Public Comment for Oversight Board case on **POSTS SUPPORTING UK RIOTS**

General Lessons

The 2024 UK summer riots underscore the importance of enforcing several existing components of Meta's community standards. The Oversight Board should solicit information on how speedily and comprehensively Meta enforced these policies during the riots, in order to draw lessons for future incidents:

- The Violence and Incitement Policy straightforwardly prohibits speech advocating attacks, including coded threats where relevant threat and contextual signals are satisfied.

- The Hate Speech Policy prohibits all attacks on the basis of religion or race (e.g., such as posts attacking users on the basis of their Muslim or Arab identities), as well as the most severe attacks against "refugees, migrants, immigrants, and asylum seekers". This includes speech casting such persons as violent criminals (e.g., terrorists).

- The Misinformation Policy prohibits "misinformation or unverifiable rumors that expert partners have determined are likely to directly contribute to a risk of imminent violence or physical harm to people."

Case 1

This post was justifiably removed for the following reasons:

- Speech "calling for more mosques to be smashed and buildings to be set on fire where 'scum are living'" plainly violates the Violence and Incitement Policy.
- Referring to migrants as "terrorist[s]" and "scum" plausibly violates the Hate Speech Policy. While that policy has a carveout for "commentary on and criticism of immigration policies", this speech does not plausibly qualify.
- The reference to the Southport murder victims, and the suggestion that more victims will come, is a classic form of dangerous speech that justifies violence as a form of preemptive self-defence against a looming threat (in this case an imagined threat) (see Howard 2019, Leader-Maynard and Benesch 2016). This speech plausibly violates the Hate Speech Policy, which prohibits dehumanizing speech casting people as "criminals" on the basis of their protected characteristics or immigration status, as well as speech "supporting harm" to such persons.

That this post was not caught by Meta's automated system indicates a problem; it does not seem to be a difficult borderline case, so it is important to explore why it was a false negative and to ensure that Meta takes steps to address and remedy this problem.

Case 2

Opposing Meta's decision, we think this post should have been removed under the Violence and Incitement Policy, under its more complex provisions:

- The Violence and Incitement Policy restricts "[c]oded statements where the method of violence is not clearly articulated, but the threat is veiled or implicit, as shown by the combination of both a threat signal and contextual signal". Meta then provides a list of potential threat signals and contextual signals.

- One of the threat signals is "[a]cts as a threatening call to action (e.g., content inviting or encouraging others to carry out violent acts or to join in carrying out the violent acts)." One of the contextual signals is "[l]ocal context or expertise confirms that the statement in question could lead to imminent violence."

- In the context of the UK riots in which Muslim migrants were subjected to ongoing attacks, we believe that these signals were established—making removal justified. Meta claims that this speech should have been allowed "because the image did not constitute calls for violence against a target." Two contextual factors make this claim implausible: first, the ongoing violent riots; second, the "EnoughIsEnough" hashtag and text overlay listing time and place to gather, which strongly suggest that the post is not simply *describing* the violent actions depicted in the post, but *"inviting or encouraging*" such actions. (In other contexts, in contrast, these signals will often not be satisfied.)

Note the fact that the post is AI-generated is irrelevant to its harmfulness (see Fisher, Howard, and Kira 2024).

Case 3

Opposing Meta's decision, we think this post should have been removed either under the Hate Speech Policy or the Misinformation Policy, for the following reasons:

- The Hate Speech Policy prohibits dehumanizing speech casting members of a protected group as "[v]iolent criminals" ("including but not limited to: terrorists, murderers, members of hate or criminal organizations"). Meta claims that the post is protected because it doesn't refer to *all* Muslims, but only *some* Muslims. But such reasoning would effectively render visually depicted dehumanization impossible, since by necessity visuals cannot depict all

members of a group. It is also well established that dehumanizing claims taking a generic form (e.g., "Muslims are terrorists") do not necessarily make a claim about *all* members of the target group (Wodak et al. 2015). Yet such claims are paradigm cases of hate speech. So the fact that the post does not refer to all Muslims is not sufficient grounds for non-removal.

- The Misinformation Policy, as noted, prohibits "misinformation or unverifiable rumors that expert partners have determined are likely to directly contribute to a risk of imminent violence or physical harm to people." In context, where tensions are high over the murder of the young children in Southport and the false rumor that Muslims were responsible (which Meta acknowledged was a "false rumor" when justifying its moderation decision to the Oversight Board) this image strongly insinuates that very rumor. At the very least, it is *borderline* to such a violation, and ought to be demoted under Meta's policy of demoting borderline content (as part of its Types of Content We Demote policies).

References

Fisher, Howard, and Kira, "Moderating Synthetic Content: The Challenge of Generative AI," *Philosophy & Technology* 37, 133 (2024)

Howard, "Dangerous Speech," *Philosophy & Public Affairs* 47, 2 (2019)

Leader Maynard and Benesch, "Dangerous Speech and Dangerous Ideology," *Genocide Studies & Prevention* 9, 3 (2015-2016)

Wodak et al., "What a Loaded Generalization: Generics and Social Cognition," *Philosophy Compass* 10, 9 (2015)

Submission Prepared By:

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

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.

Maxime Lepoutre is Lecturer in Politics and International Relations at the University of Reading.

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.