Pakistan Political Candidate Accused of Blasphemy

(2024-031-FB-MR)

Summary

The Board has upheld Meta's decision to remove a post containing an accusation of blasphemy against a political candidate. In the immediate run-up to Pakistan's 2024 elections, there was potential for imminent harm. However, the Board finds it is not clear the relevant rule under the Coordinating Harm and Promoting Crime policy, which prevents users from revealing the identity of a person in an "outing-risk group," extends to public figures accused of blasphemy in Pakistan or elsewhere. It is concerning this framing does not easily translate across cultures and languages, creating confusion for users trying to understand the rules. Meta should update its policy to make clear that users must not post accusations of blasphemy against identifiable individuals in locations where blasphemy is a crime and/or where there are significant safety risks to those accused.

About the Case

In January 2024, an Instagram user posted a six-second video of a candidate in Pakistan's February 2024 elections giving a speech. In the clip, the candidate praises former Prime Minister Nawaz Sharif, stating that "the person after God is Nawaz Sharif." The video had text overlay in which the user criticizes this praise for "crossing all limits of kufr," alleging he is a non-believer according to the teachings of Islam.

Three Instagram users reported the content the day after it was posted and a human reviewer found it did not violate Meta's Community Standards. The users who reported the content did not appeal that decision. Several other users reported the post over the following days but Meta maintained the content did not violate its rules, following both human review and automatic closing of some reports.

In February 2024, Meta's High Risk Early Review Operations (HERO) system identified the content for further review based on indications it was highly likely to go viral. The content

was escalated to Meta's policy experts who removed it for violating the Coordinating Harm and Promoting Crime policy rule based on "outing." Meta defines "outing" as "exposing the identity or locations affiliated with anyone who is alleged to be a member of an outing-risk group." According to Meta's internal guidance to reviewers, an outing-risk group includes people accused of blasphemy in Pakistan. When the video was flagged by HERO and removed, it had been viewed 48,000 times and shared more than 14,000 times. In March 2024, Meta referred the case to the Oversight Board.

Offenses relating to religion are against the law in Pakistan and the country's social media rules mandate the removal of "blasphemous" online content.

Key Findings

The Board finds that, given the risks associated with blasphemy accusations in Pakistan, removing the content was in line with the Coordinating Harm and Promoting Crime policy's rationale to prevent "offline harm."

It is not intuitive to users that risks facing members of certain religious or belief minorities relate to "outing," as commonly understood (in other words, risks resulting from a private status being publicly disclosed). The use of the term "outing" in this context is confusing, both in English and Urdu. Neither is it clear that people accused of blasphemy would consider themselves members of a "group" at risk of "outing," or that politicians would fall within an "outing-risk group" for speeches given in public, especially during elections. In short, the policy simply does not make it clear to users that the video would be violating.

Furthermore, the policy does not specify which contexts are covered by its line against outing and which groups are considered at risk. It also does not explicitly state that those accused of blasphemy are protected in locations where such accusations pose an imminent risk of harm. Meta explained that while it has an internal list of outing-risk groups, it does not publicly provide this list so that bad actors cannot get around the rules. The Board does not agree that this reason justifies the policy's overall lack of clarity. Clearly defining outing contexts and at-risk groups would inform potential targets of blasphemy allegations that such allegations are explicitly against Meta's rules and will be removed. This, in turn, could strengthen reporting by users accused of blasphemy in contexts where blasphemy poses legal and safety risks,

including Pakistan. Greater specificity in the public rule may also lead to more accurate enforcement by human reviewers.

The Board is also concerned that several reviewers found the content to be non-violating even though users repeatedly reported it and Meta's internal guidance, which is clearer, explicitly includes people accused of blasphemy in Pakistan in its outing-risk groups. It was only when Meta's HERO system identified the content, seemingly after it had gone viral, that it was escalated to internal policy experts and found to be violating. As such, Meta's at-scale reviewers should receive more tailored training, especially in contexts like Pakistan.

The Oversight Board's Decision

The Oversight Board upholds Meta's decision to remove the content.

The Board recommends that Meta:

- Update the Coordinating Harm and Promoting Crime policy to make clear that users must not post accusations of blasphemy against identifiable individuals in locations where blasphemy is a crime and/or there are significant safety risks to persons accused of blasphemy.
- Train at-scale reviewers covering locations where blasphemy accusations pose an imminent risk of harm to the person accused, providing them with more specific enforcement guidance to effectively identify, and consider nuance and context, in posts containing such allegations.

*Case summaries provide an overview of cases and do not have precedential value.

Full Case Decision

1. **Case Description and Background**

   At the end of January 2024, an Instagram user posted a six-second video clip on Instagram that shows a candidate in Pakistan's February 2024 elections giving a speech in Urdu. In the clip, the candidate praises former Prime Minister Nawaz Sharif stating

that “the person after God is Nawaz Sharif.” The video has text overlay in which the user criticizes this praise for “crossing all limits of ‘kufr,’” with “kufr” meaning not believing in Allah according to the teachings of Islam.

Three Instagram users reported the content the day after it was posted and a human reviewer found it did not violate Meta’s Community Standards. The reporting users did not appeal. A day later, two additional users reported the content. Reviewers decided the content was non-violating in all instances. Three days after it was posted, another user reported the content. Reviewers actioned the reports and found it did not violate Meta’s rules. The content was then reported by different users nine more times in the following days but Meta automatically closed these reports based on the prior decisions. All user reports were reviewed within the same day of reporting.

In early February, five days after it was posted, Meta’s High Risk Early Review Operations (HERO) system identified the content for further review based on signals indicating it was highly likely to go viral. The HERO system escalated the content to Meta’s policy experts. They removed it for violating the Coordinating Harm and Promoting Crime Community Standard for “outing: exposing the identity or locations affiliated with anyone who is alleged to be a member of an outing-risk group.” Based on Meta’s internal guidance to reviewers, an outing-risk group includes people accused of blasphemy in Pakistan. By the time the video clip was removed, it had been viewed approximately 48,000 times and shared more than 14,000 times. In late March 2024, Meta referred the case to the Oversight Board.

The Board noted the following context in reaching its decision.

The video clip was posted in the lead-up to Pakistan’s February 2024 elections, in which former Prime Minister Nawaz Sharif’s brother, Shehbaz Sharif, was elected Prime Minister for another term. The political candidate praising Nawaz Sharif in the video belongs to the same political party as the brothers.

Based on research commissioned by the Board, there were many posts online featuring the video echoing the blasphemy allegation. At the same time, there are other posts sharing the six-second video clip but which counter the accusation of “kufr,” claiming the edited clip takes the candidate’s full speech out of context. A different post that featured 60 seconds of the same speech, recorded by a different camera, was shared more than 1,000 times and viewed approximately 100,000 times.

This longer video provides fuller context to the election candidate's reference to Allah, which the text overlay to the video clip in this case had claimed is blasphemous.

Pakistan criminalizes offenses relating to religion under Sections 295-298 of the Pakistan Penal Code, including for defiling the Quran, derogatory remarks about the Prophet Muhammad, and the deliberate and malicious intention of outraging "religious feelings." Pakistan's social media rules (2021) mandate the removal of online content if it is "blasphemous," according to the penal code. This has led to people, often religious minorities and perceived critics of Islam being convicted of blasphemy for online posts and sentenced to death. According to the UN Special Rapporteur on freedom of religion or belief, religious minorities are often the target of blasphemy laws and broadly designated as "blasphemers" or "apostates," (A/HRC/55/47, para. 14). In Pakistan, Ahmadiyya Muslims and Christians are among those targets (A/HRC/40/58, para. 37). The UN Special Rapporteur has also noted that even those belonging to major religious denominations, including within Islam, who actively oppose maligning their religion through blasphemy laws also bear "an increased risk of being accused of 'betrayal' or 'blasphemy' and having retaliatory penalties inflicted upon themselves," (A/HRC/28/66, para. 7; see also public comment PC-29617). Blasphemy charges are also used to intimidate political opponents.

Blasphemy accusations have also led to mob lynchings in Pakistan, which have occurred in the country for decades, although not always implicating social media. Recent incidents include:

- In 2021, a mob attacked a sports equipment factory and beat up and burned a Sri Lankan man to death after he was accused of desecrating posters bearing the name of the Prophet Muhammad. Video footage shared on social media showed the crowd dragging the man's seriously injured body before he was burned to death as hundreds of onlookers cheered.
- In February 2023, a mob snatched a Muslim man from police custody and beat him to death after alleging he desecrated pages of the Quran. Video footage of the incident circulated on social media, showing people dragging the man by the legs while beating him with metal rods and sticks.
- In August 2023, mobs attacked Christian churches, burning them and damaging neighboring houses and household items, after two parishioners allegedly tore

pages of the Quran and wrote insulting remarks on them. Some parishioners had to flee their homes to escape.

- In February 2024, a woman was saved by the police from a potential attack by a mob after they mistook the print on her dress for Quranic verses. The mob had gathered around the restaurant where the woman was eating. She later issued a public apology.
- In May 2024, an elderly Christian man, Nazir Masih, was attacked by a mob for allegedly desecrating the Quran. The man sustained multiple head injuries and was taken to the hospital, where he died. In early June, a crowd of 2,500 people staged a rally to express support for the killing.
- In June 2024, while the Board was considering this case, a local tourist was killed and his body burned after he was accused of desecrating the Quran. The police had arrested the man but hundreds of people gathered around the police station where he was held, demanding he be handed over. The crowd then attacked the premises and dragged the man out. The crowd burned the body after beating the man to death. As of late June, police had arrested 23 people involved in the attack.

Politicians have also been the target of blasphemy-related violence. One of the most high-profile incidents involved former Governor of Punjab Salman Taseer, who was killed by his own bodyguard in 2011. Taseer had advocated for the repeal of Pakistan's blasphemy laws. The bodyguard was sentenced to death, with crowds taking to the streets to protest. After the bodyguard's execution, protestors erected a shrine around his grave. Another incident in 2011 involved unidentified perpetrators who killed the Federal Minister for Minorities Affairs Shahbaz Bhatti. Like Taseer, Bhatti had been critical of Pakistan's blasphemy laws.

In addition to UN special procedures, human rights and religious freedom organizations as well as other governments have all condemned the blasphemy-related mob violence in Pakistan resulting from accusations of blasphemy. Experts consulted by the Board confirmed that filing a police report against an accused person for blasphemy can result in their arrest to protect them from mobs. However, as the June 2024 incident has shown, police custody can be insufficient to protect accused blasphemers from mob violence.

Despite this, blasphemy prosecutions continue in Pakistan, and social media posts have formed the basis for conviction. For instance, a professor is facing the death penalty and has been imprisoned for more than 10 years for an allegedly blasphemous Facebook post in 2013. His lawyer was killed in 2014 for defending him. In 2020, police filed a blasphemy case against a human rights defender for a social media post. In March 2024, a 22-year-old student was convicted of blasphemy and sentenced to death for allegedly sending derogatory images about the Prophet Muhammad and his wives on WhatsApp. The Pakistani government has a history of monitoring online content for blasphemy and has ordered social media companies to restrict access to posts it considers blasphemous. The government has also met with Meta about posts it considers blasphemous.

Meta reported in its July 2023 – December 2023 transparency report that it restricted access in Pakistan to over 2,500 posts reported by the Pakistan Telecommunication Authority for allegedly violating local laws, including posts for blasphemy and "anti-religious sentiment." These reports only cover content Meta removes on the basis of a government request that do not otherwise violate Meta's content policies (i.e., posts flagged by the government that the company removes for violating Meta's rule on "outing" would not be included in this data). Based on the information provided by Meta on this case, there appears to be no indication that the government requested the review or removal of this content. As a member of the Global Network initiative, Meta has committed to respecting freedom of expression when faced with overbroad government restrictions on content.

2. User Submissions

The user did not provide a statement for this case.

3. Meta's Content Policies and Submissions

I. Meta's Content Policies

Coordinating Harm and Promoting Crime

The Coordinating Harm and Promoting Crime policy aims to "prevent and disrupt offline harm and copycat behavior" by prohibiting "facilitating, organizing, promoting, or admitting to certain criminal or harmful activities targeted at people, businesses, property or animals." Two policy lines in the Community Standards address "outing," the first is applied at-scale, and the second requires "additional context to enforce" (which means that this policy line is only enforced following escalation). The first policy line applies to this case. It specifically prohibits: "outing: exposing the identity or locations affiliated with anyone who is alleged to be a member of an outing-risk group." This policy line does not explain which groups are considered to be "outing-risk groups." The second policy line, which is only enforced on escalation, also prohibits "outing: exposing the identity of a person and putting them at risk of harm" for a specific list of vulnerable groups, including LGBTQIA+ members, unveiled women, activists and prisoners of war. Persons accused of blasphemy are not among the groups listed.

Based on Meta's internal guidance provided to reviewers, "outing-risk groups" under the first outing policy line include people accused of blasphemy in Pakistan, in addition to other specified locations. Moreover, "outing" must be involuntary; a person cannot out themselves (for example, by declaring themselves to be a member of an outing-risk group). To violate the policy under Meta's internal guidance, it is immaterial whether the blasphemy allegation is substantiated or whether the content misrepresents blasphemy. A mere allegation is sufficient to put the person accused within the "at-risk" group and for content to be removed.

Spirit of the Policy Exception

Meta may apply a "spirit of the policy" allowance to content when the policy rationale (the text introducing each Community Standard) and Meta's values demand a different outcome than a strict reading of the rules (set out in the "do not post" section and in the list of prohibited content). In previous decisions, the Board has recommended that Meta provide a public explanation of this policy allowance (Sri Lanka Pharmaceuticals, recommendation no. 1, Communal Violence in Indian State of Odisha). The relevant recommendations were accepted by Meta and are either fully implemented or in progress, according to the latest assessment by the Board.

II. Meta's Submissions

Meta explained that people accused of blasphemy in Pakistan were added to its internal list of "outing-risk groups" under the Coordinating Harm and Promoting Crime policy in late 2017, following violence related to blasphemy allegations in the country. As part of its election integrity efforts for Pakistan's 2024 elections, Meta prioritized the monitoring of content containing accusations of blasphemy given the high risk of offline harm, including extrajudicial violence, resulting from such allegations. Meta claims these integrity efforts resulted in the identification of the content in this case.

In its case referral, Meta noted the tension between voice and safety in this kind of content during an election period. Meta noted the public interest value in criticism of political candidates while acknowledging the various risks to safety posed by accusations of blasphemy in Pakistan, such as violence and death against politicians.

Meta found that the video clip's text overlay, which stated the electoral candidate had "crossed all limits of kufr," constituted a blasphemy allegation. For Meta, such language either suggests the political candidate has committed "shirk" – in other words, the belief in more than one God or holding up anything or anyone as equal to God – or it accuses the political candidate of violating Pakistan's blasphemy laws. In either case, Meta determined the risk of offline harm outweighed the potential expressive value of the video. The company clarified that had the video not included the text overlay, it would have remained on the platform.

Meta also provided an explanation of the HERO system it uses to detect high-risk content (in addition to user reports): the high risk of a particular content depends on the likelihood that it will go viral. Meta uses various signals to predict if content will go viral. These signals include the visibility of a piece of content on a user's screen, even partially, the post's language and the top country in which it is being shared when it is detected. HERO is not tied to a particular Community Standard. Moreover, Meta does not have a static or set definition for "high virality." Instead, high virality is informed by factors that vary across markets. As a result, the way that Meta weighs high-virality signals during high-risk events, which may include election periods, varies. Meta's internal teams may leverage view count during election periods to respond to specific risks. As such, HERO identifies content under any policy regardless of the likelihood of policy violation.

The Board asked Meta questions on the "outing-risk groups" rule in the Coordinating Harm and Promoting Crime policy and its enforcement, the company's election integrity efforts for Pakistan, and government requests for content takedowns under Pakistan's blasphemy laws and Meta's Community Standards. Meta responded to all the questions.

4. Public Comments

The Oversight Board received three public comments that met the the terms for submission. Two of the comments were submitted from Europe and one from Central and South Asia. To read public comments submitted with consent to publish, click here.

The submissions covered the following themes: Meta's content moderation of posts containing blasphemy allegations, the human rights impact of such allegations and related prosecutions in Pakistan, and the role that blasphemy accusations against public figures play in Pakistan and other countries.

5. Oversight Board Analysis

This case highlights the tension between Meta's values of protecting voice, including political criticism during elections, and of ensuring the safety of people accused of blasphemy, given threats to life and liberty that such accusations can carry in Pakistan.

The Board analyzed Meta's decision in this case against Meta's content policies, values and human rights responsibilities. The Board also assessed the implications of this case for Meta's broader approach to content governance.

5.1 Compliance with Meta's Content Policies

I. Content Rules

The Board finds that the rule in Meta's policy that prohibits revealing the identity of anyone alleged to be a member of an "outing-risk group" was not violated because it is unclear this extends to public figures accused of blasphemy in Pakistan or elsewhere.

In Pakistan, while members of certain religious or belief minorities may be considered "groups at risk" of harm, it is not intuitive that these risks relate to "outing" as commonly understood (i.e., risks resulting from a private status being publicly disclosed). Similarly, people accused of blasphemy do not necessarily consider themselves members of a "group" (compared to individuals who share a protected characteristic, which would include religious minorities). Moreover, it is not intuitive that politicians fall within an "outing-risk group" for information revealed in public speeches, especially in an election context. Indeed, other parts of Meta's rules in this area distinguish "political figures" and do not provide them protection for certain forms of "outing."

The Board finds that even if the internal guidance for reviewers contains more specific enforcement guidance listing the "outing groups" (or more accurately, contexts) covered by the policy, the public-facing policy does not contain the basic elements that would clearly prohibit the content in this case.

However, in exercising its adjudication and oversight function, the Board finds that reading the Coordinating Harm and Promoting Crime policy's prohibition in light of the policy rationale warrants the removal of the content, a conclusion that is reinforced by the human rights analysis below. According to the policy rationale, the Coordinating Harm and Promoting Crime Community Standard aims to "prevent and disrupt offline harm," including by forbidding people from "facilitating, organizing, promoting or admitting to certain criminal or harmful activities targeted at people." Meta allows for debating the legality or raising awareness of criminal or harmful activity as long as the post does not advocate for or coordinate harm. In this case, the Board finds that removing the content serves the policy rationale to prevent offline harm given the legal and safety risks that blasphemy accusations can carry in Pakistan. The user's post cannot be interpreted as raising awareness or discussing the legality of blasphemy in Pakistan. Rather, it does the opposite: it accuses someone of engaging in blasphemy in a location where they could face prosecution and/or safety risks. The accusation against the political candidate was in the immediate run-up to the February 2024

elections, when the candidate would have been actively engaged in campaigning. The potential for imminent harm, such as vigilante violence and criminal prosecution, was present. This amounted to "facilitating" a criminal or harmful activity prohibited by the Coordinating Harm and Promoting Crime policy.

A minority of the Board finds that the content should be removed on the basis of the spirit of the policy. For the minority, this policy exception should only be used on a very exceptional basis, especially to remove content. However, there are situations, such as those in this case, where it is necessary to address situations of heightened risk of harm that are not expressly prohibited in Meta's specific "do not post" rules. This is the case here, because the Coordinating Harm and Promoting Crime Community Standard does not expressly provide that blasphemy accusations in Pakistan are prohibited. However, removal conforms to the spirit of the policy overall, and its aims of reducing harm. The minority considers that when Meta removes content based on the "spirit of the policy," this should be documented, so that its use can be tracked, and then form a basis for identifying policy gaps that ought to be addressed.

5.2 Compliance with Meta's Human Rights Responsibilities

The Board finds that removing the content from the platform was consistent with Meta's human rights responsibilities, though Meta must address concerns about the clarity of its rules in this area and the speed of its enforcement.

Freedom of Expression (Article 19 ICCPR)

Article 19 of the ICCPR encompasses the freedom to "seek, receive and impart information and ideas of all kinds" and provides broad protection for expression, including "political discourse" and commentary on "public affairs." This includes ideas and views that may be controversial or deeply offensive, (General Comment 34, para. 11). The value of expression is particularly high when discussing matters of public concern, and freedom of expression is considered an "essential condition" for the effective exercise of one's right to vote during elections (General Comment 25, para. 12). All public figures, including those exercising the highest political authority such as

heads of state and government, are legitimately subject to criticism and political opposition (General Comment 34, para. 38).

Blasphemy laws are incompatible with Article 19 of the ICCPR (General Comment 34, para. 48). According to the UN High Commissioner for Human Rights, the right to freedom of religion or belief does not include the right to have a religion or a belief free from criticism or ridicule. On this basis, blasphemy laws should be repealed (see General Comment 34, para. 48 and A/HRC/31/18, para. 59-60; Rabat Plan of Action, report A/HRC/22/17/Add.4, at para. 19.) Indeed, blasphemy laws often cultivate religious intolerance and lead to persecution of religious minorities and dissent. Rather than criminalizing blasphemy and speech that reflects religious intolerance, in 2011, the international community rallied around UN Human Rights Council resolution 16/18, which set forth a useful toolkit of time-proven measures to combat religious intolerance and only resort to speech bans in cases of the risk of imminent violence.

When restrictions on expression are imposed by a state, they must meet the requirements of legality, legitimate aim, and necessity and proportionality (Article 19, para. 3, ICCPR; General Comment 34, para. 22 and 34). These requirements are often referred to as the "three-part test." The Board uses this framework to interpret Meta's human rights responsibilities in line with the UN Guiding Principles on Business and Human Rights, which Meta itself has committed to in its Corporate Human Rights Policy. The Board does this both in relation to the individual content decision under review and what this says about Meta's broader approach to content governance. As the UN Special Rapporteur on freedom of expression has stated, although "companies do not have the obligations of Governments, their impact is of a sort that requires them to assess the same kind of questions about protecting their users' right to freedom of expression," (A/74/486, para. 41).

I. *Legality (Clarity and Accessibility of the Rules)*

The principle of legality requires rules limiting expression to be accessible and clear, formulated with sufficient precision to enable an individual to regulate their conduct accordingly (General Comment No. 34, para. 25). Additionally, these rules "may not confer unfettered discretion for the restriction of freedom of expression on those

charged with [their] execution" and must "provide sufficient guidance to those charged with their execution to enable them to ascertain what sorts of expression are properly restricted and what sorts are not," (*Ibid.*). The UN Special Rapporteur on freedom of expression has stated that when applied to private actors' governance of online speech, rules should be clear and specific (A/HRC/38/35, para. 46). People using Meta's platforms should be able to access and understand the rules and content reviewers should have clear guidance regarding their enforcement.

The Board finds that the Coordinating Harm and Promoting Crime policy that prohibits identifying a member of an outing-risk group is not clear to users. First, the Board considers the use of the term "outing" to be confusing, both in English and in various languages the rule is translated into, including Urdu. Though "outing" generally refers to the non-consensual revelation of another person's private status and is commonly used in the context of the non-consensual disclosure of a person's sexual orientation or gender identity, this term is less commonly used in other contexts, such as religious affiliation or belief. Persons accused of blasphemy typically do not consider themselves at risk of "outing." In addition, the translation of the phrase "outing-risk group" into other languages is problematic. For example, the Urdu translation of the public version of the outing policy is especially unclear. An Urdu speaker in Pakistan would not understand that "شناخت ظاہر کرنا" means "outing risk." The translation also does not specify what "outing-risk" means. The Board is concerned that the current framing is not easily translated across various cultural contexts, creating potential confusion for users seeking to understand the rules. The lack of transparency is exacerbated by the fact that the Instagram Community Guidelines do not have a clear link to Meta's Coordinating Harm and Promoting Crime policy, which would make it harder for the user to know which rules apply to content accusing someone of blasphemy.

Second, the (public-facing) policy does not specify which contexts are covered by this policy line and which groups are considered at risk. This policy line does not expressly state that those accused of blasphemy, in locations where such accusations pose an imminent risk of harm, are protected. This is especially problematic for members of religious minorities who are most often the targets of blasphemy allegations, in particular where individuals may for safety reasons keep their religious affiliations or beliefs discreet and be vulnerable to "outing." It is important for these communities that the rules give them confidence that content directly endangering their safety is

prohibited. While internal guidance to reviewers is clearer, the repeated failure of reviewers in this case to correctly identify that the user's post infringed that guidance indicates that it is still insufficient.

Meta explained it does not publicly specify the list of outing-risk groups covered by the policy so that bad actors cannot get around the rules. The Board does not agree that this consideration justifies the policy's lack of clarity. Clearly defining the outing contexts and at-risk groups covered by this policy would inform potential targets of blasphemy allegations that such allegations are explicitly against Meta's rules and will be removed. In fact, Meta already does this in the other policy line against "outing" that requires additional context to enforce. That policy line, which was immaterial to this case, lists various at-risk outing groups (for example, LGBTQIA+ members, unveiled women, defectors and prisoners of war) that fall within the scope of that policy line. Applying the same approach in relation to blasphemy allegations would therefore not depart from Meta's current approach on clarity for its users on other comparable policy lines where the risks and tradeoffs appear to be similar. Providing this clarity could in turn strengthen reporting by users accused of blasphemy in contexts where this poses legal and safety risks, including in Pakistan. Greater specificity in the public-facing rules may also lead to more accurate enforcement by human reviewers.

The Board strongly urges Meta to specify in its public-facing rules that accusations of blasphemy and apostasy against individuals are prohibited in certain locations where they pose legal and safety risks. This would be not only much clearer in a standalone rule, separated from the concept of "outing" but also consistent with Meta's approach when listing at-risk groups in other parts of the Coordinating Harm and Promoting Crime policy. The Board does not contemplate every single detail to be laid out in the public-facing language of the Coordinating Harm and Promoting Crime Community Standard. But as a bare minimum, the relevant elements of the prohibited content, such as the types of groups that are protected, the types of locations the rule applies to, and the types of expression that fall under the prohibition, would give more clarity to users. This would address Meta's concern about bad actors trying to evade the rules while making potential targets of blasphemy accusations aware that this type of content is prohibited.

II. Legitimate Aim

Any restriction on freedom of expression should also pursue one or more of the legitimate aims listed in the ICCPR, which includes protecting the rights of others (Article 19, para. 3, ICCPR). This includes the rights to life, liberty and security of persons (Articles 6 and 9, ICCPR). The Board also recognizes that protecting people from offense is not considered a legitimate aim under international human rights standards. The Board has previously recognized that Meta's Coordinating Harm and Promoting Crime policy pursues the legitimate aim of protecting the rights of others in the context of elections, such as the right to vote (Australian Electoral Commission Voting Rules). The Board finds that the policy's aim to "prevent and disrupt offline harm" is consistent with the legitimate aim of protecting the rights to life, liberty and security of persons.

III. Necessity and Proportionality

Under ICCPR Article 19(3), necessity and proportionality requires that restrictions on expression "must be appropriate to achieve their protective function; they must be the least intrusive instrument amongst those which might achieve their protective function; they must be proportionate to the interest to be protected," (General Comment No. 34, para. 34).

The Board finds that removing the content in this case is consistent with the principle of necessity and proportionality, and finds the six factors for assessing incitement to violence and discrimination in the Rabat Plan of Action instructive to assess the likelihood of harm resulting from this post. Those factors are the content and form of the expression, the speaker's intent, the identity of the speaker, their reach, and the likelihood and imminence of harm.

In relation to the content and form of the expression and the speaker's intent, as outlined above, the content in the post clearly communicates a desire to accuse the political candidate of blasphemy, doing so without indication of intent to raise awareness about or debate the legality of such speech.

In relation to the identity of the speaker and the reach of their content, the Board notes that the user is not a public figure with influence over others, and has relatively few followers. Nevertheless, the account's privacy settings were set to public at the time the

content was posted, and the content was shared approximately 14,000 times by about 9,000 users. This shows that speech from non-public figures can still be disseminated very broadly on social media, and virality is challenging to predict. The reach of this post increased its potential for harm, notwithstanding that the person who posted it does not seem to be in a position of authority.

The Board considers that there is a likelihood of imminent harm given the context around blasphemy allegations in Pakistan, where such accusations pose a serious risk of physical harm and even death. That context of national legal prohibitions, related prosecutions and violence from would-be vigilantes, is set out in section 1 above (see also public comments PC-29615 and PC-29617).

Given the examination of the six Rabat factors, the Board finds it was necessary and proportionate to remove the post in question.

Moreover, the Board is particularly concerned that numerous reviewers enforcing Meta's Community Standards all found the content to be non-violating. This was despite repeated reports from users and Meta's claimed prioritization of enforcement against content of this kind as part of its election integrity efforts in Pakistan, given the high risk of offline harm it could pose. It was only when Meta's HERO system identified the content days later, seemingly after it had gone viral, that it was escalated to internal policy experts and found to violate the Coordinating Harm and Promoting Crime Community Standard. Various human reviewers missed earlier opportunities to accurately enforce against the content, indicating a need for tailored training for reviewers to understand how to spot violations in contexts like Pakistan. This is especially important in election contexts, where tensions may escalate and accurate enforcement is essential to guard against unnecessary restrictions on speech and prevent offline harm. In evaluating whether its election integrity efforts in Pakistan were successful, Meta should consider why so many reviewers failed to accurately enforce against this post, and how to ensure more effective election integrity efforts in countries with similar risks in future.

While blasphemy accusations can create significant risks for politicians in countries where blasphemy is criminalized, there can nevertheless be important discussions about blasphemy, in particular in the context of an election. Meta must be cautious to

avoid over-enforcement of the policy against content that *is not* accusing individuals of blasphemy, but rather engaging in political discussions. Such over-removal would be particularly concerning in contexts where political speech is already subject to excessive government restrictions that do not comply with international human rights law. Not all content that uses the term "kufr" will be a blasphemy accusation, as demonstrated by the variety of similar videos shared of the same events as in this case. Therefore, the Board reminds Meta that its human rights responsibilities require it to respect political expression when the content is shared to counter allegations of blasphemy or engage in discussions about blasphemy without placing individuals at risk. It is important that training to moderators emphasizes the importance of freedom of expression in this context, and allows them to escalate decisions to more specialized teams when more contextual analysis may be needed to reach a correct decision.

6. The Oversight Board's Decision

The Oversight Board upholds Meta's decision to take down the content.

7. Recommendations

Content Policy

To ensure safety for targets of blasphemy accusations, Meta should update the Coordinating Harm and Promoting Crime policy to make clear that users must not post accusations of blasphemy against identifiable individuals in locations where blasphemy is a crime and/or there are significant safety risks to persons accused of blasphemy.

The Board will consider this recommendation implemented when Meta updates its public-facing Coordinating Harm and Promoting Crime Community Standard to reflect the change.

Enforcement

To ensure adequate enforcement of the Coordinating Harm and Promoting Crime policy line against blasphemy accusations in locations where such accusations pose an imminent risk of harm to the person accused, Meta should train at-scale reviewers

covering such locations and provide them with more specific enforcement guidance to effectively identify and consider nuance and context in posts containing blasphemy allegations.

The Board will consider this recommendation implemented when Meta provides updated internal documents demonstrating that the training of at-scale reviewers to better detect this type of content occurred.

***Procedural Note:**

- The Oversight Board's decisions are made by panels of five Members and approved by a majority vote of the full Board. Board decisions do not necessarily represent the views of all Members.

- Under its Charter, the Oversight Board may review appeals from users whose content Meta removed, appeals from users who reported content that Meta left up, and decisions that Meta refers to it (Charter Article 2, Section 1). The Board has binding authority to uphold or overturn Meta's content decisions (Charter Article 3, Section 5; Charter Article 4). The Board may issue non-binding recommendations that Meta is required to respond to (Charter Article 3, Section 4; Article 4). Where Meta commits to act on recommendations, the Board monitors their implementation.

- For this case decision, independent research was commissioned on behalf of the Board. The Board was assisted by Duco Advisors, an advisory firm focusing on the intersection of geopolitics, trust and safety, and technology. Memetica, a digital investigations group providing risk advisory and threat intelligence services to mitigate online harms, also provided research. Linguistic expertise was provided by Lionbridge Technologies, LLC, whose specialists are fluent in more than 350 languages and work from 5,000 cities across the world.