PUBLIC COMMENTS ON CASE 2023-018-FB-MR VIDEO OF COMMUNAL VIOLENCE IN THE INDIAN STATE OF ODISHA

Tech Global Institute (https://techglobalinstitute.com) is a policy lab with a mission to reduce equity and accountability gaps between technology platforms and the Global Majority. In this submission, we respond to the Oversight Board's request for public comments on the Video of Communal Violence in the Indian State of Odisha with specific reference to the following issues.

How social media platforms may be used to contribute to violence and discrimination against religious and ethnic groups in India and elsewhere.

There is credible, numerous, well-documented evidence specifically in India and South Asia that demonstrate how social media platforms have reportedly been used to amplify deep rooted inter-religious and inter-ethnic tensions with the intention of causing the spread of violence on the ground. Social media platforms typically view communal violence through the lens of Hate Speech and Violence and Incitement, however misinformation and disinformation are growing concerns to discriminate and spread hate against a particular religious or ethnic group. Misinformation is prevalent through duplicate accounts, or otherwise known as SUMA (same user, multiple accounts) that share out-of-context multimedia content or unverifiable text-based allegations claiming attacks on a particular religious or ethnic group to fuel retaliatory sentiment among other religious and ethnic groups. In some instances, content is misrepresentative of norms and intent of rituals of a particular religious or ethnic group that stokes discriminatory sentiments against them. Use of deepfake and manipulated media are becoming increasingly common to exacerbate inter-religious tensions, for example, through misrepresentation of religious text and/or leaders. And, as studies show, toxic and polarized content spreads six times faster than its benign counterpart.

There is a common counterargument that religious and ethnic violence are long-standing issues in South Asia and therefore, cannot be attributed to social media platforms alone. While there is truth to this view, it is necessary to critically examine how exactly social media platforms contribute to offline violence with reference to their **design**, **governing policies** and **decision-making framework**. There is a growing body of literature that indicates that the design of social media platforms, specifically recommender engines and engagement algorithms, play a tangible role in *amplifying* violence. Events that otherwise would have remained peripheral consequently gain momentum through algorithms, thereby producing contagion effects. This is salient in context of inter-religious and inter-ethnic events in South Asia—with its shared history and culture—where a content of communal violence in an Indian state is likely to similarly stoke violence in Bangladesh and/or Pakistan. In fact, Prime Minister Hasina called on Prime Minister Modi's government to quell communal violence in India, noting that such incidents result in spillover attacks on the Hindu minorities in Bangladesh.

The most recent spate of violence in the Indian state of Haryana demonstrates the role of social media in orchestrating and enabling violent clashes. In another example from the Indian state of Manipur, multimedia content deliberately mischaracterized by bad actors triggered reprehensible abuse of women, which in turn went viral on social media. In this case, violence against Kuki women by the Meitei community was found to be triggered by misinformation in the form of a viral clip that claimed to depict Meitei women being attacked by the Kuki community. Similarly, in October 2021, communal violence broke out throughout Bangladesh during Durga Puja festival after a viral video on Facebook showed the Quran placed on the knee of a Hindu deity in Cumilla. The amplification of the video through re-shares (reportedly 56,000 times before it was taken down) and re-uploads led to en masse vandalization of 101 Hindu temples and 186 homes in 12 districts, along with country-wide demonstrations and attacks that resulted in hundreds of injuries, the death of half a dozen of Hindus and Muslims,

and thousands of arrests. Earlier the same year, Hindu temples were vandalized and 11 Muslim protesters were killed as violence spread across the country in the wake of a visit by Indian Prime Minister Narendra Modi.

Second, Meta's Violence and Incitement (V&I) and Hate Speech policies fall short of addressing content that could contribute to violence and discrimination against religious and ethnic groups by raising three critical questions: **how are protected characteristics defined under these policies**, **how should Meta (and other social media platforms) treat violence against a majority versus minority group** and **how should Meta (and other social media platforms) treat attack on concepts, practices, beliefs and institutions**. We discuss these at length in subsequent sections of this submission.

Specifically, violence against minority communities in South Asia fueled by social media content that attack concepts is a recurrent theme. Bangladeshi Muslim rioters, for example, attacked Hindu houses in Pabna in November 2013, in Cumilla in April 2014, in Rangpur in November 2017, in Bhola in October 2019, in Narail in July 2022 in retaliation to content that allegedly demeaned the Prophet Mohammed on Facebook. In September 2012, Buddhist villages in Ramu were attacked and families were forced to flee, after images on a Facebook page run by a Buddhist showed a burnt Quran. In *all* cases, while the concept of Islam was allegedly attacked in the content itself, violence was instigated on Hindus (and Buddhists) *by* Muslims, thereby illustrating the inequities in threshold for violence as well as present a classic case of Heckler's veto. Platform policies currently do not account for power dynamics and societal contexts that play a role in exacerbating discrimination against religious and ethnic groups, thereby leaving up a significant number of harmful content.

Third and final, the decision-making framework in social media platforms and their underlying principles are skewed towards the U.S. that results in lower resource allocation and autonomy for South Asia and other emerging countries. There is no evidence indicating agency is decentralized, which typically contributes to delays in decisions about addressing potentially harmful content that can contribute to violence. This is especially evident in how the clear and present danger test (Schenck v. United States (1919), n.d.), which is grounded on First Amendment rights, is applied unilaterally to content decisions elsewhere in the world. Despite well established precedent on similar pieces of content contributing to violence and discrimination against a particular religious or ethnic group, as well as local laws putting reasonable and proportionate restrictions on free speech to protect the safety of its population, social media platforms tend to assess whether the threat posed by the content is explicit and can lead to imminent harm. This process is bureaucratic, relies on a central decision-making structure and does not factor in the speed at which content depicting or implying violence, segregation or retaliation against a particular religious or ethnic group spreads. As a result, even *if* social media platforms ultimately decide to remove a piece of content, the harm is already committed.

Insights into the socio-political context regarding the treatment of religious and ethnic groups in India, including the Indian government's policies and practices.

The outbreak of violence, particularly between Hindu and Muslim groups, during religious processions is the crucial context within which this case must be understood. Outside of this context, any analysis risks underestimating the risk potential that the at-issue content bears.

First, the content at-issue depicts potent and contextually significant triggers for violence. Stone pelting during religious processions are observed to be the lynchpin trigger to Hindu-Muslim violence. More than other forms of intimidation (such as brandishing of weapons, hate speech, physical nuisance etc.), stone pelting has been observed in an overwhelming number of cases where processions turned violent. Importantly, the mere apprehension of stone

pelting has been reported as the cause behind groups taking the first strike. Therefore, the at-issue content depicts acts that (in this context) have a heightened trigger value, as opposed to other forms of intimidation or violence.

Second, bad actors have increasingly relied on social media platforms to mobilize and incite more widespread violence. 'Live' and video messages of a charged nature have been observed to elicit a response to calls for violence, that are both numerous as well as delocalized. This has resulted in violence rapidly spreading beyond the localized environment of the procession, to other parts of the state and/or country. The critical role played by the 'virality' of charged clips in escalating violence should serve as important context in this case. In this backdrop, social media content risks being identified as an identifiable channel that enables violence to spread.

Indian governments and law enforcement agencies recognize the importance of limiting the spread of communally sensitive content especially when they anticipate communal violence to break out. For instance, ahead of a controversial court ruling on an issue that has previously led to riots, Indian local administration imposed prohibitory orders directing media (and social media) outlets to desist from airing content that could incite violence. While such orders can be overbroad and inconsistent with international free speech principles, it would not be surprising for social media platforms to consider provisionally suspending or geo-blocking communally sensitive content as it would be seen as consistent with the Indian government's approach to containing communal violence.

How Meta's Violence and Incitement policy should treat video content depicting scenes of communal violence, and how to assess whether such content may cause or contribute to offline violence.

Currently, there is no universally accepted content moderation framework guiding Meta's practices or the Oversight Board's subsequent analysis, which means they are free to adopt any legal content moderation standards and practices that suit its community and culture. This introduces significant gaps in global value alignment in the debate around voice versus safety, which illustrate two vital goals for platforms.

- **Enabling voice**: Meta describes "voice" as allowing "people to be able to talk openly about the issues that matter to them, […] even if some may disagree or find them objectionable." Social media platform's free speech goals imply that users who are victims and witnesses of violence are able to portray their lived experience. This could be newsworthy, helpful in combating misinformation, and documenting the reality from a firsthand account. Having said that, decisions about specific content need to be understood fairly in its context in order to assess whether it indeed serves these free speech goals. Thus, any assessment made on the axis of enabling voice should typically be context-intensive.
- **Quelling violence**: Meta describes "safety" as a commitment to "mak[e] Facebook a safe place" and states that content that "threatens people has the potential to intimidate, exclude or silence others and isn't allowed on Facebook." Regardless of whether it is possible to arrive at conclusions as to the 'aggressor' or bad actor in every situation, the first priority must be to quell ongoing violence and prevent the spread of further violence. Barring exceptional circumstances, actioning or restricting triggering or sensitive content, based on an assessment framework that prioritizes safety, can be less context-intensive, even tending to be context-agnostic.

From a policy design perspective, choosing to prioritize a context-agnostic criteria could minimize the role of social media content in contributing to religious or ethnic violence, even in circumstances with limited available resources. This is fundamentally different from how social media platforms operate today with its existing leaning towards First Amendment rights in its policy design and enforcement. Additionally, Meta's current approach to its

Community Standards, including its **Violence and Incitement Policy** (V&I) present three critical gaps in their assessment of voice versus safety:

- **How are protected characteristics defined under these policies**: Under the status quo, protected characteristics are defined as race, gender identity, ethnicity, religious affiliation, caste, etc. However, platforms treat *all* races, gender identities, ethnicities, etc. equally without acknowledging the historical and well-established power dynamics and inequities that exist within each group in a given context and geographic region. This risks in content decisions likely providing an advantage to a dominant group within a protected characteristic under the remit of free speech.
- **How should Meta (and other social media platforms) treat violence against a majority versus minority group**: While accounting for historical inequities and where the content originated from, content depicting violence and implied violence against a majority group *within a specific geographic region* should be the threshold for enforcing it. We argue the threshold should be lower for enforcement if the content is shared in reference to a minority group *within the same geographic region*. Under the status quo, in most cases, platforms apply the same thresholds for both groups, which could contribute to exacerbated violence against historically excluded religious and ethnic groups.
- **How should Meta (and other social media platforms) treat attack on concepts, practices, beliefs and institutions**: Under the status quo, platform policies, including V&I, require a target, i.e. a person, group of persons or a physical location, to be mentioned in the content itself, which frequently is not the case with how harmful materials manifest on the platform. There are limited safeguards for attacks on the concepts, practices and beliefs of a particular religious or ethnic group that could lead to violence against them or against those that are critical of them.

Communal violence in India and other South Asian countries is rarely predictable, and therefore, Meta is often going to be in a position where it has to decide on enforcement actions for the content in a *very* short span of time, and in a dynamically developing factual context. We propose that Meta accounts for the aforementioned questions in their assessments, even determining context-specific, pre-agreed criteria, to adjust with the time pressures and avoid making sub-optimal choices. Meta would be better advised to take the more safety-first decision, grounded in Article 3 of the Universal Declaration of Human Rights, than being in a position where it is *unable* to act in time. This would also put Meta's actions in a more operationally feasible and defensible position. All of these will support the legitimate aims of public order, respect for the rights of others, including the rights to life, security, etc. (as protected by ICCPR), as well as necessity and proportionality requirements.

What policy recommendations should be proposed to Meta that are relevant to this case.

At the outset, **we support Meta's decision to remove the content** outlined in this case given the risks of offline harm and inter-religious clashes in India. We also recommend that a similar approach should be adopted elsewhere around the world where the likelihood and imminence of offline harm is real for religious and ethnic minorities.

Further to this, we recommend Meta take into account **four fundamental policy considerations** in light of the societal, cultural and political realities in South Asia and elsewhere in the world in its assessment of violence and discrimination against, among others, religious and ethnic groups.

- Within a **protected characteristic**, Meta (and social media platforms) should fundamentally lean towards safeguarding the rights, interests and safety of the less dominant or historically marginalized group, as established by credible, external research authored by experts *from the region*. This will address historic

equity within protected groups as opposed to an unilateral, even harmful, response. For example, within a man-woman or wealthy-poor construct, priority should be given to the safeguarding of the rights of the woman or the poor, even if—in certain circumstances—it results in violation of some free speech rights of the dominant group.

- When assessing violence on a particular group within a specific geographic region, especially in context of religious affiliations, ethnicity, caste, and other sensitive protected characteristics, there should be differential thresholds for what constitutes "violence" and the appropriate actions based on the **power dynamic between majority versus minority groups**. Within a specific geographic region and context, *explicit violence* and *clear depiction of implied violence* on a majority group should be subject to enforcement, while a much lower threshold (such as demeaning rhetoric that may not necessarily depict violence) should be applied to minority groups and subsequent enforcement decisions. In this particular case, Meta's decision to remove the video content showing stone pelting among Hindus (a majority group in India) which essentially is a clear case of visible, explicit violence, is accurate. In a similar backdrop, if it were a procession composed of mainly Muslims (a minority group in India) who were subject to a demeaning remark, it may *also* be interpreted as harmful and pose risks of violence—and should be treated accordingly.

- Content assessments should treat **attacks on concepts, beliefs, practices and institutions** with the same weight as it does when the content explicitly states a target, given the multifaceted nature of inter-religious and inter-ethnic tensions in South Asia and elsewhere. Specifically, assessment should not be limited to a correlation between the given concept and the group adhering to it, rather take possible retaliation constructs into consideration and resolve Heckler's veto. In case of India for example, if a content specifically attacks Hindu practices, there may be retaliatory and violent actions against Muslims that the Hindu *majority* accuses of the attack.

- In qualification of harm on a particular group, **higher weight should be given to *impact* of the content as opposed to *intent***. This streamlines how platforms, including Meta, assesses content shared in praise or support versus condemning or neutral contexts, where the latter has often contributed to harm among religious and ethnic groups in South Asia and elsewhere. While this may seem counterintuitive to First Amendment rights, it adheres to Article 3 of Universal Declaration of Human Rights and does not necessarily stifle voice. This is especially critical in decisions on adding content to matching systems to reduce re-uploads of the same or similar content.

While the aforementioned principles can be instituted within the existing frame of V&I or even Hate Speech policy, we recommend that Meta should consider instituting a new public-facing policy framework specifically to address content that contains or depicts communal violence, or otherwise capture footage from triggering events on the ground given the significance, volume and real-world harm implications of such content. Content of this nature is well suited to be culled out from the broader V&I and Hate Speech policies, since it can be defined and identified with fewer gray areas. This policy framework should be sufficiently precise, clear, adaptable, non-discriminatory and inclusive, as well as transparent and consistent in its application. Importantly, it should address explicit and borderline content that could trigger violence and discrimination against a particular ethnic, linguistic or religious group. It is crucial to acknowledge that not all catalysts for violence and discrimination exhibit the same overt and recognizable characteristics as this case or are equally applicable in all regions. Other circumstances — while less conspicuous may also inflame communal tensions.