both evidence that individuals associated with a group, movement, or campaign are engaged in some form of coordination and that the results of that coordination cause harm to others.

How do we define coordination?

Under this framework, we assess two aspects of coordination: **technical** and **social**.

- **Technical coordination** refers to the use of specific, detectable techniques of platform manipulation to engage in the artificial inflation or propagation of a message or narrative on Twitter. For example, a single individual operating multiple accounts in order to Tweet the same message from all of those accounts would constitute technical coordination. The platform manipulation and spam (https://help.twitter.com/rules-and-policies/platform-manipulation) rules describe the different forms of prohibited technical coordination on Twitter.
- **Social coordination** refers to on- or off-Twitter coordination among a group of people to amplify or propagate a specific message. For example, a group of people using a messaging application to organize a campaign to each Tweet about an issue at the same time would constitute social coordination. Another form of social coordination would be an individual using Twitter to incite their followers to say or do a specific thing, such as reply to another person with abusive messaging — a practice referred to as "dogpiling."

All forms of technical coordination are prohibited under the Twitter Rules. If we can prove with sufficient evidence (such as technical linkages between accounts) that an individual or group has engaged in technical coordination, we will always take enforcement action — typically by permanently suspending the accounts involved.

Not all forms of social coordination constitute a violation of the Rules. We are unlikely to apply this framework to groups, movements, and campaigns involved in activism, speaking truth to power, engaging in awareness campaigns, or other similar activities, unless we can establish evidence of harm.

How do we define harm?

Each of our policies is intended to prevent, mitigate, and/or respond to a specific, known harm. There are a wide variety of potential harms that may be linked to online behavior; in evaluating harm under this framework, we look at three principal forms of harm: physical, psychological, and informational. These harms are reflected across the Twitter Rules.

Physical

Harm describing a physical negative effect on someone or something, or on social groups or society. This includes harmful behavior captured in our policies on violence (https://help.twitter.com/rules-and-policies/violent-threats-glorification), terrorism and violent extremism (https://help.twitter.com/rules-and-policies/violent-groups), child sexual exploitation (https://help.twitter.com/rules-and-policies/sexual-exploitation-policy), abuse and harassment (https://help.twitter.com/rules-and-policies/abusive-behavior), hateful conduct (https://help.twitter.com/rules-and-policies/hateful-conduct-policy), and suicide or self-harm (https://help.twitter.com/rules-and-policies/glorifying-self-harm).

Psychological

Harm which focuses on an individual and their mental and emotional wellbeing and psyche, or on the wellbeing of social groups or society. This includes harmful behavior captured in our policies on abuse and harassment (https://help.twitter.com/rules-and-policies/abusive-behavior), hateful conduct (https://help.twitter.com/rules-and-policies/hateful-conduct-policy), suicide or self-harm (https://help.twitter.com/rules-and-policies/glorifying-self-harm), private information (https://help.twitter.com/rules-and-policies/personal-information), and non-consensual nudity (https://help.twitter.com/rules-and-policies/intimate-media).

Informational

Harm that adversely impacts the ability for an individual to access information fundamental to exercising their rights, or that significantly disrupts the stability and/or safety of a social group or society including medical mis-information i.e. COVID-19. This includes harmful behavior captured in our policies on civic integrity (https://help.twitter.com/rules-and-policies/election-integrity-policy), synthetic and manipulated media (https://help.twitter.com/rules-and-policies/manipulated-media), and platform manipulation and spam (https://help.twitter.com/rules-and-policies/platform-manipulation).

Additionally, we take into consideration the severity of any harms we identify. The severity can be low, moderate, or high.

Low

- The harms caused by individuals/supporters are not documented, or are limited in number/frequency, and/or are not severe in nature
- There are little to no sources that confirm that individuals/supporters linked to this group have caused harm, or what harm they have caused
- If left unchecked, the likelihood is almost nonexistent that harm will be caused

Moderate

- The harms caused by individuals/supporters are documented, but may not be particularly severe and/or may be smaller in number/frequency
- There are multiple readily-available, credible sources that establish that individuals/supporters linked to this group have caused harm
- If left unchecked, the likelihood is moderate that more harm will be caused

High

- The harms caused by individuals/supporters are extreme, either in number/frequency or severity
- There are numerous sources that confirm that individuals/supporters have caused extreme harm (frequency or severity), as well as numerous examples of what harm they've caused
- If left unchecked, the likelihood is almost certain that more harm will be caused

How do we use our framework to assess harm?

We use our framework to assess and understand the impact of a group's activity, both on- and off-platform, across the 3 types of harm and severity levels listed above. When assessing the harm that may have been caused by members of a group, we look for evidence of harms that have occurred both on and off Twitter. Typically, a group engaged in activity that has a high severity of harm in at least one category, or multiple moderate severities of harm, will be designated harmful under this framework.

For these assessments, we ask ourselves several questions, including:

- Is this an identifiable group with a unified purpose, or a one-time campaign with a clear goal?
- In what ways does the group or campaign's on- or off-Twitter behavior constitute coordinated activity?
- In what ways have individuals associated with this group caused harm across our 3 categories?
- What are the risks of applying this policy towards this activity, including potential impacts on speech or associational rights?

We keep in mind that examples of harm may differ from actor to actor, from group to group, and across geographies and cultures. Different harms may come in conflict with each other or with the perceived benefits or rights associated with the activity. When such conflicts arise, we prioritize physical and psychological safety, and therefore are most likely to intervene where the activity we've identified leads to physical or psychological harm to others.

Enforcement against coordinated harmful activity

Under this framework, we use an assessment of both coordination and harm, as defined above, to designate a group, movement, or campaign as engaged in Coordinated Harmful Activity. We review these actor-level designations on a regular basis, and may modify or withdraw this designation based on changes in a group's on- and off-Twitter behavior.

When we determine that a group, movement, or campaign meets the criteria for designation as engaged in Coordinated Harmful Activity, we may take the following actions on Tweets and/or accounts which we identify as associated with the group:

- Limiting the visibility of Tweets and/or accounts in search results, replies, and on timelines
- Preventing the recommendation of Tweets and/or accounts in email or in-product recommendations
- Preventing trends from appearing which are directly linked with groups engaged in coordinated harmful activity
- Suspending accounts whose primary use is to propagate and/or encourage engagement in the identified coordinated harmful activity

We identify accounts associated with a group on an ongoing and rolling basis, using a mixture of technology and human review, taking into consideration things such as recent on Twitter behavior or profile information. Limitations on the visibility of Tweets and/or accounts resulting from this framework expire automatically after a limited period of time, but may be re-applied (either manually or automatically) if we determine that the accounts have continued to engage in behavior associated with the designated group.

Any account engaged in coordinated harmful activity may also be subject to enforcement action under the Twitter Rules. Our enforcement actions for violations of our rules include, but are not limited to:

- Requiring the removal of specific Tweets which violate the Twitter Rules
- Suspension of accounts engaged in severe or repeated violations of the Twitter Rules

Learn more about our range of enforcement options (https://help.twitter.com/rules-and-policies/enforcement-options) and our approach to policy development and enforcement (https://help.twitter.com/rules-and-policies/enforcement-philosophy).

Share this article