Facebook: Moderating Content Showing Terrorist Attacks in Russia

July 2024

Centre for Law and Democracy
info@law-democracy.org
+1 902 431-3686
www.law-democracy.org

This Note[1] was prepared in response to a call by the Meta Oversight Board for comments on how it should apply its Dangerous Organizations and Individuals Community Standard in the context of three posts made immediately following the terrorist attacks in Moscow on 22 March 2024, which it removed as they depicted terrorist attacks on visible victims (the call for comments is available here).

Meta justified the takedowns to the Board on the basis that "quickly removing [a] moment of attack imagery on visible victims promotes safety by helping to mitigate the risk of contagion and copycat attacks while disrupting the spread of perpetrator propaganda" and that this also protects the "dignity of victims and their families". Two of the posts showed live video footage of the attack with armed individuals shooting at victims while one showed a still image of the same and then a video taken from outside of the shopping centre which was at the centre of the attack, showing it burning

Comments on the Policy Framework

There are a number of structural problems with the Dangerous Organizations and Individuals standard. International standards require restrictions on freedom of expression to be "provided by law", and this requires such laws to be sufficiently clear that those subject to them can regulate their behaviour accordingly. This part of the test is particularly relevant to Meta, and well within its power to implement across all of its content moderation standards. It is problematical that under the standard, in relation to Tier 1, removals cover "unclear or contextless references if the user's intent was not clearly indicated", which is clearly overbroad.

The category of "Glorification" is the most problematical in the Dangerous Organizations and Individuals standard. Even the title of this was tagged as problematical in the 2015 Joint Declaration on Freedom of Expression and Responses to Conflict Situations of the special international mandates on freedom of expression at the UN, OSCE, OAS and African Commission, which, in the context of terrorism, noted that "vague concepts such as glorifying', 'justifying' or 'encouraging' terrorism

should not be used".[2] But many of the terms and notions used are overbroad. For example, it covers characterising violent acts as an achievement but, often, they do represent an achievement for the perpetrators, something it is perfectly legitimate to point out.

International Human Rights Norms

International law has an important body of standards regarding when States may criminalise speech on grounds of terrorism, along with wider standards on balancing freedom of expression with the need to protect national security. Meta should seek to align with these international standards, given that Meta's objectives in moderating this type of speech align with those of (democratic) States and that these standards have been distilled from high-level sources, including peak national courts.

The essence of these standards is perhaps best captured in the following quotation from the 2016 Joint Declaration of the special international mandates:

> States should not restrict reporting on acts, threats or promotion of terrorism and other violent activities unless the reporting itself is intended to incite imminent violence, it is likely to incite such violence and there is a direct and immediate connection between the reporting and the likelihood or occurrence of such violence.[3]

It is clear even from the brief facts made available that none of the authors acted with intent to incite violence. More information and perhaps research would be needed to ascertain whether the second and third conditions were met, i.e. whether the posts were likely directly to incite imminent violence, although this seems highly unlikely. Underlying these conditions – the need for a direct and immediate connection between the reporting and the likelihood of imminent violence – is a key standard for preventing the abuse of national security restrictions on freedom of expression, namely by requiring a very close nexus between the statement and the risk of harm, including that the latter is imminent. Absent strict application of this standard, a very wide range of in fact harmless statements could be penalised on the basis that they might increase security risks.

There is, however, a great deal of difference between a State pursuing criminal charges against an individual, the primary means by which States address terrorism-inciting speech, and Meta taking down a post on Facebook. It seems reasonable for Meta to dispense with the intent requirement in this context, and potentially even to broaden out the scope of the prohibition to cover more promotional content as opposed to just incitement to violence, as the Dangerous Organizations and Individuals standard does. However, retaining the requirement of a close nexus between the statements concerned and the risk of these results is just as essentially in the Meta context as it is for States. Absent this requirement, as in the State context, speech which is perfectly innocuous or at least which has only a remote chance of promoting harm may be sanctioned.

[2] 4 May 2015, para. 3(b), https://www.osce.org/fom/154846.

[3] Joint Declaration on Freedom of Expression and Countering Violent Extremism, 4 May 2016, para. 2(d), https://www.osce.org/fom/237966.

The primary element of the Dangerous Organizations and Individuals standard relied on here – the ban on "third-party imagery depicting the moment of such attacks on visible victims" – fails to incorporate any requirement of a nexus to any specific harm. This is an inherent flaw in the standard. In some cases, such content will have a sufficient nexus to contagion, copycat offences and/or the promotion of terrorism, and it other cases it will not. The standard should be amended to reflect the need for a relevant nexus and the Oversight Board should assess whether, as a matter of fact, such a nexus existed in relation to these three posts. The call by the Board for comment on research into the harm caused by this sort of content suggests it is already thinking about this question.

Overrides

Another way to ensure an appropriate balance between respecting freedom of expression while safeguarding national security, particularly in the context of third-party reporting, is to recognise an override to the Dangerous Organizations and Individuals standard where there is a public interest in the content involved. Thus, the harshness of the general prohibition on third-party depictions of terrorist attacks could be mitigated by recognising an override whereby this rule would be waived where the post in question made a contribution to the public interest, normally via satisfying the public's right to be informed about events such as terrorist attacks. The need for information about such events was highlighted in the 2016 Joint Declaration, as follows:

> Everyone has the right to seek, receive and impart information and ideas of all kinds, especially on matters of public concern, including issues relating to violence and terrorism.[4]

The rationale for this is obvious. And the Board appears to recognise this element of the case by asking for comments on the relevance of a closed media environment in the country in question (see below).

In the area of privacy – which is at the heart of Meta's additional justification that taking down these posts "protects the dignity of victims and their families" – the need for public interest balancing when conflicts with freedom of expression arise is well established in human rights law. In the seminal *Von Hannover v. Germany (No. 2)*[5] case, the European Court of Human Rights noted that "cases such as the present one … require the right to respect for private life to be balanced against the right to freedom of expression" and then set out a list of factors to be considered, of which an "essential" one was the contribution of the statements "to a debate of general interest". This formulation was likely driven by the circumstances of that case, which involved media reports about a princess of Monaco, but bringing important facts to the notice of the public (i.e. news reporting) would equally qualify.[6]

There is nothing in the Dangerous Organizations and Individuals standard which recognises such an override. Meta does have a dedicated "newsworthiness allowance" which states: "In rare cases, we allow content that may violate Facebook Community Standards of Instagram Community Guidelines, if it's newsworthy and if keeping it visible is in the public interest." When applying this

[4] *Ibid.*, para. 1(a).

[5] 7 February 2012, Application Nos. 40660/08 and 60641/08, https://hudoc.echr.coe.int/#{%22itemid%22:[%22001-109029%22]}.

[6] *Ibid.*, paras. 106 and then 108-113.

override, Meta "weighs the public interest against the risk of harm" and it lists a number of factors it takes into account for this. Without commenting on those factors, this approach largely appears to meet the need for an override set out above.

However, in practice the newsworthiness allowance appears to provide only rather theoretical protection for newsworthy content. The page for this does not indicate how an assessment of newsworthiness is triggered but it does note that between 1 June 2021 and 1 June 2022 the allowance was applied only 68 times, which is a risibly low number given the annual number of posts on Meta's services. As such, this cannot be said to function as an effective override for the otherwise blanket ban on third-party imagery of terrorist attacks. It is not clear whether this allowance was specifically considered in these cases at either the stage of the initial removal or the internal review.

The content in question could have been important to help citizens understand what was going on, especially given the problematical nature of information flows in Russia, as well as for journalists and human rights researchers.

We recommend that something like the satirical override (found at the bottom of the standard) be added directly into the Dangerous Organizations and Individuals standard for content which provides important information to people in the context of an ongoing attack, where this overrides the risk of harm from that content.

The Situation in Russia

There is no question that Russia represents a closed media context and, indeed, it goes far beyond this and has a barrage of laws restricting online speech. The extreme nature of the local context can perhaps best be highlighted by the fact that Meta itself has been held by local courts in Russia to be an extremist organisation under the Law "On Combating Extremist Activities".[7]

Some of the more problematical Russian laws imposing broad content restrictions on online content include the "Yarovaya" Laws,[8] the "Disrespect for Authority" Laws[9] and the "Fake News" Law.[10] These laws provide for sanctions and other measures to be taken, among other things, for the dissemination of information which shows blatant disrespect for public morals, the State, official

[7] Federal Law of July 25, 2002 No. 114-FZ. A list of such organisations can be found at https://minjust.gov.ru/ru/documents/7822/, with item No. 96 covering Meta, albeit not WhatsApp, although this is only accessible to people in Russia (or via a VPN which locates you as being in Russia).

[8] Federal Law "On Amendments to the Federal Law "On Countering Terrorism" and certain legislative acts of the Russian Federation regarding the establishment of additional measures to counter terrorism and public safety", 6 July 2016, No. 374-FZ, and Federal Law "On Amendments to the Criminal Code of the Russian Federation and the Code of Criminal Procedure of the Russian Federation regarding the establishment of additional measures to counter terrorism and ensure public safety", 6 July 2016, No. 375-FZ.

[9] Federal Law of 18 March 2019 N 30-FZ "On Amendments to the Federal Law On Information, Information Technologies and Data Protection" and Federal Law of 18 March 2019 N 28-FZ "On Amendments to the Code of Administrative Offenses of the Russian Federation".

[10] Federal Law of 18 March 2019 N 27-FZ "On Amendments to the Code of the Russian Federation on Administrative Offenses" (the "Fake News" Law).

symbols or the constitution, and "inaccurate socially important information distributed under the guise of credible reports". All of these are illegitimate under international law.

As the level of practice, local sources have informed us that news reporting on the Moscow attacks was, especially early on, limited. While international news sources started reporting on the attacks immediately, there was no mention of it on Russian TV channels for about two hours and President Putin only gave a public speech about it the following day. There were attempts by Russian official sources to link the attacks to Ukraine. Even after charges were laid against two citizens of Tajikistan, official allegations still claimed that these individuals were heading to Ukraine as a safe third country following the attack. Belief that the 1999 bombings in the Russian cities of Buynaksk, Moscow and Volgodonsk were in fact organised by Russian State actors (contrary to the official version that they were orchestrated by Chechen separatists) remains reasonably widespread[11] and, extrapolating from this, some people believe that the March 2024 was as well. The point here is merely to emphasise the very shaky nature of reporting on terrorism and overall trust in that reporting in Russia.

Sanctions

It is well established under international human rights law that sanctions in expressive matters need to be proportionate.[12] As noted above, States have relatively limited options when responding to terrorist-related speech, whereas Meta has a much wider range of options, depending partly on the particular service in question. This is thus an area where Meta should take full advantage of the range of options available to it and the Oversight Board should take into account the extent to which Meta has made an effort to do that in its decisions.

The Dangerous Organizations and Individuals standard appears to envisage only one response, namely the removal of offending content. For the rationales that Meta has put forward in these cases, that largely makes sense (i.e. if the content really does pose a direct and imminent risk of further violence, adding a warning or demoting it is not sufficient). But the far more tailored approach of blurring the faces of victims would have resolved the privacy issue. In any case, consideration should be given to adding in the possibility of other measures in relation to this standard.

More concerning was the application of a rather intrusive strike in the third case. We do not have all of the facts but this appears to be an unnecessarily heavy measure. In addition, the risk of this being applied is not set out in the Dangerous Organizations and Individuals standard, which would represent a breach of the provided by law part of the test for restrictions on freedom of expression (unless it is mentioned somewhere else in Meta's rules). We also note the highly anomalous situation whereby the strike was maintained following the internal review but removed after the Board selected the case for review. This suggests some arbitrariness in Meta's actions in this area.

[11] See, for example, https://www.wilsoncenter.org/publication/foiled-attack-or-failed-exercise-look-ryazan-1999.

[12] See, for example, *Tolstoy Miloslavsky v. the United Kingdom*, 13 July 1995, Application No. 18139/91, para. 49 (European Court of Human Rights).