Twitter Transparency Report for the Regulation (EU) 2021/784 on addressing the dissemination of terrorist content online June - December 2022

Twitter was founded on a commitment to transparency. This commitment is part of our effort to serve the public conversation and to increase its collective health, openness, and civility around the world.

Twitter 2.0 - Our continued commitment to the public conversation: We are a new company embarking on a new chapter, but our steadfast commitment to the public conversation has not changed. We remain committed to providing a safe, inclusive, entertaining, and informative experience for everyone. We will continue to be transparent as we move through this transition period. And we will listen to the people who make Twitter what it is: the town square of the internet.

As part of this mission, we have a zero-tolerance approach to terrorists and violent extremists seeking to exploit our platform to further their aims. Internally, we aggressively fight online violent extremist activity and have heavily invested in technology and tools to enforce our policies, as well as in our external work with industry partners through the Global Internet Forum to Counter Terrorism (GIFCT) and Tech Against Terrorism to share information, knowledge, and best practices.

We take our responsibility seriously to combat terrorism and violent extremism online, but realise that it requires far more than simply identifying and removing content. It requires a collaborative effort from governments, industry, and civil society.

The Twitter Rules prohibit violent threats and the promotion of violent extremism. Specifically, you may not threaten terrorism and/or violent extremism, nor promote violent and hateful entities. Additionally, there is no place on Twitter for violent and hateful entities, including (but not limited to) terrorist organisations, violent extremist groups, perpetrators of violent attacks, or individuals who affiliate with and promote their illicit activities.

In this first iteration of our report that provides information about the EU Regulation on addressing the dissemination of terrorist content online ("TCO"),[1] we provide information from June 7, 2022 through December 31, 2022. We will continue to publish this report yearly on our Transparency Report page under "Reports and "Other".

[1] REGULATION (EU) 2021/784 OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL of 29 April 2021 on addressing the dissemination of terrorist content online.

Technical and Operational Measures and Capacities

General Information

Reporters within the EU can submit requests for removal of potential terrorist content in a number of ways, which are described in our Help Center page about Violent and Hateful Entities[2]. Law enforcement and government requests can also submit legal requests through our Legal Request Submissions site[3].

Operational and technical measures

We continue to aggressively fight online violent extremist activity and have heavily invested in technology and tools to enforce our policies, as well as in our external work with industry partners notably through the GIFCT and Tech Against Terrorism to share information, knowledge, and best practices. We are continuing to invest in automation to maximise our proactivity and efficiency. This includes network and behavioural analysis, technical signal analysis, keyword normalisation, and text scraping of media to automatically surface accounts that we suspect are affiliated with violent and hateful entities.

Measures to address the reappearance of previously removed content

We employ automated methods to detect and prevent violative content from being re-shared on the platform. This includes labelling violative hashes, as well as the aforementioned signals to find recidivist accounts. In addition, we may also carry out manual checks for such content.

Decision criteria and review procedure

Each request we receive is reviewed under Rules and Policies. Any content or account we determine to violate Twitter's Rules and Policies may be removed or suspended from the service. If we receive a removal order under TCO, this content may be withheld[4] in the EU as a consequence.

Organisation, Team Resources, Expertise, Training and Support

Description of the team

Twitter has built a specialised team made up of individuals who have received specific training in order to handle reports about terrorist content. This team consists of different tier groups, with higher tiers consisting of more senior, or more specialised, individuals. In the case that a more detailed investigation is required, content reviewers can escalate reports to experienced safety and legal specialists. These individuals take appropriate action after

[2] https://help.twitter.com/en/rules-and-policies/violent-entities.
[3] https://help.twitter.com/en/rules-and-policies/twitter-law-enforcement-support.
[4] https://help.twitter.com/en/rules-and-policies/tweet-withheld-by-country.

carefully reviewing the report and available context in close detail. Reports can also be discussed with in-house legal counsel.

Furthermore, all teams involved in solving these reports closely collaborate with a variety of other safety teams at Twitter who focus on safety rules and policies, site integrity, or rules and policies related to cybercrime. This cross-team effort is particularly important in the aftermath of tragic events, such as violent attacks, to ensure alignment and swift action on violative content.

Training and support

All team members working on alleged terrorism reports, i.e. all employees hired by Twitter as well as vendor partners working on these reports, are trained and retrained regularly on identifying and assessing terrorist content, including sessions on cultural and historical context. Initially when joining the team at Twitter, each individual follows an onboarding program and receives individual mentoring during this period.

Updates about significant current events or rules and policy changes are shared with all content reviewers, to give guidance and facilitate balanced and informed decision making. In the case of rules and policy changes, all training materials and related documentation is updated.

Removal orders

Twitter has put in place systems and partially automated tooling solutions, in order to expeditiously handle removal orders that Twitter may receive from authorised entities under the EU Regulation. As required, human review of these reports will be conducted within 24 hours of the initial report. Content that is found to be in violation of our policy on the TCO will be made locally unavailable, or “withheld”, in the European Union within one hour of receipt of a valid and properly scoped legal request.[5]

Methods for the automated detection of terrorist content

A vast majority of all accounts that are suspended for the promotion of terrorism are proactively flagged by a combination of technology and other purpose-built internal proprietary tools. You can learn more about our commitment to eradicating terrorist content and the actions we’ve taken here. Our continued investment in proprietary technology is steadily reducing the burden on people to report to us.

As a first line of defence, Twitter employs a combination of heuristics and machine learning algorithms to automatically detect terrorist content and accounts that may violate the Twitter rules and policies.

[5] To date, Twitter has not received any removal orders under the TCO.

Heuristics are typically utilised to react quickly to new forms of violations that emerge on the platform. They may include patterns of text or behaviour that may be typical of a certain category of violations.

Machine learning algorithms vary in complexity and based on application. For automated detection of violations of policies in tweets, we use combinations of natural language processing models, image processing models and other sophisticated machine learning methods. Inputs to the models can include the text within the tweet itself, the images attached to the tweet, and other features. For example, to detect images containing sensitive content such as gore or nudity, we use machine learning models that work with images as inputs.

Heuristics and machine learning models are trained on thousands of data points with labels (e.g. violative or not) generated by trained human content reviewers. Training data comes from both the cases reviewed by our content moderators, a random sample and various other samples of content from the platform.

Automated policy enforcement undergoes rigorous testing before being applied to the consumer product and operating at Twitter scale; and once deployed, regular checks occur to ensure expected performance standards are being met. Before any given model is launched to the platform, we verify its detection of policy violating content or behaviour by drawing a statistically significant test sample and performing item-by-item human review. Reviewers have expertise in the applicable policies and training by our Policy teams to ensure the reliability of their decisions. During this testing phase we also calculate the expected volume of moderation actions a given automation is likely to perform in order to set a baseline against which we can monitor for anomalies in the future (we call this “sizing”). Human review helps us to confirm that automations achieve a level of precision, and sizing helps us understand what to expect once the automations are launched. Once reviewers have confirmed that the detection meets an acceptable standard of accuracy, we consider the automation to be ready for launch. Once launched, automations are monitored dynamically for performance and health. If we detect anomalies in performance (for instance, significant spikes or dips against the volume we established during sizing), Engineering, Data Science and Policy teams revisit the automation to diagnose any potential problems, changing the automations as appropriate.

Our current methods of surfacing potentially violating content for review include deploying a range of internal tools prior to any reports filed. We commit to continuing to invest in technology that improves our capability to detect and remove e.g. terrorist and violent extremist content online, including the extension or development of digital fingerprinting and AI based technology solutions. Our participation in multi stakeholder communities, such as Global Internet Forum to Counter Terrorism (GIFCT) and the Christchurch Call to Action, helps to identify emerging trends in how terrorists and violent extremists are using the Internet to promote their content and exploit online platforms.

Required Data

The data set out below is required under the TCO for the reporting period June 6, 2022 to December 31, 2022.

	Number of instances	
Removals[6] following specific measures	0	
Removals following removal orders	0[7]	
Removal orders where the content has not been removed	0	
Complaints for reinstatement of removed content submitted	7,270	
Complaints for reinstatement of removed content granted	401	
Reinstatements as a result of administrative or judicial review proceedings	0	
	Number of proceedings	**Outcome**
Administrative or judicial review proceedings brought by Twitter	0	N/A

	Number of instances
Proactive referrals to law enforcement under Art. 14(5) TCO	3

[6] Removals in this section mean both removal and disabling of access.
[7] To date, Twitter has not received any removal orders under the TCO.