TQH Submission to Meta's Oversight Board on Cases involving Explicit AI Images

Overview and Executive Summary

Based in India, **The Quantum Hub** (TQH) works extensively on issues relating to technology and intermediary regulation as well as gender-responsive policymaking. Since this case lies at the intersection of our work, we make the following submission by drawing on insights from Indian social-cultural realities. Given our location and work, we have accorded primacy to Indian contextual analysis and the Indian case amongst the two identified in the problem statement.

Though our analysis of this case, we recommend that Meta:

1. Have a clear position on AI generated sexualised or derogatory imagery not being permitted on its platform, and prioritize this objective in content moderation practices;
2. Improving the design of reporting tools available on Instagram (in this case) to allow users to better navigate the tool and make more specific reports. Specifically, users must be provided ways to describe the reason for their report, and add additional context and information;
3. Consider slowing the spread of content that starts to be reported by users (particularly in this category) as an interim protective measure; and,
4. Reconsider and discontinue its policy of automatically closing appeals within 48 hours.

This document has been authored by Aparajita Bharti, Ujval Mohan and Devika Oberai

Response to the Board's Queries

Query 1	**The nature and gravity of harms posed by deep fake pornography including how those harms affect women, especially women who are public figures.**
Query 2	**Contextual information about the use and prevalence of deep fake pornography globally, including in the United States and India.**

(*addressed together below*)

Deepfake pornography or synthetically generated **sexualised or derogatory content (*for brevity, referred to collectively as SDC in this Submission*)** is the latest facet of a long standing online gender based violence (OGBV) pandemic that women around the world have been combating. Data supports that these digital creations are used most often to target women, often resulting in severe repercussions. Different studies independently arrive at the same conclusion — that an overwhelming majority of deep fake content (upwards of 90%) targets women by generating

SDC. In the misuse of synthetic content generators, there exists a troubling capacity to diminish and intimidate individuals, particularly women. Creation of SDC in concerning volumes therefore presents not just ethical dilemmas but also credible threats of **real world harm** and stifling women's expression, particularly in Indian settings.

Contextual factors: Dominant social norms in India place a premium on womens' modesty, reputation and 'honour'. Bad actors often target notions of 'decency' as a means to attack women, ranging from threats to sexual violence, disclosure of intimate details or imagery, aspercions of promiscuity, etc. As a result, individuals targeted on this front are more likely to incur social censure. SDCs form a new weapon to target these vulnerabilities and damage women's social capital while living in Indian realities. It is plausible that women targeted by SDCs experience ostracisation, shaming, and secondary harassment. For this reason, Indian laws (under the Indian Penal Code, Information Technology Act, and to an extent the Indecent Representation of Women Act) criminalize content that fits the SDC description. In fact, the gendered harms of deep fakes have also been expressly recognised by the Indian government, which is exploring regulatory solutions to prevent deep fake proliferation.

Perpetrators can exploit deep fake technology to threaten, blackmail, and manipulate victims, exacerbating the harm inflicted. This technology poses a significant threat as perpetrators can leverage deepfakes to instigate and perpetuate cycles of abuse, similar to other forms of non-consensual intimate image sharing. The social pressure, especially outside cosmopolitan or urban contexts, can often be so strong as to result in discrimination in professional or social settings, and severely damage personal relationships. This places SDC squarely on the OGBV continuum, representing not only an act of violence itself but also a catalyst for escalating threats against women.

SDC drives up online toxicity: In online contexts, SDCs can be a potent kernel that attracts and fuels sexist narratives and harmful online engagement. The tendency for toxicity to take over even benign content must be well understood and appreciated in crafting appropriate policy responses. Deepfake videos featuring Swift on Twitter(X) accumulated over 27 million views and over 260,000 likes within a span of 19 hours before the account responsible was suspended. Subsequently, X even blocked searches for 'Taylor Swift' on the platform. Bad actors can use their networks to widely disseminate SDC and use reactions and comments to bully the subject. When coupled with users' ability to identify the subject and their place of residence and/or work, the online harassment can quickly translate to threats of real world harm. Overall, SDCs contribute heavily towards a climate of apprehension for women both online and offline. SDC should therefore be considered high-risk content and a form of OGBV that must be proscribed.

Heightened vulnerabilities of public figures and politically active/opinionated women: These risks are exacerbated in the case of women who are public figures (such as politicians, activists, journalists, entertainment sector celebrities), who are already at the forefront of sexist attacks and reprisals when commenting on subjects in a charged social-political environment. There is a documented track record of Indian women public figures experiencing threats of aggravated violence (death, rape, etc.) and toxic abuses as a direct response to the expression of their political or social views online. Abusive online behavior against an Indian journalist has also resulted in police action. In the context of alarming levels of online toxicity directed at vulnerable groups, AI tools and

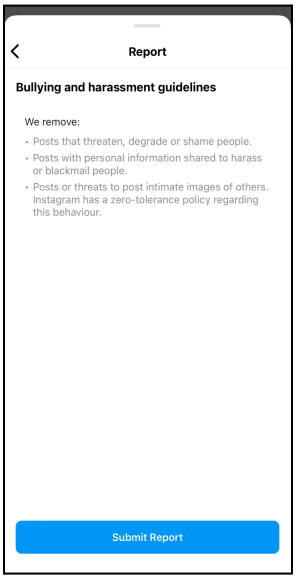
SDC fan the flame by providing inauthentic material that spurs further harassment. For instance, images of an Indian filmmaker were doctored by swapping his image with that of a female model in order to trigger homophobic and derogatory comments aimed at his identity.

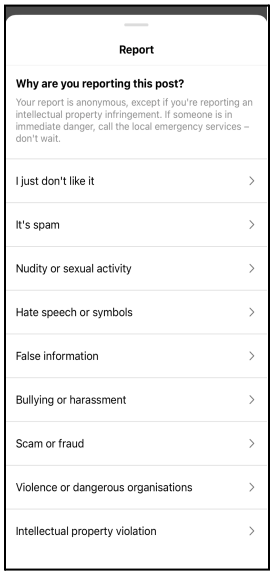
In this context, SDC serves as a formidable barrier for women with dissenting views or those from marginalized communities, preventing them from speaking up without fear of manipulation or retaliation — thus pushing women away from the political or public arena. **Without the strong assurance of countermeasures against SDC, women (especially young women) can be forced to consider leaving the public arena and stop airing their views to avoid reputational harm.**

A policy on SDCs envisioned by Meta should be crafted from the perspective of an ally to women and other victims of SDCs, rather than an impartial observer of the abusive potential of SDC. As a first step, this involves developing a clear and normative understanding of SDCs to be critically harmful, and prioritizing the prevention of SDC on Meta as the objective of any content moderation policy.

Traditional GBV is historically underreported on account of survivors having to bear the 'cost' of seeking justice or redressal that is often disruptive to their lives. Online platforms have the unique opportunity to change this paradigm as it applies to certain forms of OGBV, including SDC abuse. Changes in Meta's reporting and content moderation practices that result from this case should lead to increased ease of reporting SDC and enhanced efficiency in SDC removal.

Query 3	**Strategies for how Meta can address deepfake pornography on its platforms, including the policies and enforcement processes that may be most effective.**





Meta's content moderation efforts should aim to prevent the discovery of SDC on its platforms. At a time where most deep fakes fall within this category, achieving this objective requires a combination of effective user reporting, Meta's timely reviews, and automated moderation techniques.

We find that the way Instagram allows users to report SDC does not help users identify the most appropriate reason for their report, nor does it allow users to provide any details or explanations that can add much needed context to their report.

Though the problem statement clarifies that the content was found to violate the Bullying and Harassment Standard, the corresponding reporting label **does not indicate** that Meta removes

"derogatory sexualised photoshop or drawings" under this policy. Rather, the tool only refers to "*threats to post intimate images of others*". Understood normally, this does not describe the SDC content at-issue.

On the other hand, when a user begins to report SDC, at least three broad categories ("Nudity or sexual activity" "Bullying or harassment", and arguably "Hate speech or symbols") all appear to be viable heads to describe the content report, with **none of these containing descriptions that correctly describe the content**.

Furthermore, it is concerning that the 'Bullying' label enumerates content of differing risk potential (posts that 'shame' other people are relatively low risk by comparison to NCII, that is featured in the same list.) Even so, users have no ability to specifically identify the reason that most accurately describes the reason for their report. This is unlike other popular platforms (like X) that allow users to be more specific while reporting content.

Finally, users have no **ability to provide additional details, information or context to their report**. Allowing users to briefly describe their reasons for reporting content will provide Meta crucial context that could help action contextually high-risk content before it can create further harm. Today, there is also technical capability in automated systems to scan user-written descriptions and assess the category, nature, and urgency of a report, that can benefit both human moderators as well as automated moderation systems. By limiting users from providing such information, there is lost potential in building robust content reporting and assessment frameworks.

Since Meta relies significantly on user reports to action 'bullying' content, design inefficiencies in its reporting tool compromises the first layer of defense in preventing the damaging content of this nature from reaching a wider audience.

Recommendation: Empowering users to provide the greatest level of detail (as they are able or willing to) in their reports will help Meta speed up and appropriately prioritize its content moderation efforts. Given that the overwhelming majority of synthetic content is of SDC description, Meta should provide a specific way to report such high priority content in a manner that is clearly labeled, well understood by users, and with cross links to other plausible categories (even if the user inadvertently chooses another reporting head).

Query 4	**Meta's enforcement of its "derogatory sexualised photoshop or drawings" rule in the Bullying and Harassment policy, including the use of Media Matching Service Banks.**

Meta's inability to review and action the content in the first case of the problem statement reveals gaps in its enforcement of the Bullying and Harassment Standard. Aside from a flawed appeal mechanism and ineffective reporting options (*covered elsewhere in this Submission*), the omission to review the complaint/appeal could be attributed to the improper prioritization of reviewer resources and failure to implement temporary or interim restrictions on the content pending review.

SDC bears the risk of doing real world harm to users that is identical or closely resembles the harmful effects of NCII (non-consensually shared intimate imagery). **Therefore, in the broader gradation of priorities for content moderation, SDC should occupy a high-priority position, ideally, around the same level of attention that is paid to NCII reports. Instituting interim measures to prevent SDC's harm even until it reaches a content moderator can be a useful approach. Neglecting to slow content that received 'community notes' on X has been assessed as a limitation in its crowdsourced content moderation efforts.**

While one should always be cautious of recommending interim measures that can restrict users from expressing themselves to as large a group as they desire, we believe there are exceptions to this rule. SDC is an apt candidate for content that should ideally be slowed down pending review.

1. Content that is wrongly flagged under a label meant for SDC is most likely to be found to fall within **benign themes of artistic expression** or health or educational related content. Normally, there is no urgency associated with speech for such expression, and therefore, there is **no countervailing interest** that is impacted if its circulation is limited, and subsequently restored.
2. Further, SDC is likely to attract engagement that exacerbates violence and harassment intended towards women. So long as the content remains freely available on the platform, reposts (on IG stories), comments, user-to-user sharing are all avenues by which bad actors can maximize the reach of the material which all contribute to the ultimate harm borne by the user/individual in question. Therefore, delayed action or poor practices in response to reports under this label serves to **increase the likelihood of toxic discourse** on Meta platforms - which is better addressed if temporary restrictions are placed on the velocity of the spread of the post.

Therefore, interim or protective measures that automatically kick in upon receipt of a predetermined level of reports under this flag should be considered. **Without slowing suspicious content down, countermeasures such as fact checks (in case of misinformation) were not found to be effective. This coupled with our suggestion of improving the quality of user reporting by giving additional details can help Meta appropriately prioritize content for review.**

Finally, Meta has disclosed that a bulk of its enforcement of the Bullying and Harassment policy activity is proactive. As covered in response to the preceding query, we believe that Meta's proactive enforcement of the Bullying and Harassment policy must include automated tools. Meta's automated content moderation tools have shown some promise in correctly identifying intended content on the platform. More sophisticated versions of such tools can be effective in proactively identifying content that fits the SDC description. Incorporating AI in the identification process based on existing banks of SDC reported content could increase the accuracy of such exercises, with diminished risk of legitimate speech being censored.

We caution against over-reliance on Media Matching Service Banks (MMSB) as a way for Meta to proactively identify SDC. While MMSB can be effective in preventing secondary transmission of offending content (like in the second case of the problem statement), they are likely ineffective at preventing SDC from Meta unless the exact same image has been reported and found to be

offending. AI generated sexual imagery, by its nature, is likely to be novel and plenty given that it is easy and cheap to generate at scale.

Query 5	**The challenges of relying on automated systems that automatically close appeals in 48 hours if no review has taken place.**

We strongly advocate for Meta to reconsider and discontinue the policy of automatically closing appeals, especially when Meta itself fails to address the user complaint. From a user perspective, automatic closures do not contribute to resolution of complaints, but rather increase the burden on affected users to repeatedly track and report triggering content. The system does not bring with it any benefits that are immediately clear, when viewed from the perspective of healthy content moderation practices.

Having said that, we appreciate considerations of resource limitations that come with having to review a large number of reports that need appropriately trained human resources to address. Rather than closing the complaints/appeals, Meta should consider apportioning reviewer time to close out high priority user reports, and have automated tools / technical aids that can work to streamline reviewers' burden.

Even considered from a regulatory perspective, automated closure of complaints (when no review is conducted) is more likely to be treated as Meta's inaction. In the context of content which is likely illegal or harmful, regulators are less likely to be sympathetic to Meta's inaction, as compared to even slower review and content action.

Founded in **2017**, **The Quantum Hub** (TQH) is a multi-sectoral public policy research and consulting firm based out of **New Delhi, India**. Within our technology policy practice, we have been working on various digital economy and governance issues with a variety of stakeholders, and have closely tracked discussions around data protection and online safety. Within our gender practice, we work on women's labour force participation, women's representation in private and public sector leadership and women's overall health and wellbeing. We also work at the intersection of these two practices to track and study women's role and participation in an increasingly digital world.