Twitter

Bundestag Digital Committee - Follow up Questions

3 April 2023

Bundestag Digital Affairs Committee
Platz der Republik 1
11011 Berlin
Germany

Deutscher Bundestag
Ausschuss für Digitales

Ausschussdrucksache
20(23)143
03.04.2023

Twitter International Unlimited Company
Global Government Affairs
One Cumberland Place
Fenian Street Dublin 2 D02 AX07
Ireland

April 3rd 2023

Dear Members of the Bundestag Digital Committee,

We hope this letter finds you well and thank you again for the opportunity given to Twitter to intervene at the hearing organised by your Committee.

On behalf of Twitter, we would like to address the questions you have raised concerning our content moderation policies and practices. We appreciate your interest in understanding our efforts to ensure a safe, inclusive, and transparent platform for users in Germany and around the world.

In the interest of clarity and organization, we have provided responses to your questions below, and grouped our answers in four different categories :

1. Twitters' content moderation policies and enforcement measures
2. The company's transition
3. Twitter's compliance's work with the EU Digital Services Act (DSA) and local laws
4. Twitters' cooperation with law enforcement authorities

We hope that you will find these answers helpful and that they will adequately complete the information we provided during the hearing. We would like to stress that we remain available for any follow-up questions in case it is needed.

Twitter

Bundestag Digital Committee - Follow up Questions

3 April 2023

1. Twitter's content moderation policies and enforcement measures

General policies and enforcement actions :

Twitter's purpose is to serve the public conversation. Violence, harassment and other similar types of behavior discourage people from expressing themselves, and ultimately diminish the value of global public conversation. Our rules are to ensure all people can participate in the public conversation freely and safely.

Twitter has three categories of rules for safety, privacy and authenticity. The Twitter Rules and Policies are publicly accessible on our Help Center, and we are making sure that they are written in an easily understandable way. We are also keeping our Help Center regularly updated anytime we are modifying our rules.

Additionally, you will find explanations in our Help Center on our policy development process and rules enforcement philosophy. Creating a new policy or making a policy change requires in-depth research around trends in online behavior, developing clear external language that sets expectations around what's allowed, and creating enforcement guidance for reviewers that can be scaled across millions of Tweets.

We gather input from around the world so that we can consider diverse, global perspectives around the changing nature of online speech, including how our rules are applied and interpreted in different cultural and social contexts. We then test the proposed rule with samples of potentially abusive Tweets to measure the policy effectiveness and once we determine it meets our expectations, build and operationalise product changes to support the update. Finally, we train our global review teams, update the Twitter Rules, and start enforcing the relevant policy.

When it comes to the enforcement of our rules, we empower people to understand different sides of an issue and encourage dissenting opinions and viewpoints to be discussed openly. This approach allows many forms of speech to exist on our platform and, in particular, promotes counterspeech: speech that presents facts to correct misstatements or misperceptions, points out hypocrisy or contradictions, warns of offline or online consequences, denounces hateful or dangerous speech, or helps change minds and disarm.

Thus, context matters. When determining whether to take enforcement action, we may consider a number of factors, including (but not limited to) whether:

- The behavior is directed at an individual, group, or protected category of people;
- The report has been filed by the target of the abuse or a bystander;
- The user has a history of violating our policies;
- The severity of the violation;
- The content may be a topic of legitimate public interest.

When we take enforcement actions, we may do so either on a specific piece of content (e.g., an individual Tweet or Direct Message) or on an account. We may employ a combination of these options. In some instances, this is because the behavior violates the Twitter Rules. Other times, it may be in response to a valid and properly scoped request from an authorized entity in a given country.

To enforce our rules, we are using a combination of machine learning and human review. Our systems are able to surface content to human moderators who use important context to make decisions about potential rule violations. This work is led by an international, cross-functional team with 24-hour coverage and the ability to cover multiple languages. We also have an appeals process for any potential errors that could occur.

For manifestly illegal content such as, for instance, child sexual exploitation and terrorist content, we are relying more and more on technology, which enables us to scale enforcement of our rules for this kind of content, that has no place on Twitter.

Of the unique Twitter accounts we suspended for child sexual exploitation content during our last reporting period (July to December 2021), more than 91% were surfaced proactively and removed by a combination of technology solutions, including PhotoDNA and internal proprietary technical tools. These tools and initiatives support our efforts in surfacing potentially violative content for further review and, if appropriate, removal.

During the past year, the overwhelming majority (approximately 96%) of accounts violating our Terrorist Organisation policy were suspended proactively using our internal proprietary tools and automation and, since August 2015, we have suspended **more than**

1.8M accounts for violations related to the promotion of terrorism. Although we have been dramatically increasing our proactivity rates, we continue to see an overall downward trend in the number of violating accounts. This likely reflects the changing behavior patterns of bad actors and improvements in our defenses, for example making it harder for bad actors to compromise accounts.

In our last transparency report published on 28 July 2022 and covering the period between July-December 2021, we announced that **we suspended 33,693 unique accounts** for violations of the policy during this reporting period. **Of those accounts, 92% were proactively identified and actioned.** Over this reporting period, we observed a **25% decrease** in the number of accounts actioned for violating our terrorism/violent extremism policy.

Moderation of content :

Our rules have largely remained unchanged in relation to the company's transition and our Trust & Safety team continues its diligent work to keep the platform safe and to challenge those who break Twitter's rules. Also, Twitter remains strongly committed to content moderation.

By way of an illustration, we have been fighting more aggressively against Child Sexual Exploitation (CSE) on our service, as our last figures show. Today more than ever, Twitter's number one priority is tackling child sexual exploitation and we have absolutely zero tolerance to this kind of content which corresponds to the more severe violation of our rules. In the majority of cases, the consequence for violating our child sexual exploitation policy is immediate and permanent suspension. In addition, violators will be prohibited from creating any new accounts in the future. When Twitter is made aware of content depicting or promoting child sexual exploitation, including links to third party sites where this content can be accessed, they will be removed without further notice and reported to the National Center for Missing & Exploited Children (NCMEC).

Our last enforcement figures show that :

- Over the week-end of 3rd-4rd December 2022, we took action on almost **44 000 accounts**, **including 1300 profiles** that had been trying to evade our detection

methods. One such measure that we've recently implemented has reduced the number of successful searches for known CSE patterns **by over 99% since December 2022.**

- In January, we suspended roughly **404 000 accounts that created, distributed, or engaged with this content**, which represents a **112% increase** in CSE suspensions since November 2022.

Another example is our work to fight the prevalence of hate speech on the platform. Recently, we have been partnering with an independent third-party, Sprinklr, to evaluate the reach of hate speech on Twitter. Sprinklr's AI-powered model found that the reach of hate speech on Twitter remains at very low levels (even lower than our own model quantified).

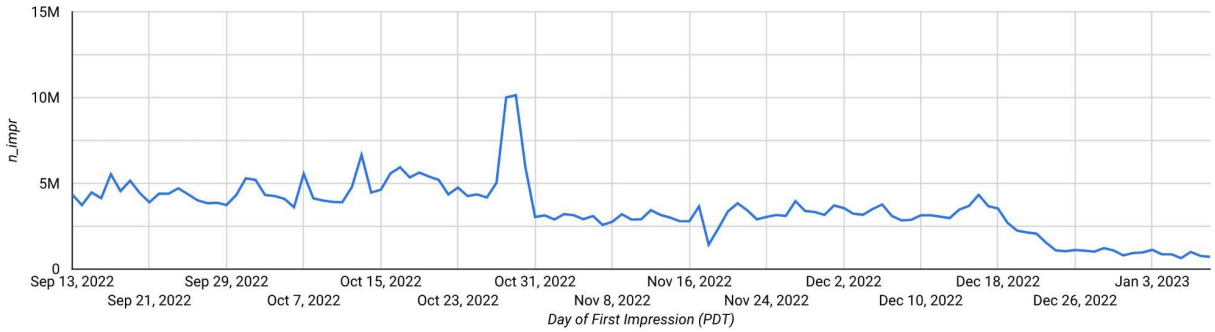
To quantify hate speech, Twitter & Sprinklr start with 300 of the most common English-language slurs. We count not only how often they're tweeted but how often they're seen (impressions). Our models score slur Tweets on "toxicity," the likelihood that they constitute hate speech and our focal metric is hate speech impressions, not the number of Tweets containing slurs. Most slur usage is not hate speech, but when it is, we work to reduce its reach. **Sprinklr's analysis found that hate speech receives 67% fewer impressions per Tweet than non-toxic slur Tweets.**

According to our own analysis, we continue to see a decrease in the reach of this content, **with these impressions accounting for less than 0.01% of all English-language Tweet impressions on the platform** since our last update. The chart below shows you the number of hate speech impressions in English speaking languages on Twitter between September 2022 and January 2023 :

Twitter

Bundestag Digital Committee - Follow up Questions

3 April 2023



Most recent policies changes :

Twitter's rules have remained largely unchanged. Since the transition, we have updated few of our policies, and updated our Help Center accordingly :

- **Reinstatement of accounts** : One of the main policy changes that Twitter has been doing in the enforcement of our policies is that a small number of accounts previously suspended for non-criminal activity have been restored. However, we are still keeping tweets that have harmful content or abusive language filtered so they do not become automatically visible to the world once we reinstate the account. We will continue to restore accounts that were suspended for non severe rules violations and move towards using visibility filtering and desamplification as a mechanism to reduce the reach of content without having to remove it entirely from the platform. This is consistent with a gradual move that the company has been doing on content moderation, and which consist in using a wider range of remediation options, tailored to the severity of the violation. As a result, people that are actively searching for the content will be able to find it but Twitter will not be amplifying it, striking a better balance between safety and free speech. It goes without saying that if these accounts break the law or severely break our rules, they will be suspended again, with a possibility for permanent suspension.

- **Violent Speech Policy Update** : In February 2023, we launched an updated Violent Speech policy, which prohibits violent threats, wishes of harm, glorification of violence, and incitement of violence. Healthy conversations can't thrive when violent speech is

used to deliver a message. As a result, we have a zero tolerance policy towards violent speech in order to ensure the safety of our users and prevent the normalisation of violent actions. In most cases, we will suspend any account violating this policy. For less severe violations, we may require the user to delete the content before they can access their account again.

As you are aware, Twitter Trust & Safety Council has been deprecated. We indeed announced that Twitter may form a Content Moderation Council in the future. This is a format that we have been discussing with civil society groups, and we are still considering putting it in place. Obviously, we continue to be in regular contact with Civil Society Organizations (CSOs) on a bilateral basis and really value the cooperation with them and their expertise. In addition, as part of our compliance work with the DSA, we will be putting in place a priority reporting channel for CSOs that will be recognized as Trusted Flaggers under Article 22 of the DSA.

Cooperation with the research community and API access :

Twitter's cooperation with the research community is essential and we intend to continue working with researchers and academics going forward. For different reasons, we are currently pausing our free access to the Twitter API, both v2 and v1.1. A paid basic tier will be available instead. In the past, academics were getting access through a range of different paths, but there actually was never a bespoke 'academic' API. Some institutions were already paying, and in some cases it was a nominal fee agreed more than a decade ago. There were key challenges associated with this model :

- Twitter incurs the cost for the infrastructure and the compliance around access to the data.
- Others were using different APIs, and sometimes unfortunately, in a way that violated Twitter's policies (e.g. using multiple API keys to access more data) or were doing research without proper review by Twitter.
- It was very difficult to audit who had access to the APIs, so not only some uses violated the policies, but also used the data for commercial purposes.

We believe that it is fair that people using Twitter data make a contribution to Twitter's infrastructure and compliance costs. There are many third parties who are able to offer access to Twitter data and it is highly unlikely a research project needs access to the full commercial

firehose. Academic access can be limited in type and in periods of time, such as for instance a period established in the research project that will be pre-approved by Twitter. Depending on the project, access may not need to be in real time, meaning access could be given to data from a previous period of time, in order to avoid potential misuse of the data, for commercial real time social listening purposes. **In addition, in some cases and especially in the European Union, the DSA will mandate Twitter to give access to data to vetted research institutions and CSOs**.

Community based approach to fight misinformation

The centerpiece of Twitter's new approach to offering context and surfacing credible information is Community Notes. We believe that this product represents a fundamental shift in how we mitigate disinformation.

Community Notes aims to create a better-informed world by empowering people on Twitter to collaboratively add helpful notes to Tweets that might be misleading. Contributors can leave notes on any Tweet and if enough contributors from different points of view rate that note as helpful, the note will be publicly shown on a Tweet.

We believe that Community Notes is an inherently scalable and localised response to the challenge of disinformation. It will be particularly relevant in the case of AI-powered images as it will enable users to quickly debunk such misleading content. By making this feature an integral and highly visible part of the Twitter product, and by ensuring that the user interface is simple and intuitive, we are investing in a tool that can be truly global in its application. It also reduces our reliance on forms of content moderation that are more centralised, manual and bespoke; or which require intensive and time-consuming interactions with third parties.

In practice, contributors write and rate notes. Contributors are people on Twitter who sign up to write and rate notes. The more people that participate, the better the program becomes.

- Only notes rated helpful by people from diverse perspectives appear on Tweets : In other words, Community Notes doesn't work by majority rules. To identify notes that are helpful to a wide range of people, notes require agreement between contributors who have sometimes disagreed in their past ratings. This helps prevent one-sided ratings.

Twitter

Bundestag Digital Committee - Follow up Questions

3 April 2023

- Twitter doesn't choose what shows up, the community does : Twitter doesn't write, rate or moderate notes (unless they break the Twitter rules.) We believe giving people a voice to make these choices together is a fair and effective way to add information that helps people stay better informed.
- Open-source and transparent : It's important for people to understand how Community Notes works, and to be able to help shape it. The program is built on transparency: all contributions are published daily, and our ranking algorithm can be inspected by anyone.

It is important to keep in mind that Community Notes complements and does not replace the work that our threat disruption unit within our Trust & Safety teams does to keep the platform safe from manipulation, disinformation campaigns and spam.

We have been evaluating the effectiveness of Community Notes so far and found that, according to the results of four surveys run at different times between August 2021 and August 2022, a person who sees a Community Note is, on average, **20-40% less likely to agree with the substance of a potentially misleading Tweet** than someone who sees the Tweet alone. Survey participation ranged from 3,000 to more than 19,000 participants, and the results were consistent throughout the course of the year, even as news and Tweet topics changed. We also see that Community Notes informs sharing behaviour. Analysing our internal data, we've found that a person on Twitter who sees a note is, on average, **15-35% less likely to choose to Like or Retweet a Tweet than someone who sees the Tweet alone.** In our most recent survey, notes were found to be informative regardless of a person's self-identified political party affiliation — there was no statistically significant difference in average informativeness across party identification.

All Community Notes contributions are publicly available on the Download Data page of the Community Notes site so that anyone has free access to analyse the data, identify problems, and spot opportunities to make the product better.

At the moment, Community Notes are publicly visible to everyone. Users in the US, the UK, Ireland, Canada, Australia and New Zealand can now contribute to the program. Over the coming months, users in more markets will be able to contribute notes and the product will be

localised further, and notably in European Union countries. We currently have around 20,000 contributors and we aim to expand this number by 10% each week.

Over time, users in any country, writing in any language, should be able to contribute to Community Notes and the most helpful contributions will be surfaced to inform readers. Eventually, we can see a future where attempts to spread disinformation are consistently flagged by conscientious users seeking to share important context and facts with citations. We think that this will help improve critical thinking on the platform and encourage users to be more responsible in their use of the product.

This is an open and transparent process. That's why we've made the Community Notes algorithm open source and publicly available on GitHub, along with the data that powers it so anyone can audit, analyse or suggest improvements.

2. The company's transition :

In any corporate change of ownership, there is a period of transition and this is no different. We are only a few months into the process, and we are please asking our partners and stakeholders to understand that we will still need some time to adjust to new structures and priorities. Unfortunately, Twitter made a loss in 8 of the last 10 years. As a result, the very difficult choice of cost cutting was necessary to protect the service and the business. It's also a recognition of the current economic environment and other technology companies have taken similar measures.

This process will include changes to our leadership and organisation, in addition to new modes of product experimentation and development. We will continue to regularly share information on this process as it develops and remain fully available to our stakeholders and partners along this process.

It is important to note that in the context of the company's transition, the Health team has been one of the least impacted, with staffing levels remaining sufficient to moderate our service and enforce Twitters' rules consistently. In addition, while we have seen departures, we intend to invest in hiring for these teams. The people and agents actually in charge of

moderation are still in place, and we retain strong teams to be able to process law enforcement requests within normal time frames in Europe and notably in Germany. These teams and their work is vital to Twitter and it is essential that they can continue to do their work as well as possible in the context of the company's transition.

Unlike many companies, the public nature of Twitter means that we're doing our work in the open. We're moving quickly, but we remain open to feedback from our partners, decision makers and regulators, and the people on Twitter - indeed, the people using Twitter have always been part of the product and policy development process. Many of Twitter's most iconic features have come from our community of users. We believe this is also part of our policy of placing the user at the center of what we build and do.

When we make mistakes, we will move quickly to acknowledge them and fix them. For example, we quickly responded to issues with impersonation by allowing everyone to report such cases and increased our proactive efforts to help detect new violations of our existing parody, commentary, and fan account policy.

3. Our compliance's work with the DSA and local laws :

Our approach to regulatory compliance remains unchanged. We will respect the laws of the jurisdictions where Twitter is available, including the Digital Services Act (DSA), for which important and intense compliance work is underway. We are regularly meeting with the European Commission and we are consistently going through the process of DSA compliance. As part of this process and its timeline, you will see that we published by mid-February, as required by the DSA, our number of Twitter Average Monthly Active Recipients of Service (AMARS) in the EU, which places us under the category of Very Large Online Platform (VLOP) under the DSA. This is the first step of our DSA compliance work and by mid-June the DSA will fully apply to Twitter.

Like any corporate CEO, Elon has been meeting, and will continue to do so, with policymakers and government representatives to communicate Twitter's view on policy and regulations, including the German Minister for Digital Affairs, whom he met in Twitter's San Francisco

headquarters, and the EU Commission for Internal Market, Thierry Breton, whom he met twice, including for working sessions as part of the DSA compliance work.

Additionally, the Global Government Affairs team continues to engage actively in this work around the world, and, even if we are currently a smaller team, we want to remain available to decision makers, regulators and partners globally.

4. Our cooperation with law enforcement authorities :

Cooperation with law enforcement authorities around the globe is crucial to Twitter. As you know, we also work closely with law enforcement around the world - and Germany is no exception - and we do our best to assist them in identifying users whose content may be in violation of local laws. Any law enforcement authority or agency can find guidelines on our Help Center for law enforcement and can reach out to Twitter using a dedicated form.

Twitter's International Unlimited Company which is headquartered in Dublin, Ireland reviews and processes law enforcement requests for user data. Twitter receives and responds to requests related to user data from EU law enforcement agencies and judicial authorities wherever there is a valid legal process. We have existing processes in place, **including a dedicated online portal for law enforcement**, and expert teams with global coverage across all timezones that review and respond to reports in any language.

Law Enforcement from across the world can use our dedicated portal to submit their legal demands and can request the following information from Twitter :

- Information requests (IRs) : request for user personal and private information (account info, device info and IP data).
- Content removal requests (RRs) : request to remove Twitter content based on Twitter's Terms and Services or local laws.
- Preservation requests : Request to preserve data for 90 days for the purposes of an investigation.
- Emergency requests : Process through which, when there is an imminent threat to life or serious bodily harm, Twitter may disclose user information without receiving a legal process.

Twitter

Bundestag Digital Committee - Follow up Questions

3 April 2023

We regularly train German law enforcement officers and provide recommendations on how to best use this portal to facilitate communication and speed up processes and we have dedicated processes for both information, content removal and data preservation requests. For information requests, we have an emergency request channel. “Emergency request” refers to the process through which, when there is an imminent threat to life or serious bodily harm, Twitter may disclose user information without receiving a legal process.

From July to Dec 2021 - our latest reporting period - Twitter received and responded to 634 Information Requests from Germany. For this period, our compliance rate for information requests - including both IR and ER requests - in Germany **reached around 45%**, slightly above our global compliance rate of 40%. Among these requests, **we received 37 emergency requests and disclosed information in around 27% of cases.**

From July to Dec 2021, Twitter received and responded to **17 Removal Requests from Germany with a 76,4% compliance rate,** up from the previous reporting period.

We accept requests from law enforcement to preserve records, which constitute potentially relevant evidence in legal proceedings. We will preserve, but not disclose, a temporary snapshot of the relevant account records for 90 days pending service of valid legal process.
If law enforcement requires more time to obtain a court order or other process, they can submit a preservation extension request prior to the expiration of the 90 days. From July to Dec 2021, **Twitter received 13 preservation requests from Germany.**

In conclusion, Twitter is committed to maintaining an open and safe platform for users while respecting local laws and regulations. We appreciate the opportunity to address your concerns and are dedicated to continuously improving our policies and processes in response to feedback from users, stakeholders, and regulatory authorities.

We look forward to engaging in further discussions and collaborations with the Bundestag Digital Committee and other relevant stakeholders to enhance the safety, transparency, and inclusiveness of our platform. Should you have any additional questions or require further clarification, please do not hesitate to reach out.

Twitter

Bundestag Digital Committee - Follow up Questions

3 April 2023

Thank you for your attention to this matter.

Mit freundlichen Grüßen,

Twitter