Summary of Integrity Institute Feedback on TikTok Transparency Reports

Submitted March 13, 2023

Community Guidelines Enforcement Report:

Overview

We generally think the transparency center is a good start. It is usable and fairly easy to understand. But the data included in it isn't at parity with other platforms and isn't enough for people outside Tiktok to assess whether Tiktok is responsibly designed and taking responsible steps to mitigate harms on the platform.

- **The data currently provided is not sufficient if TikTok wants to offer meaningful transparency** beyond the surface level. (See detailed suggestions under Question 4.) Some of the main overall suggestions were:
    - **Too many totals, not enough rates**: Overall, the report relies heavily on totals which are not as helpful without understanding what the rest of the platform metrics are.
    - **Report should be paired with (or include) specific, privacy-protecting raw datasets** that let outsiders verify and audit the claims being made (specific suggestions for this data are below under Question 4). Inclusion of such datasets are a prerequisite for meaningful transparency.
- **There is a big delta between what is contained in this report and what Integrity Institute members have put forward as a consensus for transparency, contained in these 2 decks:**
    - Metrics & Transparency
    - Transparency in Ranking and Design
- **Downloadable data format:** A zip file with a separate CSV file for each data table would likely be more useful than the Excel spreadsheet, because those are open formats.

- The section on covert operations is great and we are happy to see the targeted countries and communities included.
- **Overall the data was usable, the charts were helpful, particularly the interactive element**.
- Some specific questions/ areas where more clarity could be helpful:
    - In Video Removal by Policy, clarify if this data includes videos that were later restored?
    - In Removal Rate by Quarter/Policy, clarify if this data includes videos that were later restored?
    - More explanation needed in the "Safety" section: Why are remaining markets (other 10% of removal volume) not included? Some explanation behind this decision could be helpful.

Suggestions to make the report more comprehensive:

- **The scope of this report is limited to content and accounts that were removed from the platform, and the data provided felt surface-level**. There is room to offer more explanation, context or more data on the scale, cause and nature of violating content on the platform.
- On scale:
    - The report addresses some of this by providing the number of video removals and the fraction of removed videos relative to all published.
    - But more information is needed about the reach of the violating content. In particular:
        - Prevalence of impressions on harmful content
        - The reach of harmful content over 7, 30, and 90 day windows
        - The distribution of frequency of exposures for users
            - How many users had 0, 1, 2, 3, 4, 5+ harmful exposures?
        - How many views did the removed videos have before they were removed? How many users viewed them?
        - The time delay between harmful content being posted and moderated
- On cause:
    - Where did the harmful exposures come from - e.g:
        - What percentage of harmful exposures took place in algorithmic feeds?

        - What percentage were from creators the users follow?
        - What percentage of those follows were from a recommendation?
        - What percentage of harmful exposures are on public content?
        - What percentage of harmful exposures are from creators who have previous offenses?
        - How many harmful exposures were from ads?
        - What was the reach of the ads before removal? How many impressions/views?
    - Meaningful transparency includes information on underlying systems, and such data could include:
        - The Top N most important features in the ranking system
            - N should be > 10
            - "Importance" of the features should be assessed using accepted practices for the model design
        - A list of ML models and what they try to predict, with special attention to any ML models that involve predicting user actions
        - The top-line objectives for the ranking systems and their specific definitions
        - Full disclosure if there are any different ranking processes for content topics and how content classifiers impact ranking
        - How you prevent bad actors from "gaming" the ranking systems and make the system adversarially robust
- On nature:
    - Would also be useful to provide content data sets to put removals in context and give auditable set of data to show how enforcement of policies work. E.g.:
        - The Top N pieces of public content (data should include all public data with the content)
        - A random sample of N impressions on public content
            - N should be at least 10,000, preferably released on weekly basis
            - Data should include all public post content
            - Data should include key ML model scores for the content (specifically any engagement predictions)
    - Basic demographic statistics on the viewers of harmful content

    - Additional data could be provided on the targets of harmful content, rather than the viewers
    - Transparency into policies are also important and could include:
        - Core metrics that are used in experimentation processes and their exact definitions
        - An outline of their processes for determining product or ranking changes
            - Specifically for features to reduce harms on the platform
            - And for "normal" features
        - Their process for platform changes around significant events (Elections)
        - Staffing levels on integrity and trust and safety teams
        - Platforms should release how they assess content quality
            - Specifically any quality assessments related to integrity
            - Should include positive definitions of content quality as well as negative
- Additional suggestions by section:
    - **"Security" section**: suggest sharing more data around security (e.g., what percentage of users use 2FA).
    - **Ads:**
        - Why were ads removed (under what policies? Were they harmful, or did they violate other guidelines?)
        - Say more about targets of ads, as well as origins.
        - Unclear what "removed due to account actions" means.
        - Suggest to publish an Ads library (if this already exists, would suggest adding to the Ads section of this report)
    - **Convert influence ops:**
        - **Disclose privacy-preserving datasets of removed operations (content and behavior) to vetted researchers and journalists**; this would help increase trust and improve the research community's ability to track info ops.
        - **Include total views of network and impressions from non-followers** (a sign of how much the particular campaign was able to break through into more organic consumption)

- **Disclose repeat offenders**. (Currently the report mentions when additional accounts from prior ops were taken down, but doesn't disclose which ops).

- **Latest Data:**
  - **Total videos removed/total videos, by quarter**
    - This gives some information on the prevalence of violating content, but no information on the reach (and potential impact).
    - Could also be useful to include some information on the origins of violating videos, e.g., of the videos removed, how many (or what percentage) were removed from accounts with multiple violations?
  - **Total videos removed/restored, by type and quarter**
    - How many videos restored were removed by automation?
    - How many videos were flagged for appeal?
    - What is the rate of restoration to appeal?
  - **Total video removal, by policy**
    - Include data on reach (impressions, views, etc) before removal.
    - Where possible, include some analysis for context (e.g., why might removals of videos of minor safety have changed by -~10% from first to second quarter in 2020)
  - **Total video removal and rates, by sub-policy**
    - Include videos removed by automation
  - **Removal rate, by quarter/policy**
    - Include data on reach (impressions, views, etc) before removal.
  - **Total account removal, by quarter and reason**
    - Would be useful to include data on the reach of these accounts.
    - How many had high follower counts or high engagement?
  - **Removal volume and rates, by country**
    - Possible to start sharing information from all countries? Even smaller markets may have big societal impact.
  - **Fake engagement and Ads Policy Enforcement charts:**

- Add more analysis/context to the these numbers (e.g., possible explanations for why there are spikes in some quarters)

Comments on Government Removal Requests, Information Requests & Intellectual Property Removal Reports:

- **Downloadable data format:** A zip file with a separate CSV file for each data table would likely be more useful than the Excel spreadsheet, because those are open formats.
- **On the Government Remove Requests report:**
  - There was some confusion about the meaning behind "total requests received" in the GRFCR map tab of the Excel file. Our understanding was that each request refers to an instance where a government submits a request, which can include multiple accounts and videos, but we couldn't find this specifically clarified.
  - Suggest to add more context/ case studies to illustrate the kinds of requests and put data in context:
    - **There is no context for the government removal requests**, so it is hard to evaluate or draw any conclusions. For example, breaking out these numbers to give more detail:
      - "Accounts actioned for community guidelines violations" : what category of violation? Under what policy?
      - "Accounts actioned due to local law violations" : could be specific, or break out into legal categories (e.g., copyright)
  - **Collaborate with public organizations working on government requests**. For example, the Lumen database, based at the Berkman Klein Center at Harvard, which collects and analyzes removal requests.
  - The inclusion of case studies illustrating kinds of government requests received would also help add more context and be more illustrative of what is happening.
  - Another possibility is to disclose when possible requests to remove accounts of journalists, human rights defenders, etc.