OVERSIGHT BOARD PUBLIC COMMENT

CASES CONCERNING THE ASSASSINATION OF CANDIDATE RUNNING FOR MAYOR IN MEXICO

These cases raise the question of when, if at all, imagery depicting episodes designated "violating violent events" (VVEs) should be permitted on Meta's platforms. Here we offer analysis on how to think through this question–appealing both to Meta's own values and IHRL principles. We then apply the analysis to the four cases.

Summary of Core Insights

- Meta should amend its policy on third-party sharing of VVE imagery so that it is permitted when done for a clearly legitimate purpose, such as news reporting, awareness raising, or condemnation.
- Meta should continue to restrict third-party sharing of VVE imagery when it is done for clearly illegitimate purposes, such as supporting or glorifying the violence.
- Users should be instructed in the policy to indicate clearly their purpose in sharing VVE imagery when doing so, and ideally should be prompted to do so when they haven't indicated their purpose or when their indicated purpose is ambiguous.
- Meta should not rely on the newsworthiness allowance to decide when third-party sharing of VVE imagery is allowed. Relying on *ad hoc* and *ex post* newsworthiness allowances risks under-protecting and chilling legitimate sharing, and is incompatible with a commitment to legality.

Analysis

When, if ever, should imagery depicting *violating violent events* (VVEs) be permitted? One implausible answer to this question is that such imagery should *never* be permitted. Instructively, this is not the approach Meta has taken, in two respects. First, while imagery posted by perpetrators, their representatives, and third parties is restricted, the policy (as we read it) does not restrict *victims* of such events from posting imagery of their experiences (though curiously this is not stated explicitly). Second, as the Board notes in its call for comments, Meta sometimes applies its "newsworthiness allowance" to cases in which third parties share such imagery. This reflects the insight, with which we agree, that there can be *legitimate purposes* for sharing this content. These purposes include *raising awareness, condemnation, and responsible news reporting.* Sharing the actual imagery of VVEs can often serve these aims, helping audiences grasp the seriousness of what has occurred, more effectively than mere textual description. Meta's

own commitment to *voice*, and commitment to respecting free speech under IHRL, provide strong reasons to protect the sharing of violating violent events for these legitimate purposes.

Yet a policy of allowing all imagery depicting VVEs would also be misguided. Meta is reasonable to generally disallow perpetrators of these events, their representatives, or third parties to post this content in order to promote and glorify criminal violence–which can risk inspiring copycat attacks. Meta's own commitment to *safety* provides strong reason to disallow the sharing of violating violent events for these illegitimate purposes. (Meta should retain records of this footage, however, given its relevance for criminal prosecutions.)

Given the preceding analysis, Meta should alter its rules to allow explicitly the sharing of VVE imagery for legitimate purposes, while continuing to ban the sharing of VVE imagery for illegitimate purposes. The policy already contains an exception for sharing information or referencing VVEs for legitimate purposes of reporting, discussing, and condemning; this should be expanded explicitly to encompass sharing third-party imagery.

At present legitimate sharing of VVEs by third parties is not allowed, unless a newsworthiness allowance is applied. But relying on the newsworthiness allowance is insufficient, for two reasons. First, it is applied *ex post* only to some escalated content, and so much legitimate awareness-raising and condemnation will be restricted by default. Those who seek to share such content might rationally self-censor out of fear that they will receive a strike penalty on their account. Second, because it is applied in an *ad hoc* manner, it does not adequately account for the IHRL requirement of legality, which requires users to have reasonable grounds to understand and predict how the content they post will be treated.

How should Meta distinguish the legitimate sharing of this content from the illegitimate sharing of it? We think users can reasonably be expected to *indicate their purpose* when sharing content. For example, the kinds of captions involved in neutral news reporting or condemnation plainly indicate legitimate purposes, whereas the kinds of captions involved in glorification and support plainly indicate illegitimate purposes.

What about grey-area cases in which users' purpose is ambiguous--for example, when no purpose is indicated? Insofar as VVEs are uploaded to an MMS bank, it is conceivable that these users could receive a notification requiring them to indicate their purpose. What should Meta do if users' indicated purpose remains ambiguous? We think there can be reasonable disagreement about what Meta should do in these cases. One reasonable option is to remove content for which the purpose is ambiguous. Another

option is simply to allow such cases, erring on the side of protecting voice. Our research team is admittedly divided on which is the best option.

Finally, we note that even when VVE imagery is shared with a clearly legitimate purpose, it may be gratuitously gruesome or be requested to be removed by a family member (reflecting Meta's own commitments to privacy and dignity). Thus we think Meta's distinct policy on Violent and Graphic Content, which limits gratuitous imagery and empowers family members, should constrain any legitimate sharing of VVE imagery. The Dangerous Organizations and Individuals Policy should be amended to cross-reference the Violent and Graphic Content Policy and clarify their relation.

Cases

What are the implications of our analysis for the cases?

Case 1. Meta was wrong to remove the post in this case. Given the fact that the user was clearly sharing the content for news reporting purposes (evidenced by the caption and the identity of the user as a news organisation), it was shared for an entirely legitimate purpose that a well-designed policy would protect. While Meta's decision to remove the post was plausibly consistent with its current policy, the policy should be amended.

Case 2. Meta was right to leave this content up, albeit it arrived at that decision in the wrong way and for the wrong reason. The content was left up because a newsworthiness allowance was applied to it; so, the content was deemed violating, but Meta declined to remove it since it was in the public interest for people to see it. Yet given that the content was shared for an entirely legitimate purpose – news reporting and awareness raising – it should have been protected at the outset. As we note above, relying on *ad hoc* and *ex post* newsworthiness allowances risks under-protecting and chilling legitimate sharing, and is incompatible with a commitment to legality. The policy itself should be amended to protect explicitly the legitimate sharing of such content.

Case 3. In this case, the user's purpose in sharing the content was ambiguous. What should we interpret the speaker to be saying by directing people to Telegram to see an "uncensored" version of the video? On the one hand, we might charitably interpret the user to be raising awareness about the episode, perhaps motivated by the view that a full appreciation of an episode involves seeing it in all its grim detail (available only on sites with more permissive rules). On the other hand, we might interpret the user to be seeking to glorify violence by directing users to sites that eschew concerns for safety, dignity, or privacy. This case also raises distinct questions about Meta's policy on posts that direct users to more permissive sites, and about the importance of relying on news organisations' editorial decisions when deciding how much footage of VVEs to allow. In any case, the important point to stress is that the case is ambiguous. As we note above,

in cases of ambiguity, there is reasonable disagreement about what policy is best, given the tradeoff between the values of voice and safety. One option is to place a clear expectation on users to indicate their legitimate purpose in sharing the content, such that they cannot reasonably complain when it is removed in ambiguous cases. Ideally users would be prompted when uploading a video that matches the MMS bank.

Case 4. While Meta may have applied its current policy accurately in removing these posts, the policy is flawed for the same reasons as we offered in Case 1. Given that the post was shared for legitimate purposes of news reporting/awareness-raising, and is clearly indicated as such, a defensible policy would allow it.

* * *

Submission Prepared By:

Jeffrey Howard is Professor of Political Philosophy & Public Policy at University College London, where he directs the Digital Speech Lab, and Senior Research Associate at the Institute for Ethics in AI at Oxford University.

Ricki-Lee Gerbrandt is Research Fellow in Law and Platform Governance at University College London, based in the Digital Speech Lab.

Tena Thau is Research Fellow in Political Philosophy at University College London, based in the Digital Speech Lab.

About the Digital Speech Lab

The Digital Speech Lab hosts a range of research projects on the proper governance of online communications. Its purpose is to identify the fundamental principles that should guide the private and public regulation of online speech, and to trace those principles' concrete implications in the face of difficult dilemmas about how best to respect free speech while preventing harm. It is funded by a Future Leaders Fellowship awarded to Jeffrey Howard from UK Research and Innovation. Thanks to UKRI (grant reference MR/V025600/1) for enabling this work. Find out more at www.digitalspeechlab.com.