Notes from USG Meeting with YouTube on Violence and Atrocity Prevention
Pilot re: Bangladesh/Burma
April 20, 2023

~~(SBU)~~ Draft Agenda
~~(SBU)~~ Engagement Framework

Attendees from USAID:

- DRG: Vera Zakem
- ITR/DRG: Garth Willis
- CPS: Amy Malessa
- Asia Bureau: ?

Other USG Attendees:

- State Department and National Security Council

Background from State Department, USAID and National Security Council:

- Recapped the same toplines from Meta call
- Importance of proactive approach to violence and atrocity prevention, intersects with human rights and information integrity
- Importance of empowering USAID staff and partners on-the-ground to prevent and respond to violence
- WH emphasized that State/USAID will take the lead on this initiative moving forward - the goal is to make these engagements more intentional and consistent
- Hope to consolidate our response across the USG so that RFIs/asks do not fall through the cracks moving forward
- Will start with Burma and then move into Bangladesh as pilots - political toplines provided on the context from unclassified reporting

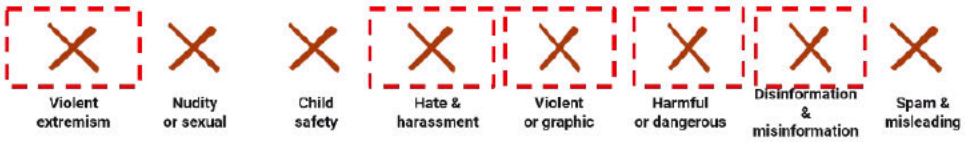
Presentation from YouTube - "How YouTube Works: Key Events Initiative" (see slides):

- Government and Public Affairs Team across Google / YouTube - whole team dedicated to foreign affairs and national security, including staff dedicated to Asia Pacific
- YouTube provided toplines on the penetration of YouTube services - number of videos, channels, etc.
- YouTube has global, guiding principles (e.g. Community Guidelines) to remove, raise, and reward users
- YouTube has content moderation standards - apply equally to all speakers, in all languages, at all times
- Trust and Safety Teams across the globe - locations of staffing not disclosed
- YouTube has a variety of policies to prevent offline harm…

Managing content responsibly

Remove

We remove content that violates our policies, covering 8 core verticals



Confidential & Proprietary | 5

- YouTube policies evolve in the face of geopolitical competition and conflict

Managing content responsibly

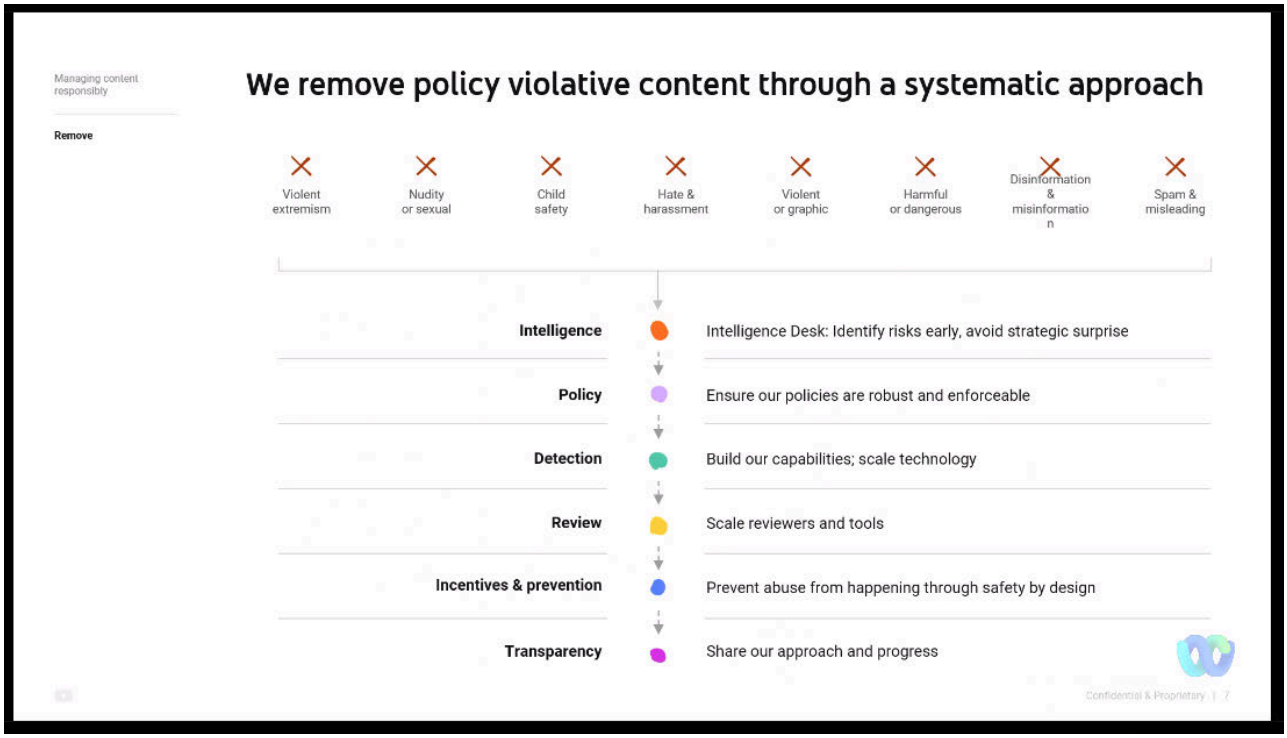
Remove

Combination of policies work together to prevent major violent events

Violent extremism	Violent or graphic	Disinformation & misinformation	Hate & harassment	Harmful or dangerous
• Content produced by violent extremist criminal, or terrorist organizations • Content praising or memorializing prominent terrorist, extremist, or criminal figures in order to encourage others to carry out acts of violence • Content praising or justifying violent acts carried out by violent extremist, criminal, or terrorist organizations • Content aimed at recruiting new members to violent extremist, criminal, or terrorist organizations • Content depicting hostages or posted with the intent to solicit, threaten, or intimidate on behalf of a criminal, extremist, or terrorist organization • Content that depicts the insignia, logos, or symbols of violent extremist, criminal, or terrorist organizations in order to praise or promote them • Content that glorifies or promotes violent tragedies, such as school shootings	• Inciting others to commit violent acts against individuals or a defined group of people. • Encouraging others to go to a particular place to commit violence, to perform violence at a particular time, or to target individuals or groups with violence. • Footage, audio, or imagery involving road accidents, natural disasters, war aftermath, terrorist attack aftermath, street fights, physical attacks, immolation, torture, corpses protests or riots, robberies, medical procedures, or other such scenarios with the intent to shock or disgust viewers. • Footage or imagery showing bodily fluids, such as blood or vomit, with the intent to shock or disgust viewers. • Footage of corpses with massive injuries, such as severed limbs	• Voter suppression: Content aiming to mislead voters about the time, place, means, or eligibility requirements for voting, or false claims that could materially discourage voting. • Candidate eligibility: Content that advances false claims related to the technical eligibility requirements for current political candidates and sitting elected government officials to serve in office. Eligibility requirements considered are based on applicable national law, and include age, citizenship, or vital status. • Incitement to interfere with democratic processes: Content encouraging others to interfere with democratic processes. This includes obstructing or interrupting voting procedures. • Distribution of hacked materials: Content that contains hacked info, the disclosure of which may interfere with democratic processes. • Election integrity: Content advancing false claims that widespread fraud, errors or glitches occurred...	• Age • Caste • Disability • Ethnicity • Gender Identity and Expression • Nationality • Race • Immigration Status • Religion • Sex/Gender • Sexual Orientation • Veteran Status • Victims of a major violent event and their kin	• Extremely dangerous challenges • Dangerous or threatening pranks • Instructions to kill or harm

Confidential & Proprietary | 6

- This is how policies are implemented…



- YouTube policies are constantly under review to make sure they are in line with global standards, domestic laws, and able to be enforced
- YouTube uses classifiers are updated in local languages to account for coded language - need to protect political speech, remove hate speech (can be gray space)
- YouTube uses machine learning / artificial intelligence to automate the detection of problematic content that violates global standards - used for both text, images and videos
- YouTube users can appeal violations for further review
- Below is a screenshot to demonstrate the scale of content removal - see transparency report for more information about removals - mostly scams / spam content

Managing content responsibly

Remove

- Intelligence
- Policy
- Detection
- Review
- **Transparency**

Sharing our progress on removing harmful content (Q4 '22)

5.6M+ | Videos removed

6.4M+ | Channels removed

1.9B+ | Comments removed



Source: Based on YouTube Community Guidelines Enforcement Report, Oct-Dec2022

- The next frontier is in raising the good stuff (not taking down the bad stuff)… elevating authoritative sources, legitimate news sources

Managing content responsibly

Raise

Our approach to raising authoritative content

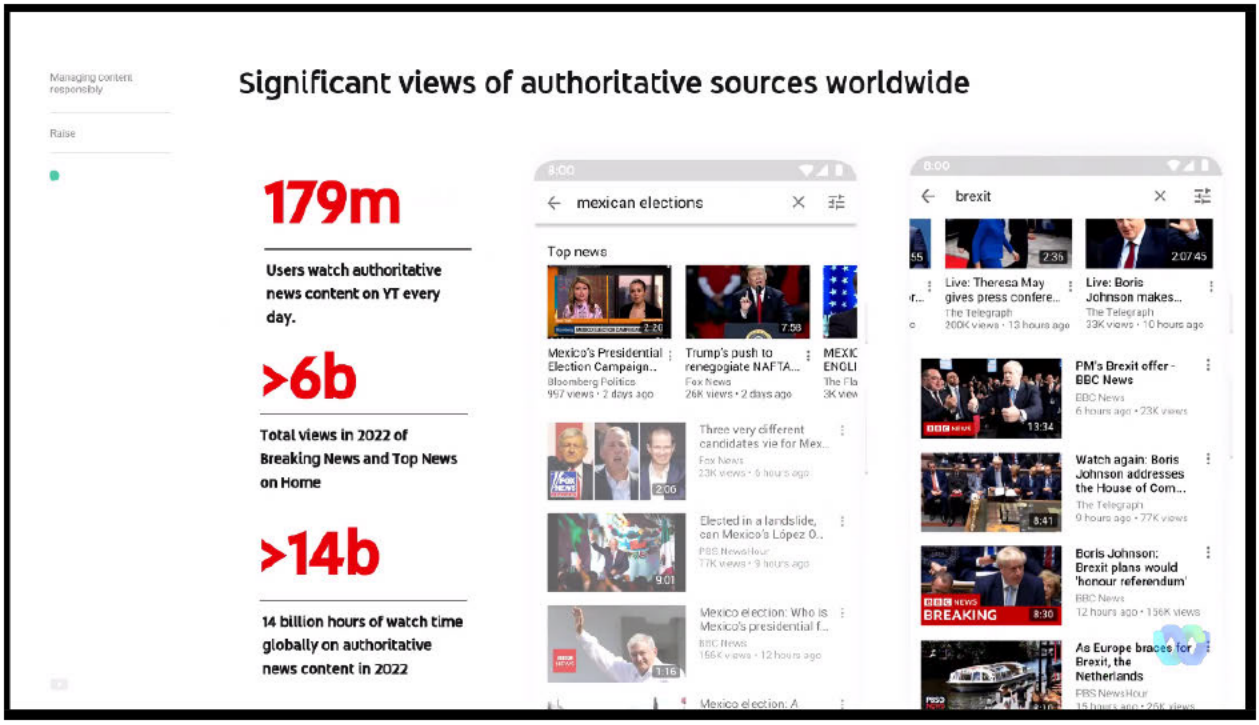


Elevate authoritative sources

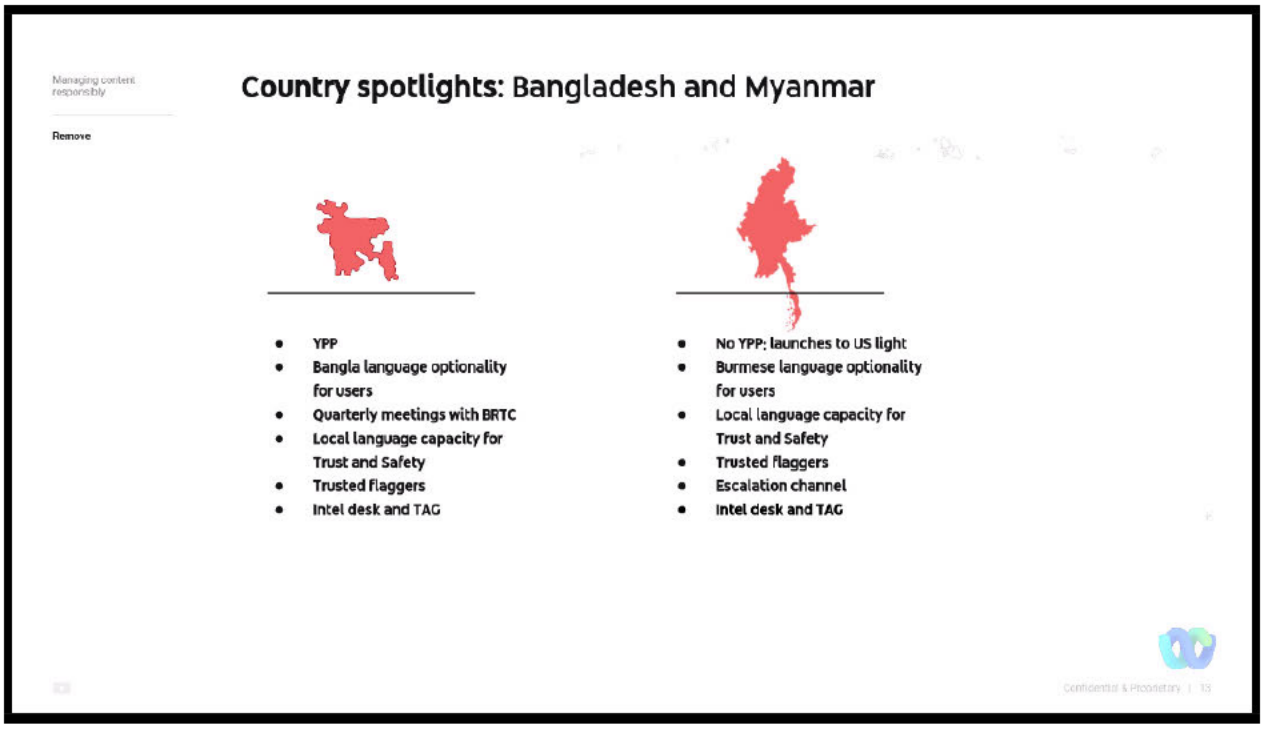


Provide context





Spotlights on Bangladesh and Burma:



- YouTube is launched in Bangladesh, YouTube Partnership Program is operational (can monetize content)

- YouTube has local language UX/UI available in Bangladesh, including a dedicated a local Trust and Safety Team
- YouTube has quarterly meetings with regulator in Bangladesh
- YouTube has rolled out their Trusted Flaggers program in Bangladesh as well
- YouTube also has their Threat Analysis Group (TAG) in Bangladesh
- YouTube is talking internally about how to prepare for elections in Bangladesh, Burma, and Thailand - to help to detect and mitigate threats early
- YouTube has policies related to elections and rule of law - to prevent electoral manipulation and irregularities
- YouTube puts in place a Watch Desk before an election, surge 3rd party content moderators, and scrub for the impersonation of politicians
- YouTube will also surge local language capacity six months before an election
- YouTube is not officially launched in Burma - no YouTube Partnership Program (no monetization)
- YouTube recognizes that it is a problem for civil society, journalists, etc. to not be able to monetize their content in Burma (creators keep the majority of the revenue when monetized)
- YouTube has a light version of YouTube in Burma; Burmese language UX/UI available, but doesn't feature local news, sports, creators, etc. - it's bare bones
- But, YouTube does have local language capacity in Burma
- YouTube also has Trusted Flaggers, a dedicated threat escalation channel, and the more official Threat Analysis Group active in Burma
- YouTube is concerned about channels that have been blocked for violating Terms of Service and are not creating "copycat" channels
- YouTube has a whole system designed to detect - in automated and human ways - users / content who are trying to get around content removals (e.g. circumvention)
- YouTube is seeing much less content related to the dehumanization of the Rohingya than hyperbolic pro-Government, pro-authoritarian content
- YouTube does a lot of work around elections - "Global Election Watchers" - includes Google Search - to help people take part in the democratic process
- YouTube has an election policy related to election results - not miss-stating election results or claiming victory, etc.

State / USAID Questions and Discussion:

- How does YouTube decide which events require additional resources and/or focus?
    - The Intel Team inside YouTube is constantly looking at open source intelligence to understand and respond to threats, inform YouTube resource decisions
    - YouTube had a proactive and thorough approach in Brazil but still couldn't predict the level of violence
    - YouTube has a Major Event Policy - Holocaust, Ukraine, mass/school shootings, etc. - allows YouTube to remove content that claims the event didn't happen
    - Working together the Community Guidelines are flexible enough to respond to unexpected events
- How quickly can YouTube respond when there are unexpected events?

    - On misinformation, there is a Hit Pause Policy - train users on what misinformation is and how users can respond
    - YouTube has Trusted Flaggers - civil society organizations - to help them detect events quickly, plan in advance, and respond as quick as possible
- What is the process by which YouTube shares threat / intel analysis between platforms? How can USG align with those efforts?
    - Among the industry, there's a lot of sharing between Trust and Safety Teams and the more centralized Intel Team, Threat Analysis Groups, etc.
    - GIFCT is a good example of a consortium that helps to bring together industry partners to prevent violent extremism / live-streaming of violence
    - More than 99% of violating / violent content is removed before there are 10 views, but they are still trying to get better
    - YouTube is working on "age-gating" to prevent certain users (under 18 yo) from viewing violent content
        - But, HR defenders will push back because of this is documenting HR violations - need to make nuanced determinations and responses on violent content
- When do you decide if there's an opportunity to provide in-country training, etc.?
    - There is interest by YouTube in providing localized training through Trusted Flaggers, etc.
    - Constantly looking for new sources of information to inform their local Trust and Safety Teams
- How can USAID/State help to make sure the right people are connected on-the-ground?
    - Talk to YouTube's local counterparts who cover Burma and Bangladesh - sit in Singapore and cover Asia Pacific - welcome a follow-up conversation
    - Work through Ria / Josh (as "traffic cops") to set-up a deeper dive on both Burma and Bangladesh
- Would it be helpful to have these conversations more broadly, e.g. with other industry partners?
    - More regular touchpoints would be useful - build off initial GEC engagements as a starting point - good model because it's a regular / standing meeting
    - It would be useful to have conversations about broad trend lines / threats across all industry platforms, esp. open source intelligence
    - But, it would also be useful to have product / solution-specific conversations by platform (e.g. concerns related to specific channel / video features, go to YouTube directly)
    - Open lines of communication can be just as important or even more important than resources during times of crisis
- What has YouTube learned about how to detect problematic content over video?
    - YouTube is detecting some patterns in video content related to hate speech - uses not just text, also image detection
    - YouTube has a policy related to violent imagery - to detect violent videos that are also silent
    - Age-gating is also used – leaving violent content in a news report as an example but making it limited by age. We get criticized for leaving this content and for taking down this content. But it is a way to protect youth, without removing content.

-
- It's useful for YouTube to have concrete examples (e.g. channels / content) of what State / USAID finds most problematic, especially from the Field
    - For example, YouTube will get examples of VE content from NCTC and will respond to it - need to be careful this is not construed as USG-requested content removal
- Note: State will look at building out best practices / lessons learned when it comes to engaging technology companies on these issues. Looking to set up a process for communication, especially in a crisis situation.
    - YouTube agreed that more lines of communication are of course beneficial – suggested to think about how to do this and highlighted that would be good to have these lines of communication specific to YouTube as they are a unique platform