Referring to Designated Dangerous Individuals as "Shaheed"

Policy Advisory Opinion

PAO-2023-01

Executive Summary

The Board finds that Meta's approach to moderating content that uses the term "shaheed" to refer to individuals designated as dangerous substantially and disproportionately restricts free expression. Meta interprets all uses of "shaheed" referring to individuals it has designated as "dangerous" as violating and removes the content. According to Meta, it is likely that "shaheed" accounts for more content removals under the Community Standards than any other single word or phrase on its platforms. Acts of terrorist violence have severe consequences – destroying the lives of innocent people, impeding human rights and undermining the fabric of our societies. However, any limitation on freedom of expression to prevent such violence must be necessary and proportionate, given that undue removal of content may be ineffective and even counterproductive.

The Board's recommendations start from the perspective that it is imperative Meta take effective action to ensure its platforms are not used to incite acts of violence, or to recruit people to engage in them. The word "shaheed" is sometimes used by extremists to praise or glorify people who have died while committing violent terrorist acts. However, Meta's response to this threat must also be guided by respect for all human rights, including freedom of expression.

On October 7, 2023, as the Board was finalizing this policy advisory opinion, Hamas (a designated Tier 1 organization under Meta's Dangerous Organizations and Individuals policy) led unprecedented terrorist attacks on Israel that killed an estimated 1,200 people and resulted in roughly 240 people being taken hostage (Ministry of Foreign Affairs, Government of Israel). According to news reports, as of February 6, 2024, at least 30 of the estimated 136 hostages remaining in Hamas captivity in early January are believed to have died. Meta immediately designated these events a terrorist attack under its Dangerous Organizations and Individuals policy. Israel quickly initiated a military campaign in response to the attacks. That military campaign had killed more than 30,000 people in Gaza as of March 4 (UN Office for the Coordination of Humanitarian Affairs, drawing on data from the Ministry of Health in Gaza). Reports from January indicated that 70% of fatalities were estimated to be women and children.

Following these events, the Board paused publication of this policy advisory opinion to ensure its recommendations were responsive to the use of Meta's platforms and the word "shaheed" in this context. This additional research confirmed the Board's recommendations to Meta on moderating the word "shaheed" held up, even under the extreme stress of such events, and would ensure greater respect for all human rights in Meta's response to crises. At the same time, the Board underscores that Meta's policies in this area are global and their impact extends far beyond this conflict. While acknowledging the salience of recent events in Israel and Palestine, the Board's recommendations are also global and not limited to any particular context.

In the Board's view, Meta's approach to moderating the word "shaheed" is overbroad, and disproportionately restricts freedom of expression and civic discourse. For example, posts reporting on violence and designated entities may be wrongly removed. Meta's approach also fails to consider the various meanings of "shaheed," many of which are not intended to glorify or convey approval, and lead all too often to Arabic speakers and speakers (many of them Muslim) of other languages having posts removed, without that removal serving the purposes of the Dangerous Organizations and Individuals policy. Moreover, Meta's policies prohibit, for example, the glorification, support and representation of designated individuals, organizations and events, as well as incitement to violence. These policies, enforced accurately, mitigate the dangers resulting from terrorist use of Meta's platforms. Accordingly, the Board recommends that Meta end its blanket ban on use of the term "shaheed" to refer to individuals designated as dangerous, and modify its policy for a more contextually informed analysis of content including the word.

Background

In February 2023, Meta asked the Board whether it should continue to remove content using the Arabic term "shaheed," or شهيد written in Arabic letters, to refer to individuals designated under its Dangerous Organizations and Individuals policy. "Shaheed" is also a loanword (meaning many non-Arabic languages have "borrowed" this Arabic-origin term, including by adapting its spelling).

The company describes the word "shaheed" as an "honorific" term, used by many communities across cultures, religions and languages, to refer to a person who has died unexpectedly, such as in an accident, or honorably, such as in a war. The company acknowledges the term has "multiple meanings" and while there is "no direct equivalent in the English language," the common English translation is "martyr." Noting that "in English the word 'martyr' means a person who has suffered or died for a justified cause and typically has positive connotations," Meta states that "it is because of

this use that we have categorized the term [“shaheed”] as constituting praise under our [Dangerous Organizations and Individuals] policy.”

Meta’s presumption that referring to a designated individual as “shaheed” always constituted “praise” under the Dangerous Organizations and Individuals policy resulted in a blanket ban. Meta has acknowledged that because of the term’s multiple meanings it “may be over-enforcing on significant amounts of speech not intended to praise a designated individual, particularly among Arabic speakers.” In addition, Meta does not apply the Dangerous Organizations and Individuals policy exceptions allowing the use of “shaheed” to “report on, condemn or neutrally discuss designated entities.” This has continued under the latest updates to the policy – made in December 2023 – that now prohibit “glorification” and “unclear references” instead of “praise,” which has been removed completely.

Meta started a policy development process in 2020 to reassess its approach to the term “shaheed” because of these concerns. However, no consensus was reached within the company and no new approach agreed.

When requesting this policy advisory opinion, Meta presented three possible policy options to the Board:

1) Maintain the status quo.
2) Allow use of “shaheed” in reference to designated individuals in posts that satisfy the exceptions to the “praise” prohibition (for example, reporting on, neutrally discussing or condemning), so long as there is no other praise or “signals of violence.” Some examples of these signals proposed by Meta included a visual depiction of weapons, or references to military language or real-world violence.
3) Allow use of “shaheed” in reference to designated individuals so long as there is no other praise or signals of violence. This is regardless of whether the content falls under one of the exceptions listed above, in contrast to the second option.

The Board did consider other possible policy choices. For the reasons given throughout this policy advisory opinion, the Board’s recommendations align closely with the third option, though fewer signals of violence are adopted than Meta proposed in its request, and there is a requirement for the broader application of policy exceptions for reporting on, neutrally discussing and condemning designated entities and their actions.

Key Findings and Recommendations

The Board finds that Meta's current approach to the term "shaheed" in connection to individuals designated as dangerous is overbroad, and substantially and disproportionately restricts free expression.

"Shaheed" is a culturally and religiously significant term. At times it is used to indicate praise of those who die committing violent acts and may even "glorify" them. But it is often used, even with reference to dangerous individuals, in reporting and neutral commentary, academic discussion, human rights debates and even more passive ways. Among other meanings, "shaheed" is widely used to refer to individuals who die while serving their country, serving their cause or as an unexpected victim of sociopolitical violence or natural tragedy. In some Muslim communities, it is even used as a first (given) name and surname. There is strong reason to believe the multiple meanings of "shaheed" result in the removal of a substantial amount of material not intended as praise of terrorists or their violent actions.

Meta's approach of removing content solely for using "shaheed" when referring to designated individuals intentionally disregards the word's linguistic complexity and its many uses, treating it always and only as the equivalent of the English word "martyr." Doing so substantially affects freedom of expression and media freedoms, unduly restricts civic discourse and has serious negative implications for equality and non-discrimination. This over-enforcement disproportionately affects Arabic speakers and speakers of other languages that have "shaheed" loanwords. At the same time, other ways of implementing the Dangerous Organizations and Individuals policy would still enable Meta to advance its value of safety and its goal of keeping material glorifying terrorists off its platforms. The current policy is therefore disproportionate and unnecessary.

To align its policies and enforcement practices around the term "shaheed" more closely with human rights standards, the Board recommends the following recommendations (see section 6 for recommendations in full):

1. **Meta should stop presuming that the word "shaheed," when used to refer to a designated individual or unnamed members of designated organizations, is always violating and ineligible for policy exceptions.** Content referring to a designated individual as "shaheed" should be removed as an "unclear reference" in only two situations. First, when one or more of three signals of violence are present: a visual depiction of an

armament/weapon, a statement of intent or advocacy to use or carry an armament/weapon, or a reference to a designated event. Second, when the content otherwise violates Meta's policies (e.g., for glorification or because the reference to a designated individual remains unclear for reasons other than use of "shaheed"). In either scenario, content should still be eligible for the "reporting on, neutrally discussing and condemning" exceptions.

2. **To clarify the prohibition on "unclear references," Meta should include several examples of violating content, including a post referring to a designated individual as "shaheed" combined with one or more of the three signals of violence specified in recommendation no. 1.**

3. **Meta's internal policy guidance should also be updated to make clear that referring to designated individuals as "shaheed" is not violating except when accompanied by signals of violence, and that even when those signals are present, the content may still benefit from the "reporting on, neutrally discussing or condemning" exceptions.**

If Meta accepts and implements these recommendations, under its existing rules, the company would continue to remove content that "glorifies" designated individuals, characterizes their violence or hate as an achievement, or legitimizes or defends their violent or hateful acts, as well as any support or representation of a designated dangerous entity. The Board's proposed approach that results from these recommendations would be for Meta to stop *always* interpreting "shaheed" in reference to a designated individual as violating, only removing the content when combined with additional policy violations (e.g., glorification) or as an "unclear reference" due to signals of violence. Such content would still need to be eligible for the "reporting on, neutrally discussing and condemning designated individuals" policy exceptions.

The Board also recommends that Meta:

4. **Explain in more detail the procedure by which entities and events are designated under its Dangerous Organizations and Individuals policy to improve transparency around this list.** Meta should also publish aggregated information on the total number of entities within each tier of its designation list, as well as how many entities were added and removed in the past year.

5. **Introduce a clear and effective process for regularly auditing designations and removing those no longer satisfying published criteria to ensure its Dangerous Organizations and**

Individuals entity list is up-to-date, and does not include organizations, individuals and events that no longer meet Meta's designation definition.

6. **Explain the methods it uses to assess the accuracy of human review and the performance of automated systems in the enforcement of its Dangerous Organizations and Individuals policy.** Meta should also periodically share the outcomes of performance assessments of classifiers used in enforcement of this policy, providing results in a way that can be compared across languages and/or regions.

7. **Clearly explain how classifiers are used to generate predictions of policy violations and how Meta sets thresholds for either taking no action, lining content up for human review or removing content.** This information should be provided in the company's Transparency Center to inform stakeholders.

Policy Advisory Opinion in Full

1. Meta's Request

I. The Request for a Policy Advisory Opinion

1. In its request (available in English and Arabic), Meta asked the Board whether it should continue to remove content using "shaheed" (شهيد) (and its singular/plural forms, including loanword variants in Arabic and other languages) to refer to individuals designated as dangerous under its Dangerous Organizations and Individuals policy, or whether a different approach would better align with the company's values and human rights responsibilities. Meta also requested guidance on similar content issues that may arise in the future.

II. Meta's Approach

2. Meta's Dangerous Organizations and Individuals policy previously prohibited "praise, substantive support or representation of designated entities and individuals" and of designated "violating violent events." On December 29, 2023, Meta updated this policy, removing the prohibition on praise and adding prohibitions on "glorification" and "unclear references." The Board extended its considerations to assess any impact these policy changes would have on its findings and recommendations.

3. This request for a policy advisory opinion concerned the prohibition on praise of designated individuals, and not the parallel prohibitions on substantive support or representation. Compared to representation and substantive support, Meta considered "praise" to be the least severe violation. Meta reserves "Tier 1" for the most dangerous entities, including terrorist, hate and criminal organisations. When Meta designates a "violating violent event," the perpetrators of those events are also designated as Tier 1 dangerous individuals, and glorification (previously "praise") of them is therefore prohibited. Designated events include "terrorist attacks, hate events, multiple-victim violence or attempted multiple-victim violence, serial murders or hate crimes." While Meta controls its own list of designated entities, which is not public, designations are in part based on the U.S. government's designation lists (i.e., Meta's designations list will be at least as extensive as the U.S. government lists it derives from). The Community Standards explain that Tier 1 entities include "specially designated narcotics trafficking kingpins (SDNTKs)," "foreign terrorist organizations (FTOs)" and "specially designated global terrorists." These U.S. government lists are publicly available here, here and here respectively.

4. Prior to the December 29, 2023, policy update, Meta's definition of "praise" included speaking "positively about a designated entity or event," giving a designated entity "a sense of achievement," legitimizing "the cause of a designated entity by making claims that their hateful, violent or criminal conduct is legally, morally or otherwise justified or acceptable" and aligning "oneself ideologically with a designated entity or event." This definition, or series of examples, was added to the policy to improve clarity following an Oversight Board recommendation in one of our first cases (Nazi Quote decision, recommendation no. 2). Following the December 2023 policy update, "glorification" is defined as "legitimizing or defending the violent or hateful acts of a designated entity by claiming that those acts have a moral, political, logical or other justification that makes them acceptable or reasonable," or "characterizing or celebrating the violence or hate of a designated entity as an achievement or accomplishment." More broadly, Meta says in its policy update that it will "remove unclear or contextless references if the user's intent is not clearly indicated." The policy specifies that this will include "unclear humor" and "captionless or positive references that do not glorify the designated entity's violence or hate." However, the policy does not provide examples of the kinds of posts that would violate this rule. Meta has informed the Board that it continues to remove all content referring to designated individuals as "shaheed." For such references, the company presumes "shaheed" is violating in all contexts, essentially resulting in a blanket ban on the term when used to refer to a designated individual. The scope of this prohibition can be illustrated in part by consulting the abovementioned U.S. designation lists, as it is prohibited under Meta's policy to refer to all

persons (and members of organizations) on its own derivative list as “shaheed” (or any other translation of the term “martyr”). This includes entities across continents, and is not limited to terrorist organizations or organizations of one particular ideology.

5. The company describes the word “shaheed” as an “honorific” term, used by many communities across cultures, religions and languages. The company acknowledged that the term has “multiple meanings” and was “used to describe someone dying unexpectedly or prematurely, at times referring to an honorable death, such as when one dies in an accident or in a conflict or war.” Meta stated that while there was “no direct equivalent to the term in the English language,” the common English translation is “martyr.” In sending its request to the Board, prior to the December 2023 policy changes, Meta stated that “we presume that the word means ‘martyr’” and “it is because of this use that we have categorized the term as constituting praise under our [Dangerous Organizations and Individuals] policy.” The only reference to this rule in the public-facing Community Standards was as an example of phrases that would violate the praise prohibition. Among other examples, Meta included a phrase that called a convicted U.S. terrorist a “martyr.” In its Arabic version of the policy, Meta used the same example, translating the word “martyr” as “shaheed.” The “martyr”/ “shaheed” example was removed in the December 2023 update of the public-facing Community Standard, although Meta has informed the Board that it continues to be violating and moderators are instructed to remove it.

6. Meta does not apply the Dangerous Organizations and Individuals policy exceptions to the use of “shaheed” in reference to a designated person. These exceptions otherwise allow for “reporting on, neutrally discussing or condemning” designated individuals. Removal of content that includes “shaheed” as “praise” of a designated individual would result in severe “strikes” for users, with accumulated severe strikes leading more swiftly to sanctions such as account or page suspension or disablement. The inapplicability of exceptions to the rule on “shaheed” is not explained in the public-facing Community Standards.

7. Meta has explained that its treatment of “shaheed” advances its value of promoting safety because this content could, in its view, “contribute to a risk of offline harm.” At the same time, Meta acknowledged that the English word “martyr” is not an adequate translation of “shaheed.” Given the multiple meanings of “shaheed” and difficulties in accounting for context and users’ intentions at scale, Meta accepted that it has been removing speech that “does not contribute to a risk of harm” and that is “not intended to praise a designated individual, particularly among Arabic speakers.” For example, the removal of content when “shaheed” is used in news reporting

or to neutrally discuss the premature death of a designated individual, rather than to praise (or glorify) them or their conduct.

8. On August 29, 2023, Meta updated the Dangerous Organizations and Individuals policy allowance to include definitions of "news reporting," "neutral discussion" and "condemnation," as well as illustrative examples of each exception. At the same time, Meta expanded its description of this allowance to recognize that users may reference designated entities "in the context of social and political discourse." The illustrative examples of permitted content were all removed as part of the December 2023 policy update, although the allowance itself remains. The examples were reinserted in a policy update on February 8, 2024. However, neither the December 2023 nor the February 2024 updates make it explicit that unclear or contextless references to designated individuals or organizations will not benefit from these exceptions, since they are defined by a lack of clearly demonstrated intent in the post itself.

9. Meta initiated a policy development process in 2020 to reassess its approach to the term "shaheed." This included a research review and stakeholder consultation. Meta described as a key finding of this stakeholder engagement that the meaning of "shaheed" depends on context, and "that in some instances the term had become desensitized and disconnected from praise." At the same time, without doubt there are instances when "shaheed" is used and understood as praise of a designated individual. Determining the differences in intent in a single post is inherently challenging especially at-scale. During this process, as outlined in its request, Meta identified two policy options as possible alternatives to the use of its current treatment of the word "shaheed." However, there was no consensus among stakeholders regarding which option was better and Meta did not settle on a new approach. The company emphasizes that due to the volume of content on its platforms, a key practical concern is whether enforcement of any modified policy is workable at-scale.

III. Policy Changes That Meta Requested the Board to Consider

10. Meta presented the following policy options for the Board to consider, prior to its replacement of the prohibition on "praise" with a prohibition on "glorification" and "unclear references":
    1) Continue to remove all content that uses "shaheed" to refer to an individual designated as dangerous under the Dangerous Organizations and Individuals policy.
    2) Allow content that refers to a designated individual as "shaheed" when the following conditions are met: (i) it is used in a context that is permitted under the Dangerous Organizations and Individuals policy (e.g., condemnation, news reporting, academic debate,

social and political discourse); (ii) there is no additional praise, representation or substantive support of a designated individual (e.g., the post does not explicitly praise the perpetrator of a terrorist attack or legitimize their violence); and (iii) there is no signal of violence in the content. The signals, as proposed by Meta, are: a visual depiction of an armament; statement of intent or advocacy to use or carry an armament/weapon; reference to military language; reference to arson, looting or other destruction of property; reference to known real-world incidents of violence; and statements of intent, calls to action, representing, supporting or advocating violence against people.

3) Remove content that uses "shaheed" to refer to an individual designated as dangerous under Meta's Dangerous Organizations and Individuals policy only when there is additional praise, representation or substantive support, or signals of violence. These signals are the same as laid out under option two.

11. Both the second and third options seek to arrive at a more contextualized understanding of the use of "shaheed," and do not seem to diverge much in their intended outcomes. The two options appear closer still insofar as Meta has expanded the scope of the contextualized exceptions that it makes to its prohibition on praise (now, glorification) generally (see para. 8 above). As the Board understands, after asking Meta about differences in application and the outcome of these two options, the key difference is the second option would require Meta to seek out and confirm that one of the exceptions (reporting on, neutral discussion, condemnation) applies to the content, while the third option would disregard that step and only consider whether or not the post also has additional praise (now "glorification" or "unclear references") or one of the six listed signals of violence. Regarding all of Meta's proposed options, the policy would be applied in a way that any content naming or depicting designated dangerous individuals that is ambiguous or unclear in its intent is by default treated as violating, thus placing the burden of clarity of intent on the user. The December 2023 changes incorporate this application into the Dangerous Organizations and Individuals Community Standard, explaining that for Tier 1 entities and designated events the policy prohibits "unclear or contextless references," which include "unclear humor, captionless or positive references that do not glorify the designated entity's violence or hate." For example, a photograph of a designated individual, without any further words or commentary, would be removed under the policy prohibiting unclear references because the user's intent was not sufficiently explicit.

12. While the Board considered the options presented by Meta, it also considered others and took the December 2023 policy changes into account. In addition, given that Meta's request directly asked the Board for policy recommendations to address similar challenges in the future, the

Board assessed connected issues relating to Meta's enforcement and transparency practices in areas evident in this policy advisory opinion, but which also have implications for freedom of expression and other human rights more broadly.

IV. Questions the Board Asked Meta

13. The Board asked Meta 41 questions in writing. Questions related to the policy rationale behind the Dangerous Organizations and Individuals Community Standard, evidence of harm that could derive from allowing "praise" on Meta's platforms, Meta's human and automated enforcement processes, Meta's designation process and list of designated entities, Meta's strikes system and the implications in practice of adopting policy option two or three. In October 2023, the Board asked follow-up questions about content trends for the word "shaheed" related to the October 7 terrorist attacks on Israel and the ongoing military response by Israel, and whether Meta's analysis of the policy options it presented to the Board in its request had changed in light of the current crisis. The Board asked three additional questions about Meta's December 29, 2023 policy update. A total of 40 questions were answered and one was partially answered. Meta only provided the list of Tier 1 designated entities to the Board and did not share lists for Tier 2 and 3 entities, explaining that "praise" is only prohibited when referring to Tier 1.

2. Stakeholder Engagement

14. The Oversight Board received 101 public comments that met the terms for submission. A total of 72 comments were submitted from the United States and Canada; 15 from the Middle East and North Africa; eight from Europe; three from Asia Pacific and Oceania; two from Latin America and the Caribbean; and one from Central and South Asia. All were received prior to April 10, 2023. To read the public comments submitted with consent to publish, click here.

15. The submissions covered many issues. Many explained the multiple meanings of the word "shaheed" and, consequently, the negative impact of Meta's default treatment of this term as "praise" on free expression, in particular political speech and human rights documentation. Submissions also touched on concerns about the use of automation in enforcement, as well as Meta's designation list, transparency issues and potential bias in the designation process. Other submissions expressed concern that policy changes could lead to the normalization of terrorist groups and increase violence, particularly in Israel and the Occupied Palestinian Territories.

16. The Board held three regional stakeholder roundtables for Southwest Asia and North Africa, Sub-Saharan Africa and Southeast Asia. In addition, there were two subject matter roundtables, one

on automation in content moderation and one on counterterrorism and human rights. Participants reinforced that the word "shaheed" has many meanings. Martyrdom is one possible meaning, including for the death of individuals in the commission of acts of terrorism, but "shaheed" is also often used in other contexts, such as for describing victims of violence in those attacks. Many participants, including impacted community members, counterterrorism experts and content-moderation experts, expressed concerns about bias in the policy and discussed how it negatively impacts free expression, particularly for Arabic speakers and other communities that use "shaheed." Other themes included the lack of evidence demonstrating a causal link between the use of "shaheed" in reference to designated individuals and real-world harm, which was also emphasized by national security and counterterrorism experts. Equally, there were concerns that not moderating the term at all would allow the normalization of designated individuals and their organizations, which could use social media for recruitment and other forms of substantive support. Further topics included concerns about the quality of automation that Meta uses in content moderation of the term and calls from participants for more transparency around its use, as well as around Meta's list of designated entities and designation processes.

17. For a report on our Stakeholder Engagement Roundtables, please click here (for Arabic version, please click here).

3. Oversight Board Authority and Scope

18. Meta may request policy advisory opinions from the Board (Charter Article 3, Section 7.3) and the Board has discretion to accept or reject Meta's requests (Bylaws Article 2, Section 2.1.3). These opinions are advisory (Charter; Article 3, Section 7.3). Meta is required to respond to this opinion within 60 days of publication (Bylaws Article 2, Section 2.3.2). The Board monitors the implementation of recommendations Meta has committed to act on and may follow up on any prior recommendations in its case decisions.

4. Sources of Authority and Guidance

I. Prior Oversight Board Recommendations

19. In previous cases, the Board has recommended that Meta clarify and narrow the scope of the Dangerous Organizations and Individuals policy, and improve due process and transparency around enforcement.

20. In relation to improving the clarity of the policy and narrowing its scope, the Board has recommended that Meta:
    - Narrow the definition of "praise" in the Known Questions guidance for reviewers (Mention of the Taliban in News Reporting, recommendation no. 3).
    - Revise its internal guidance to make clear that the "reporting" allowance in the Dangerous Organizations and Individuals policy allows for positive statements about designated entities as part of reporting, and how to distinguish this from prohibited "praise" (Mention of the Taliban in News Reporting, recommendation no. 4).
    - Add criteria and illustrative examples to its policy to increase understanding of the exceptions for neutral discussion, condemnation and news reporting (Shared Al Jazeera Post, recommendation no. 1).
    - Update the policy rationale to reflect that respect for freedom of expression and other human rights can advance Meta's value of safety and to specify in greater detail the "real-world harms" the policy seeks to prevent and disrupt when the value of voice is suppressed (Öcalan's Isolation, recommendation no. 4).
    - Explain how users can make the intent behind their posts clear so their posts can benefit from policy exceptions (Öcalan's Isolation, recommendation no. 6).
    - Explain and provide examples of the application of key terms in the Dangerous Organizations and Individuals policy, including the meaning of "praise" and to provide clearer guidance to users on how to make their intent apparent (Nazi Quote, recommendation no. 2).

21. In relation to Meta's strikes system, the Board has recommended that Meta:
    - Explain its strikes and penalties process for restricting profiles, pages, groups and accounts on Facebook and Instagram in a clear, comprehensive and accessible manner (Former President Trump's Suspension, recommendation no. 15).
    - Make its public explanation of its two-track strikes system more comprehensive and accessible, providing more background on "severe strikes" (Mention of the Taliban in News Reporting, recommendation no. 2).
    - Provide users with accessible information on how many violations, strikes and penalties have been assessed against them, as well as the consequences that will follow future violations (Former President Trump's suspension, recommendation no. 16).

22. In relation to transparency, the Board has recommended that Meta:
    - Share its list of designated entities publicly, or at least provide illustrative examples of designated entities (Nazi Quote, recommendation no. 3).

- Improve its enforcement reporting by including numbers of profile, page and account restrictions (in addition to content removal decisions), with information broken down by region and country (Former President Trump's Suspension, recommendation no. 17).
- Include more comprehensive information on error rates for enforcing rules on "praise" and "support" of dangerous individuals and organizations, broken down by region and language (Öcalan's Isolation, recommendation no. 12).
- Increase public information on error rates and make it viewable by country and language for each Community Standard (Punjabi Concern Over the RSS in India, recommendation no. 3).

23. In relation to automation, the Board has recommended that Meta:
    - Inform users when automation is used to take enforcement action against their content, including accessible descriptions of what this means (Breast Cancer Symptoms and Nudity, recommendation no. 3).
    - Expand transparency reporting to disclose data on the number of automated removal decisions per Community Standard, and the proportion of those decisions subsequently reversed following human review (Breast Cancer Symptoms and Nudity, recommendation no. 6).
    - Publish error rates for content mistakenly included in Media Matching Service banks of violating content, broken down by each content policy, in its transparency reporting. This reporting should include information on how content enters the banks and the company's efforts to reduce errors in the process (Colombia Police Cartoon, recommendation no. 3).
    - Provide a public explanation of the automatic prioritization and closure of appeals (Iran Protest Slogan, recommendation no. 7).

24. To view the implementation status of these previous recommendations at the time this opinion was finalized, please click here (for Arabic version, please click here).

II. Meta's Values and Human Rights Responsibilities

25. The Board's analysis and recommendations in this policy advisory opinion were informed by Meta's values and human rights responsibilities.

26. Meta describes voice as a "paramount" value, noting it can be limited in service of four other values, the most relevant of which is safety for this policy advisory opinion. To protect the value of safety, Meta "remove[s] content that could contribute to a risk of harm to the physical security

of persons." It also does not allow "content that threatens people" as it "has the potential to intimidate, exclude or silence others."

27. On March 16, 2021, Meta announced its Corporate Human Rights Policy, in which it outlines its commitment to respecting rights in accordance with the UN Guiding Principles on Business and Human Rights (UNGPs). The UNGPs, endorsed by the UN Human Rights Council in 2011, establish a voluntary framework for the human rights responsibilities of private businesses. These responsibilities mean, among other things, that companies should "avoid infringing on the human rights of others and should address adverse human rights impacts with which they are involved," (Principle 11, UNGPs). Companies are expected to: "(a) Avoid causing or contributing to adverse human rights impacts through their own activities, and address such impacts when they occur; (b) Seek to prevent or mitigate adverse human rights impacts that are directly linked to their operations, products or services by their business relationships, even if they have not contributed to those impacts," (Principle 13, UNGPs).

28. As Meta's request to the Board acknowledges, its content-moderation practices can have adverse impacts on the right to freedom of expression. Article 19, para. 2 of the International Covenant on Civil and Political Rights (ICCPR) provides broad protection for this right, given its importance to political discourse, and the Human Rights Committee has noted that it also protects expression that may be deeply offensive (General Comment No. 34, paras. 11, 13 and 38). 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). These requirements are often referred to as the "three-part test." The Board uses this framework to interpret Meta's voluntary human rights commitments, in relation both to the individual content decision under review and to Meta's broader approach to content governance. As the UN Special Rapporteur for freedom of opinion and 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," (Report A/74/486, para. 41).

29. The right to freedom of expression is guaranteed to all people equally. Any restrictions on this right must be non-discriminatory, including on the basis of religion or belief, language spoken or national origin (Articles 2 and 26, ICCPR).

Legality (Clarity and Accessibility of the Rules)

30. Any restriction on freedom of expression should be accessible and clear enough in scope, meaning and effect to guide users and content reviewers on what content is permitted or prohibited. A lack of clarity or precision can lead to inconsistent and arbitrary enforcement of the rules (UN Special Rapporteur report A/HRC/38/35, para. 46). The Board has previously criticized the lack of clarity of the Dangerous Organizations and Individuals policy and made recommendations for Meta to improve its policy (Öcalan's Isolation, recommendations no. 4 and no. 6; Shared Al Jazeera Post, recommendation no. 1; Mention of the Taliban in News Reporting, recommendations no. 3 and no. 4). Meta subsequently implemented the Board's recommendations to clarify its definition of praise.

31. Meta's December 2023 update to the Dangerous Organizations and Individuals policy, however, introduces new legality concerns. Although Meta provides examples of posts that violate the prohibition on "representation," "support" and "glorification," it does not provide examples of violating "unclear references." The Board's recommendations in this policy advisory opinion aim to further improve clarity and accessibility of Meta's rules.

Legitimate Aim

32. Article 19, para. 3, ICCPR provides that any restriction on expression must pursue a legitimate aim, which includes protection of the rights of others as well as broader societal interests, such as national security (see also General Comment 34, paras. 21 and 30). Meta's policy rationale for the Dangerous Organizations and Individuals Community Standard explains that it seeks to "prevent and disrupt real-world harm," which the Board has found in various cases to be in line with the legitimate aim of protecting the rights of others, including the right to life (Article 6, ICCPR). The Board has also previously recognized that praise of designated entities may pose risks of harm to the rights of others and that seeking to mitigate those harms through the prohibition on praise in the Dangerous Organizations and Individuals Community Standard has a legitimate aim (Mention of the Taliban in News Reporting).

Necessity and Proportionality

33. Any restrictions on freedom of 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 34, para. 34).

34. UN Security Council Resolution 1624 (2005) calls upon states, as may be necessary and appropriate, and in accordance with their obligations under international law, to "prohibit by law incitement to commit a terrorist act or acts, (para. 1a), and subsequent resolutions have expressed concern about the use of the internet by terrorist organizations (UNSC Resolution 2178 (2014) and UNSC Resolution 2396 (2017)). While these resolutions reaffirm that states should address and prevent terrorism in line with their obligations under international human rights law, the UN Special Rapporteur on the promotion and protection of human rights and fundamental freedoms while countering terrorism has raised concerns about both the vagueness of laws attempting to implement their obligations, including those relating to online expression, and also their focus on the content of speech rather than on the speaker's intent or the impact of that speech on others (Report A/HRC/40/52, para 37). The Special Rapporteur concluded that criminalizing terrorist propaganda "requires the reasonable probability that the expression in question would succeed in inciting a terrorist act, thus establishing a degree of causal link or actual risk of the proscribed result occurring," (*ibid.*). The Joint Declaration on the Internet and on Anti-Terrorism Measures of the UN Special Rapporteur on freedom of expression, the OSCE Representative on freedom of the media and the OAS Special Rapporteur on freedom of expression, from 21 December 2005, states that without clearly showing that the speech in question constitutes a direct and intentional call for others to engage in terrorist acts, and directly increases the likelihood of a terrorist act occurring, a state may not restrict and sanction speech by punitive measures (p. 39). A high threshold is also set out by the Special Rapporteur on freedom of expression in report A/HRC/17/27 (May 16, 2011, para. 36), noting that expression should only be limited for national security or counterterrorism purposes when "(a) the expression is intended to incite imminent violence; (b) it is likely to incite such violence; and (c) there is a direct and immediate connection between the expression and the likelihood or occurrence of such violence." According to the Human Rights Committee, when counterterrorism measures prohibit certain speech, terms such as "encouragement," "glorification" and "praise" must be "narrowly defined" and must not "unduly restrict the crucial role of the media in informing the public about acts of terrorism," (General Comment 34, paras. 30 and 46).

35. While these principles and standards are an important starting point for the Board's analysis, the obligations and limitations that international human rights law imposes on states are not identical to the responsibilities and discretion that a private company may have in this sphere. The application of criminal sanctions by a state is not equivalent to moderating content on a social media platform, and the Board recognizes that Meta, as a private company rather than a state, can and sometimes does take an approach to expression that is more restrictive than what would be justifiable by a state, taking into account both its company values (see paras. 27 - 28

above) and the distinctive challenges of moderating content at-scale. According to the UN Special Rapporteur on freedom of expression: "When company rules differ from international standards, the companies should give a reasoned explanation of the policy difference in advance, in a way that articulates the variation," (report on online hate speech of the Special Rapporteur on freedom of expression, A/74/486, October 9, 2019, para. 48). In many of its prior decisions, the Board has explored how to appropriately translate international standards designed for states to a company's human rights responsibilities, when assessing the requirement of proportionality, as it also seeks to achieve with this policy advisory opinion.

36. Prior Board decisions relating to the Dangerous Organizations and Individuals Community Standard have also examined Meta's human rights responsibilities around "praise" of designated entities. While preventing the abuse of Meta's platforms by designated entities that seek to incite violence, and to recruit or engage in other forms of material support, is a legitimate aim, the Board has found various instances in which the breadth or imprecision of the prohibition on "praise" has unnecessarily and disproportionately restricted user expression. For example, the Board overturned the removal of an Urdu-language newspaper's post about a Taliban announcement on plans to re-establish women and girls' education, finding it was not "praise" (Mention of the Taliban in News Reporting). It similarly reversed Meta's removal of a user's post misattributing a quote to Joseph Goebbels (Nazi Quote) because there was sufficient context to make clear the post was not praising Nazi ideology but engaging in political discussion in the United States. The Board also overturned Meta's decision to remove a post sharing an Al Jazeera news article reporting a designated terrorist group's threat of violence, which should have remained on the platform, raising concerns about the potentially discriminatory impacts of Meta's policies (Shared Al Jazeera Post). In summary and expedited decisions, the Board has also reversed or prompted reversal of several Meta decisions to remove Facebook and Instagram posts initially removed as praise of designated individuals (Anti-Colonial Leader Amílcar Cabral; Lebanese Activist; Responding to Antisemitism; Federal Constituency in Nigeria; Girls' Education in Afghanistan; Mention of Al-Shabaab; Praise Be to God; and Hostages Kidnapped From Israel). Many of these decisions are still relevant, notwithstanding Meta's recent changes to the Dangerous Organizations and Individuals policy. The Board will continue examination of this policy in future cases, including in relation to the new provisions on glorification and "unclear references."

6. Recommendations and Analysis

37. The Oversight Board issues seven recommendations for Meta across content policy, enforcement and transparency. The Board believes these recommendations are operable at-scale and will

advance Meta's adherence to its values of voice and safety, while also increasing its respect for freedom of expression and other human rights, as well as transparency.

6.1 Meta's "Shaheed" policy

Recommendation 1 – Content Policy: Meta should stop presuming that the word "shaheed," when used to refer to a designated individual or unnamed members of designated organizations, is always violating and ineligible for policy exceptions. Content referring to a designated individual as "shaheed" should be removed as an "unclear reference" in only two situations. First, when one or more of three signals of violence are present: a visual depiction of an armament/weapon, a statement of intent or advocacy to use or carry an armament/weapon, or a reference to a designated event. Second, when the content otherwise violates Meta's policies (e.g., for glorification or because the reference to a designated individual remains unclear for reasons other than use of "shaheed"). In either scenario, content should still be eligible for the "reporting on, neutrally discussing and condemning" exceptions.

The Board will consider this recommendation implemented when Meta publicly updates its Community Standards to specify that references to designated individuals as "shaheed" are not allowed when one or more of the three listed signals of violence are present.

Recommendation 2 – Content Policy: To clarify the prohibition on "unclear references," Meta should include several examples of violating content, including a post referring to a designated individual as "shaheed" combined with one or more of the three signals of violence specified in recommendation no. 1.

The Board will consider this recommendation implemented when Meta publicly updates its Community Standards with examples of "unclear references."

Recommendation 3 – Enforcement: Meta's internal policy guidance should also be updated to make clear that referring to designated individuals as "shaheed" is not violating except when accompanied by signals of violence, and that even when those signals are present, the content may still benefit from the "reporting on, neutrally discussing or condemning" exceptions.

The Board will consider this recommendation implemented when Meta updates its guidance to reviewers allowing "shaheed" to benefit from the reporting on, neutrally discussing or condemning exceptions, and shares this revised guidance with the Board.

38. While it is imperative Meta seeks to prevent its platforms from being used to incite acts of terrorist violence – a legitimate aim of its content moderation policies – Meta's human rights responsibilities require any limitations on expression to be necessary and proportionate, not least to respect the voice of people in communities impacted by violence. Even though one meaning of "shaheed" does correspond to the English word "martyr" and is used in that way, the Board finds it is not necessary or proportionate for Meta to remove all content solely for use of the word "shaheed" when referring to designated individuals. This is because categorical prohibition fails to account for the term's linguistic complexity, leads to over-removal of speech, unduly restricts media freedom and civic space, and has serious negative implications for equality and non-discrimination. In addition, should this recommendation be accepted and implemented, the details of Meta's policy prohibiting "glorification" and "unclear references" more generally will continue to be applicable. Meta should therefore cease to remove content based solely on the presence of the word "shaheed." Instead, Meta should adopt a more contextual approach, and define clearly and narrowly the signals of violence that will lead to "shaheed" being interpreted as harmful. As explained below, the Board is only fully endorsing two of the six signals Meta proposed in its request, and recommends the narrowing of a third. The other three that Meta proposed – and which are provided as guidance to reviewers enforcing the separate Violence and Incitement policy – are too broad for the purposes of the Dangerous Organizations and Individuals policy. At the same time, the Board recommends that Meta apply exceptions for content that reports on, condemns or neutrally discusses a designated entity when signals of violence are present. The above recommendations apply to "shaheed" in its singular and plural forms, and to variants of "shaheed" in the many languages that have adopted the term.

39. The Board consulted experts and stakeholders to assess the potential harms of allowing content that uses the term "shaheed" to refer to designated entities on Meta's platforms. Some stakeholders were concerned about Meta's platforms facilitating propaganda and recruitment efforts of terrorists as well as fostering discrimination or violence, including against Jewish people (see public comments, e.g., PC-11123 – CyberWell Ltd; PC-11153 – Stop Antisemitism Now; PC-11194 – Committee for Accuracy in the Middle East; PC-11068 – Canadian Antisemitism Education Foundation). While there is no research showing any concrete causal connection between "shaheed" and an increase in such violence or discrimination (as noted previously in Meta's referral at page 10, and which followed from expert and stakeholder engagement that Meta itself conducted prior to requesting the policy advisory opinion), there are individual cases linking a desire for martyrdom with violent acts. More generally, concerns that social media can

be used by violent organizations to recruit and promote the commission of terrorist acts, to normalize terrorism and facilitate extremism are widespread. The extensive stakeholder engagement and expert consultations carried out by the Board, which included a roundtable with counterterrorism experts (see Stakeholder Engagement Roundtables), confirmed these assumptions and concerns.

40. Numerous expert and stakeholder submissions to the Board also described negative impacts on freedom of expression resulting from removing content that uses the term "shaheed" (see public comments, e.g., PC-11196 – Integrity Institute; PC-11164 – SMEX, PC-11183 ECNL EFF, PC-11190 – Brennan Centre; PC-11188 – Digital Rights Foundation). This includes the removal of speech using the term "shaheed" not to praise or glorify designated entities but to report on violence by terrorist or other designated organizations, or to engage in neutral political or academic discussion of designated individuals.

41. The following examples may illustrate Meta's approach when content referring to a designated individual as "shaheed" is removed:
    - A government shares a press release on Meta's platforms that confirms the death of a designated individual, using the honorific term "shaheed" alongside their name. This post would be removed as Meta presumes the term "shaheed" to be violating, regardless of context indicating this is neutral discussion.
    - A user shares a photo of a protest they are attending, without a clear caption explaining the protest's purpose. Several placards name a deceased designated individual with the phrase "shaheed." This post would be removed as Meta presumes the image featuring placards naming the designated individual, combined with the term "shaheed," is violating.
    - A human rights defender shares a post including a description of a summary execution of a designated individual by the state, referring to them as "shaheed" and condemning a government's counterterrorism policy. This post would be removed as Meta presumes the term "shaheed" to be violating, regardless of the reporting context.
    - A concerned community member posts a complaint about the state of a local road, which is named after a designated individual and includes the honorific term "shaheed." This would be removed as Meta presumes the term "shaheed" to be violating, regardless that the mention of the designated individual is incidental to a discussion about local matters.
    - A family member refers to a loved one killed in a terrorist attack as "shaheed" and condemns the attackers, who are designated individuals. Although this post would be non-violating, the combination of the use of "shaheed" in a context where the perpetrators are also named is illustrative of content that invites erroneous removal, given Meta's categorical approach.

42. Some stakeholders expressed concern that a less restrictive approach to the use of "shaheed" could have the cumulative effect, at scale, of normalizing terrorism. While it cannot be denied that in some cases "shaheed" is intended to and does constitute a form of approval or endorsement of a person and their violent acts, the Board does not consider the policy's blanket ban on the word to be justified on those grounds. In light of the information received by the Board from stakeholders, including counterterrorism experts and from Meta's prior research and stakeholder consultations, the Board concludes that concerns about the possibilities that such cases could have substantial cumulative harmful effects are outweighed by the very tangible negative impact of the existing policy on freedom of expression. This is especially the case because, even if the Board's recommendations are implemented, significant guardrails remain in place to advance the goals of preventing violence and other harms resulting from allowing terrorists and their supporters to freely use Meta's platforms.

43. The Board also considered that "shaheed" has various meanings and is often used in ways that do not incite violence or hatred but rather to report on matters of public interest. Enforcing a blanket ban means removing a lot of content that does not pose the harms that Meta's policies seek to mitigate against. While moderating content at-scale sometimes requires accepting a percentage of enforcement errors, Meta has a responsibility to weigh trade-offs based on all its human rights responsibilities. Based on all these considerations, the Board reaches the conclusion that its proposed approach best reconciles Meta's overall human rights responsibilities.

Cultural and religious significance of "shaheed" and diversity of meanings

44. "Shaheed" is a culturally and religiously important term used in many different contexts and has various meanings. Public comments submitted to the Board and research commissioned by the Board provided further insights on the term. "Shaheed" can mean "to bear witness" and "to testify." It is also used as an Islamic first name and in some regions, including West Asia and North Africa, as a last name (see public comments, e.g., PC-11196 – Integrity Institute; PC-11164 – SMEX). "Shaheed" is also used to refer to people who die while performing a religious duty. In armed conflicts or violent attacks, including incidents in which designated organizations are involved, people sometimes commemorate the victims of violence and terrorism by referring to them as "shaheed" (see public comments, e.g., PC-11164 – SMEX; PC-11197 by the New Zealand government acknowledged victims of the Christchurch shooting as "shuhada," which is the plural of "shaheed"; PC-11196 from the Integrity Institute highlights that marginalized Muslim communities, including the Rohingya, in some countries referred to refugees who have been

forced out of Myanmar due to persecution based on religion or belief as "shaheed"). Quantitative research on content that has been posted and remains on Meta's platforms, which the Board commissioned from Memetica, indicated that "shaheed" is widely used to refer to individuals who die while serving their country, serving their cause or as an unexpected victim of sociopolitical violence or natural tragedy. While those posts did not necessarily associate "shaheed" with a designated individual, the examples cited earlier (see para. 41) illustrate how the many meanings and uses of "shaheed" are also relevant in that context.

45. Following the October 2023, Hamas-led terrorist attacks on Israel and ensuing Israeli military action in Gaza, the Board commissioned additional research into content trends around use of "shaheed." Out of more than 12,000 public posts reviewed in this analysis, only two were supportive of Hamas. Nearly all uses related to descriptions of Palestinian casualties in the Gaza strip, without mention of any designated organization or individual. Posts included Palestinians mourning the loss of their loved ones, while others showed wounded Palestinian children and/or the aftermath of airstrikes in Gaza. Additional questions were asked to Meta, since researcher access to content trends is limited to public content that remains on the platforms, and does not include content removed by Meta for violating its Dangerous Organizations and Individuals policy. Meta's response reinforced the findings of the research. In the company's view, content trends during the crisis did not demonstrate any change in the use or understanding of the word "shaheed" compared to before these events.

46. These various meanings indicate that Meta's default presumption that the use of "shaheed" is violating and leads to real-world harm, and should therefore always be removed, imposes global costs on free expression that must be addressed. Meta explained that "shaheed" is likely the most moderated word on its platforms and various stakeholders observed that content reporting on designated entities and violence is often removed by Meta (policy advisory opinion request, at page 3). This approach to "shaheed" is an example of the company prioritizing bright-line (obvious) distinctions for enforcement at-scale. While in principle such distinctions may sometimes be appropriate because content moderation at-scale often entails difficult trade-offs, the Board concludes in this case that the default presumption that "shaheed" must always be removed in connection with a designated individual is not justified and must be dropped. As Meta admits, its approach fails to account for linguistic and cultural context by design and results in significant over-enforcement, in two ways. First, it results in removal of content that is not intended as praise or glorification but that nevertheless violates the policy because Meta has chosen to presume "shaheed" is always violating. Second, it fosters higher rates of erroneous removal of content that uses "shaheed" in connection with persons who are *not* designated

individuals because the approach entrenches a categorical approach and discourages more contextual evaluations.

47. Notwithstanding the various meanings of "shaheed," the Board notes that the examples previously included under Meta's Dangerous Organizations and Individuals "praise" prohibition were not contextualized for different markets, with overly simplistic direct translations of those examples. The only indication in the Community Standards of Meta's approach to content containing "shaheed" was an example of a phrase that calls a convicted terrorist a "martyr," listed under its more broadly stated prohibition on praise. This same example was used across languages and for Arabic was translated to "shaheed," despite the terms not necessarily being equivalent. The Board notes that this example has been removed in the December 2023 update to the policy. While the translation in this example was not necessarily inaccurate, using that single example to assert that "shaheed" always equates to "martyr" is not accurate.

48. For the same reasons in reverse, the Board's conclusions on "shaheed" in this policy advisory opinion should not be automatically transferred to the term "martyr" and all translations of that term into other languages. "Martyr" and "shaheed" are not directly equivalent, as the former is less semantically variable in the English language, within and across regions where English is spoken. Across languages, Meta will have to assess which terms meet Meta's threshold for "glorification," keeping in mind the guidance set out in this policy advisory opinion. In any case, the Board does not endorse blanket prohibitions that isolate and prohibit the use of "shaheed" from the broader context of posts in which it is used. Consistent with Oversight Board precedent, these posts should be assessed in their entirety. Additionally, consistent with its public-facing policies, Meta should ensure policy exceptions allow people to "report on, condemn or neutrally discuss" designated individuals when the term "martyr" is used (and genuinely corresponding terms of glorification or positive reference in other languages). This underscores the importance of internal guidance for reviewers being adapted to local context and the linguistic specificities of the language being reviewed (see similar concerns expressed in the Board's Myanmar Bot decision, which were the basis for recommendation no. 1 in that case).

49. In several cases, the Board has pushed Meta to recalibrate the trade-off between having bright-line rules that are enforceable at-scale and policies that give greater attention to context for enforcement, often favoring freedom of expression and user voice (e.g., to respect reporting on designated entities, see Mention of the Taliban in News Reporting and Shared Al Jazeera Post). While prior cases have not dealt directly with the use of "shaheed," the Board has cautioned against presumptions of ill-intent when people reference designated entities (see Nazi Quote).

Various public comments highlight how Meta's presumption that "shaheed" always equals "praise" could, as a result of mistakes made by moderators or automation at-scale, lead to over-enforcement against people commemorating the victims of terrorist attacks (see public comments, e.g., PC-11164 – SMEX, p.1; comment PC-11197 by the New Zealand government). Removing such content may even work against the purposes of the Dangerous Organizations and Individuals policy because it may prevent the discussion of terrorist violence and efforts to counter the ideologies behind such violence (see e.g., PC-11164 – SMEX).

Respect for media freedom and civic space

50. The Board is especially concerned that Meta's approach impacts journalism and civic discourse. Those effects are acute in locations where violence by terrorist or other designated organizations is more prevalent, and in places where designated entities have territorial control or otherwise hold political power. Various public comments raised these concerns (see public comments, e.g., PC-11196 – Integrity Institute; PC-11157 – Palestine Institute for Public Diplomacy). Media organizations and other commentators may be wary of reporting on designated entities because they will want to avoid content removals that can result in severe sanctions, including the removal of pages or groups, or disabling of accounts. In some cases, this might be unpredictable – for example, reporting on a protest against a government actor during which people are displaying signs of support for designated individuals who have been killed by that same actor. Linguistic, cultural or religious conventions could dictate that "shaheed" is used to refer to persons killed in these situations, including people who died while committing acts of violence.

51. It is important to note the complexity of situations in which designated entities might be engaged in violence, including in situations that may qualify as armed conflicts, resistance to foreign occupation and civil unrest. The media can face significant challenges in reporting on those situations, as outlets need to be sensitive to local linguistic, cultural or religious conventions while also complying with Meta's content policies. These concerns are compounded by the lack of transparency around Meta's list of designated entities and the fact that users do not actually know who is on the list. Reporting on those situations may also often include signals of violence, such as the depiction of an armament, which outside the context of the reporting on, neutral discussion or condemnation exceptions, can be an indicator of a policy violation. It is therefore important that Meta applies a policy exception in instances in which content reports on, neutrally discusses or condemns a designated entity.

52. The prohibition may also make it more difficult for users to express their views and criticism when living in contexts where designated entities operate or hold political power. To avoid direct

confrontation and safety risks, users in these contexts may formulate commentary in a respectful manner that should not necessarily be interpreted as praise. Sanctioning such speech is especially concerning, as social media provides an important forum for debate, particularly in conflict zones and countries where press freedom is limited.

Equality and non-discrimination

53. Because of the widespread and highly variable use of "shaheed," particularly among Arabic speakers and speakers (many of them Muslim) of other languages that have "shaheed" loanwords, Meta's policy has a disproportionate impact on these communities. This raises significant concerns with respect to the company's responsibility to respect all users' rights to equality and non-discrimination in the exercise of their rights to freedom of expression. This was a focus of many public comments (see e.g., PC-11183 ECNL EFF, PC-11190 – Brennan Centre; PC-11188 – Digital Rights Foundation; PC-11196 – Integrity Institute; see also the BSR report on Meta's Impact in Israel and Palestine). The UN Development Programme (UNDP), in its reporting on "interrupting the journey to extremism," highlights that key drivers of extremist violence are, among other root causes, "inequality, exclusion, lack of opportunities and perceptions of injustice." The Board is concerned that Meta's policy approach may compound experiences of marginalization and exclusion for certain communities, and even be counterproductive to the stated aim of reducing violence. The UNDP highlights that in relation to content moderation on online platforms, companies should ensure that prevention of violent extremism work "does not unintentionally lead to the stigmatization and targeting of individuals."

Less intrusive means of preventing harm

54. An assessment of the necessity and proportionality of the policy also requires an analysis of whether there are less restrictive means to achieving the legitimate aim in question. The Board's recommendations in this opinion would, if implemented, require Meta to stop categorically removing all uses of "shaheed" to refer to designated individuals. Nevertheless, content using this term could still be removed in narrower circumstances where the connection to harm is less ambiguous and more pronounced. The recommendations would allow Meta to continue to remove content referring to designated individuals as "shaheed" when either accompanied by other policy violations (e.g., glorification or unclear references), or the content includes signals of violence. This will continue to provide an important constraint, while reducing the adverse impacts of the policy, including over-enforcement against specific linguistic and religious groups.

55. The presence of signals of violence alongside a reference to a designated individual as "shaheed," provide a clearer indication the post is intended to refer positively to the individual

because of its association with violence (provided the exceptions for reporting on, neutrally discussing or condemning do not apply). In its request, Meta proposed the following signals of violence, borrowing them from internal guidance for its Violence and Incitement policy: (1) a visual depiction of an armament; (2) a statement of intent or advocacy to use or carry an armament/weapon; (3) reference to military language; (4) reference to arson, looting or other destruction of property; (5) reference to known real-world incidents of violence; and (6) statements of intent, calls to action, representing, supporting or advocating violence against people. The Board finds that some of these signals provide helpful indicators of content that could cause harm, which cannot be presumed from the use of "shaheed" alone. However, other signals are not considered so helpful here. The public-facing Violence and Incitement policy refers to "temporary signals of a heightened risk of violence" three times, each in relation to rules designed to address risks of violence in specific locations and at time-bound events, such as at polling stations and at protests. The wholesale application of all these signals to guide enforcement references to "shaheed" in all contexts, for which the likelihood of harm is not comparably likely or imminent, risks disproportionate enforcement.

56. The Board finds that in the broader context of the Dangerous Organizations and Individuals policy, a narrower set of signals is appropriate. The signal of a visual depiction of an armament, and the signal of a statement of intent or advocacy to use or carry an armament or weapon, should both be taken as indicators of content that are more likely to contribute to harm. The Board finds that the signal of "reference to known real-world incidents of violence" is too broad, however, encompassing too many scenarios removed from the violence of designated organizations or individuals. Therefore, it should be defined more narrowly and restricted to events designated under Meta's Dangerous Organizations and Individuals Policy, which include terrorist attacks, hate events, multiple-victim violence or attempted multiple-victim violence, serial murders and hate crimes. The three signals recommended by the Board allow for greater contextual nuance to help ensure less ambiguous uses of "shaheed," which more expressly refer to acts of terrorist violence, are still removed.

57. The remaining signals proposed by Meta of "references to military language" and "to arson, looting or other destruction of property" are too broad and would lead to imprecise enforcement. The final proposed signal of "statements of intent, calls to action, representing, supporting or advocating violence against people" would independently lead to content violating the Violence and Incitement Community Standard – and therefore would be duplicative and unnecessary if included as a signal of violence.

58. Distinct from Meta's third policy option, the Board notes that even when references to designated individuals are made alongside these signals of violence, they may still not intend to glorify those individuals or their acts. In its request (at pages 9 – 10), Meta explains the potential for over-enforcement, referencing media reporting practices in Pakistan. In the Board's view, it would be common where referring to the death of a designated individual for media to include imagery showing them with armaments, and/or to show imagery of the destruction they have caused, while not glorifying those actions. For this reason, it is necessary that at-scale reviewers are trained to leave content up if there is clear intent to report on, condemn or neutrally discuss the entity.

59. Moreover, the Board's recommended policy changes would leave intact a wide range of other content policies to protect against terrorists and their supporters using Meta's platforms to cause real-world harm. Posts including "shaheed" in reference to a Tier 1 designated individual would be removed in the following circumstances:
    - When accompanied with one or more of the abovementioned three signals of violence.
    - When the post includes any other Dangerous Organizations and Individuals policy violation (e.g., for representation, support or glorification of a designated individual). Glorification, for example, would include legitimizing or defending the violent or hateful acts of a designated individual, or celebrating their violence or hate as an achievement.
    - When the post includes any other policy violation, including for hate speech or violence and incitement.
    - When the post has "unclear or contextless" references. This captures "unclear humor, captionless or positive references that do not glorify the designated entity's violence or hate."

60. The Board's recommendations will therefore not prevent Meta from requiring users to ensure the intent of their posts are clear when referring to a designated Tier 1 individual as "shaheed." It only prevents Meta from relying on the term "shaheed" alone to assert that a post is violating. Meta's active involvement in the Global Internet Forum to Counter Terrorism and use of its hash-matching database also ensure content that may result in real-world harm will be removed. The Board's recommendations will allow for clearer, more contextualized and proportionate enforcement against content that could increase risks of violence, while enhancing respect for freedom of expression and non-discrimination.

Evaluation of alternative policy options

61. For these reasons, the Board concludes that continuing with the policy status quo, which Meta presented as one option for the Board to consider, would entrench unjustified restrictions on free expression that disproportionately affect Arabic speakers, relevant linguistic communities and Muslims. Meta's second proposed policy option is similar to its third option, which the Board recommends, insofar as it seeks to contextualize use of the word "shaheed" and allow it to be used in connection with dangerous individuals in several ways. In fact, according to Meta the August 29, 2023 changes to the policy allow for more social and political discourse in certain instances, including in relation to peace agreements, elections, human rights related issues, news reporting and academic, neutral and condemning discussions. This makes the policy outcomes sought from the second and third options even closer. The most significant difference between them regards their technical feasibility and practical implementation at-scale. The second option requires Meta first to affirmatively establish that one of the allowed exceptions applies, and then to examine whether it contains policy violations or signals of violence; the third option goes directly to the examination of whether policy violations or signals of violence are present. According to Meta, the second option would be significantly more cumbersome to design and implement technically because it requires consideration of a broad set of possible exceptions (even broader following recent policy changes), each of which is highly dependent on an evaluation of context. The difficulty of making those assessments accurately and consistently in turn means that it is likely to require much more human review and result in substantially higher numbers of enforcement errors.

62. The Board considers it preferable for Meta to rely on an approach that examines more uniformly all uses of "shaheed" in relation to designated individuals, beyond existing exceptions, and removes those uses when accompanied by other policy violations or one of the three signals of violence listed in recommendation no. 1. Changing the policy to adopt a position closer to Meta's third proposed option to treat "shaheed" as a prohibited reference only when other policy violations or specific signals of violence are present, as the Board recommends, will help reduce false positives (mistaken removals of non-violating content), better protect public interest speech, media freedom and civic discourse, and reduce the negative impact on the right to equality and non-discrimination among affected groups. In response to the Board's questions following the October 2023, Hamas-led attacks on Israel and the ensuing conflict, Meta confirmed that these events did not change its analysis of the scalability of the options initially presented in its request. In the Board's view, the proposed recommendations, based on Meta's third option, would be most resilient to potential errors when a crisis of this magnitude unfolds,

provided the policy exceptions for "reporting on, neutral discussion and condemnation" remain available. The Board acknowledges that making these policy exceptions available does add some greater complexity to the scalability of the policy proposed by Meta's third option, but concludes they are essential to the adequate protection of freedom of expression in connection with moderating content using the word "shaheed."

63. A minority of Board Members disagree with this conclusion and would prefer to recommend either maintaining the status quo or that Meta adopt the second proposed option. Some emphasize that the word "shaheed" is, in fact, used by terrorist organizations to signify praise and glorification of people who commit violent acts, and this can be an incentive to radicalization and recruitment. For these Board Members, these facts alone represent a grave enough danger of real-world harm to warrant maintaining the current categorical prohibition, even at the recognized expense to freedom of expression. Others regard the absence of adequate data on the extent of real-world harm that can be linked to the use of "shaheed" on Meta's platforms, and on the prevalence of over-enforcement taking place under the policy, as a reason for adopting a more cautious approach, preferring to err on the side of greater safety. Some Board Members also believe that the second option – in which Meta would only consider the meaning and use of "shaheed" when specific allowable exceptions are first found to be applicable – to be a more narrowly tailored compromise position, notwithstanding the implications of its technical difficulties and feasibility at-scale. The Board considered and discussed these alternatives extensively. For the reasons outlined in the preceding sections, the majority reach the conclusion that the recommendations based on Meta's third option present the preferable balance.

Strikes and penalties

64. While maintaining other restrictions under the Dangerous Organizations and Individuals policy may be justified, Meta must ensure these restrictions and any sanctions imposed for violations are proportionate. Meta has previously informed the Board that it always applies severe strikes for all violations of its Dangerous Organizations and Individuals policy, including for its previous prohibition on "praise." Severe strikes can lead more quickly to account level penalties, such as time-bound feature limits, account suspension and permanent disabling of profiles or pages. These measures can severely constrain users' rights to freedom of expression (also emphasized in public comments, see e.g., PC-11190 – Brennan Centre). The Board previously recommended that Meta should improve clarity and transparency of its strikes system (Mention of the Taliban in News Reporting, recommendation no. 2; Former President Trump's Suspension, recommendation no. 15). In response, Meta provided information on the strikes system in its Transparency Center and allows users to review what penalties Meta has applied to their

accounts. Meta recently updated its system for standard strikes to make it more proportionate, but this does not clarify for the public whether Meta's approach to "severe strikes" has also been adapted.

65. Meta shared that it is working on changes to the enforcement system related to the Dangerous Organizations and Individuals policy, though its explanation of its December 2023 policy update does not explain if or how its enforcement system has or will change (e.g., in relation to severe strikes). The Board welcomes changes to the enforcement system if it results in more proportionate penalties for content accurately identified as violating, in line with previous Board recommendations. Severe strikes are disproportionate penalties for more mild and more semantically ambiguous violations of the Dangerous Organizations and Individuals policy, resulting in the policy continuing to have a high rate of over-enforcement and unfair treatment of users.

6.2 Improve Transparency and Auditing of the List

Recommendation 4 – Transparency: To improve the transparency of its designated entities and events list, Meta should explain in more detail the procedure by which entities and events are designated. It should also publish aggregated information on its designation list on a regular basis, including the total number of entities within each tier of its list, as well as how many were added and removed from each tier in the past year.

The Board will consider this implemented when Meta publishes the requested information in its Transparency Center.

Recommendation 5 – Enforcement: To ensure the Dangerous Organizations and Individuals entity list is up-to-date and does not include organizations, individuals and events that no longer meet Meta's definition for designation, the company should introduce a clear and effective process for regularly auditing designations and removing those no longer satisfying published criteria.

The Board will consider this implemented when Meta has created such an audit process and explains the process in its Transparency Center.

66. According to Meta's previous Dangerous Organizations and Individuals policy, "praise" and use of the term "shaheed" was only prohibited if it referred to an entity designated by Meta as a Tier 1 dangerous individual (including perpetrators of designated "violating violent events"). Following the December 2023 update, the policy now prohibits "unclear references" and glorification of all Tier 1 designated entities. "Unclear references" can include "unclear humor, contextless references and positive references that do not glorify the designated entity's violence or hate." The potential adverse impacts of such prohibitions on freedom of expression largely depend on what entities Meta designates as Tier 1 organizations and which "violating violent events" it designates. In roundtables and public comments, many stakeholders criticized a lack of transparency and due process around the list, arguing that Meta should publish it (see public comments, e.g., PC-11164 – SMEX, p.3; PC-11157 – Palestine Institute for Public Diplomacy). One stakeholder was concerned about sharing the list broadly, arguing this could create safety concerns. Since the list is so relevant to understanding the scope of Meta's policies, the company provided the Board with its list of Tier 1 designated organizations and individuals and an explanation of its designation processes, which the Board studied in-depth, as discussed below.

67. As noted above, many of the lists Meta derives its Tier 1 designations from are made public by the United States government. These include "specially designated narcotics trafficking kingpins (SDNTKs)," "foreign terrorist organizations (FTOs)" and "specially designated global terrorists." These lists are extensive, include entities (and their members) in highly politicized contexts across several continents and are not limited to terrorist entities. This breadth and diversity are indicative of the extent to which Meta's Dangerous Organizations and Individuals policy prohibitions may reach, including the more categorical prohibition on "shaheed" in reference to designated individuals – and their impact.

68. Meta explained that its process for reviewing and approving entities to be added to the Dangerous Organizations and Individuals list "relies on consultation with a range of internal experts, the assessment of internal data and external research." Meta also has a process in place to document "additional members, aliases, symbols, slogans, sub-groups or media wings related to already designated organizations and individuals." In 2022, based on Meta's assessment of the Dangerous Organizations and Individuals threats to its platforms, the company designated fewer than 1,000 entities. Criminal entities made up the largest number of designations, followed closely by terrorism entities and then hate entities. Meta has a delisting policy process that typically examines whether an entity continues to meet the policy threshold for designation as a dangerous organization or individual, and considers "active steps that an entity has taken to cease violent acts and pursue peace." The process is currently undergoing redevelopment, but it

was applied fewer than 10 times in 2022. Meta explained that it is auditing previous designations as part of a continuous effort to keep them as up-to-date as possible and accurately reflect threats from dangerous organizations and individuals.

69. Meta explained that sharing its list of designated entities or events publicly, or notifying entities when they are added to the list, could pose some risks to the effectiveness of enforcement and to the safety of many Meta employees. When entities have been designated and have discovered this, some have brought legal action against Meta (see the *Facebook v. CasaPound* case, April 29, 2020, Court of Rome, Italy, and analysis here). The Board notes, that in some cases, Meta has disclosed designations of some groups in response to media queries, and designations have been shared through Oversight Board decisions.

70. The Board previously recommended that Meta share its list of designated entities publicly, or at least provide illustrative examples (Nazi Quote, recommendation no. 3). Meta has not published the list and provided no further updates on this recommendation following a feasibility assessment. If Meta continues to decline to implement this recommendation, it should at a minimum take other measures to improve transparency around the list. Publishing aggregated data, improving transparency on the designation process and making the de-designation process more effective could improve user awareness of Meta's rules and processes. This would also contribute to greater scrutiny and accountability of Meta's designations and the consequences for freedom of expression and other human rights. Meta might also consider facilitating researcher access to data to enhance transparency while still preserving the list's confidentiality.

6.3 Data to Assess Enforcement Accuracy and Testing of Classifiers

Recommendation 6 – Transparency: To improve transparency of Meta's enforcement, including regional differences among markets and languages, Meta should explain the methods it uses to assess the accuracy of human review and the performance of automated systems in the enforcement of its Dangerous Organizations and Individuals policy. It should also periodically share the outcome of performance assessments of classifiers used in enforcement of the same policy, providing results in a way that allows these assessments to be compared across languages and/or regions.

The Board will consider this implemented when Meta includes this information in its Transparency Center and in the Community Standards Enforcement Reports.

Recommendation 7 – Transparency: To inform stakeholders, Meta should provide explanations in clear language on how classifiers are used to generate predictions of policy violations. Meta should also explain how it sets thresholds for either taking no action, lining content up for human review or removing content by describing the processes through which these thresholds are set. This information should be provided in the company's Transparency Center.

The Board will consider this implemented when Meta publishes the requested information in its Transparency Center.

71. Many stakeholders, at roundtables and in public comments, asserted that Meta's enforcement of its Dangerous Organizations and Individuals policy, and treatment of the word "shaheed" in both human and automated review, often resulted in enforcement errors that disproportionately affected Muslims and relevant linguistic communities (see public comments, e.g., PC-11183 ECNL EFF, PC-11190 – Brennan Centre; PC-11188 – Digital Rights Foundation; PC-11196 – Integrity Institute, see also the BSR report on Meta's Impact in Israel and Palestine). Meta also acknowledged in its request that this is a consequence of its policy on the word "shaheed." To improve the transparency, accuracy and fairness of Meta's human and automated enforcement processes, the Board makes the two recommendations above.

72. Meta explained that it measures human and automated enforcement accuracy on a regular basis. Human reviewers are subject to periodic audits by Meta's outsourcing partners, and those audits are subject to additional evaluation by Meta's Global Operations team.

73. The Board previously recommended that Meta improve its transparency reporting to increase public information on error rates by making this information viewable by country and language for "praise" and "support" of dangerous organizations and individuals (Öcalan's Isolation, recommendation no. 12) and for each Community Standard (Punjabi Concern Over the RSS in India, recommendation no. 3). Meta explained that, after a feasibility assessment, it has declined to implement this recommendation (Öcalan's Isolation, recommendation no. 12; Meta Q4 2021 Quarterly Update on the Oversight Board, at page 21), and has shifted focus towards long-term efforts to define accuracy metrics based on profile, page and account restrictions, as well as location rather than language (Punjabi Concern Over the RSS in India, recommendation no. 3; Meta Q2 2023 Quarterly Update on the Oversight Board, at page 59). Mindful of widespread concerns over disproportionate enforcement of the Dangerous Organizations and Individuals policy, which Meta's position on "shaheed" illustrates, and the deficiency in Meta's disclosure of

useful enforcement data, the Board reiterates these recommendations and the importance of the company improving transparency of enforcement by location.

74. The data Meta currently shares in its Transparency Center does not provide sufficient insights on the accuracy of human review and the performance of automated systems. To allow scrutiny of its methods, Meta must explain how it assesses the accuracy of human review and the performance of automated systems in the enforcement of its Dangerous Organizations and Individuals policy. To improve transparency on the performance of classifiers, Meta should share metrics that include precision (indicating the proportion of content correctly identified as violating out of all content flagged as violating) and recall (indicating the percentage of violating content the classifier identifies as violating out of all violating content that exists). Publishing information on the accuracy of human review across different regions, and the performance of classifiers in different languages, can contribute to accountability on Meta's human rights obligations. The Board therefore reiterates its previous recommendations and urges Meta to implement the proposed changes.

75. Meta relies on automated systems, including classifiers, to enforce its Community Standards. Classifier models automatically categorize data into one or more "class" (e.g., violating or not violating) based on a set of training data. Understanding how these systems function and how accurate they are is essential for evaluating how Meta moderates content at-scale. Meta provides some basic information on how it uses automated systems to detect violating content in the enforcement section of its Transparency Center, while in the features section Meta explains how it ranks and curates content. Meta already publishes information in its Community Standards Enforcement Reports on the percentage of content it removes before it is reported (i.e., on the basis of automated detection). This data indicates the importance of automated detection but falls short of providing detailed metrics about the use and accuracy of Meta's automated systems, especially on automated removals. This set of recommendations aims to strengthen transparency around Meta's use of algorithms.

76. In response to questions asked by the Board, Meta explained that the Dangerous Organizations and Individuals classifier for "General Language – Arabic" conducts an initial review for enforcement based on Meta's policies. Classifiers aim to detect content that may violate those policies. The Board finds that Meta should explain this reliance on classifiers in detail in the enforcement section of the Transparency Center to its users, consistent with its prior decisions (see Breast Cancer Symptoms and Nudity, recommendation no. 3; Colombia Police Cartoon, recommendation no. 3). In particular, it is important there is transparency on approximate

confidence thresholds for actioning or not actioning content, and for enqueuing (add to a queue) it for human review, as well as considerations or factors that determine those scores.

77. Meta explained that its accuracy testing does not focus on individual violation types within the policy, but instead focuses on overall accuracy rates across an entire policy area. Meta shared that it "conducts regular metric monitoring and audit accuracy rates at scale." The Board previously explained that it is not sufficient to evaluate the performance of Meta's enforcement of Community Standards as a whole (see Wampum Belt decision). Systems that perform well on average could potentially perform quite poorly on subcategories of content, such as content using the term "shaheed," on which incorrect decisions have a particularly pronounced human rights impact. It is therefore essential for Meta to demonstrate that it undertakes due diligence to identify and minimize the potential negative effects of its systems on human rights, and that it shares information allowing scrutiny of these effects.

78. Many stakeholders argued that Meta's algorithmic enforcement systems failed to account for context, were more inaccurate in non-English languages and that this had resulted in disparate impact against Muslim communities (see public comments, e.g., PC-11190 – Brennan Centre).

79. Some stakeholders recommended that Meta stop using automation to moderate the term "shaheed" entirely, as automation has generally failed to account for context to date (see public comments, e.g., PC-11164 – SMEX, p.4; PC-11157 – Palestine Institute for Public Diplomacy). Others argued that, considering the severe penalties for content violating the Dangerous Organizations and Individuals policy, automation should only be used to queue content for review but not to automatically remove content (see e.g., PC-11183 ECNL EFF, p. 5). Stakeholders suggested that any post that uses "shaheed" to refer to a designated person should "ideally be reviewed by human moderators who are familiar with the local context from which the post is originating," (see PC-11188 – Digital Rights Foundation, p.2). Some suggested that content should be geo-tagged, so that human and automated review could better account for context (see e.g., PC-11165 – Taraaz, p. 1). Stakeholders also called on Meta to provide more information on the accuracy of their automated enforcement and invest more money in improving their automated systems in non-English languages (see e.g., PC-11164 – SMEX, p.5; PC-11190 – Brennan Centre). They also pushed for Meta to report its content moderation activities "consistently across languages, including comprehensive data on user reports, action rate, types of action, efficacy of mitigation techniques, training information and appeal rates," (see e.g., PC-11196 – Integrity Institute).

80. Some experts disagreed with the use of "shaheed" in relation to a designated entity as a singular point of reference for automated systems to remove content from the platform, but they would encourage using the term as part of a more layered approach to moderation (e.g., as a signal for content to be enqueued for review against Meta's policies, but not removed for including this reference alone).

81. The Board acknowledges that relying on algorithmic systems is necessary when moderating content at-scale, and many of these proposed approaches would not be feasible at-scale and using automation. But the Board also concludes that Meta needs to take further steps, including the ones recommended above, to ensure the transparency and fairness of those systems.

***Procedural Note:**

The Oversight Board's policy advisory opinions are prepared by panels of five Members and approved by a majority of the Board. Board decisions do not necessarily represent the personal views of all Members.

For this policy advisory opinion, independent research was commissioned on behalf of the Board. The Board was assisted by an independent research institute headquartered at the University of Gothenburg, which draws on a team of more than 50 social scientists on six continents, as well as more than 3,200 country experts from around the world. The Board was also assisted by Duco Advisors, an advisory firm focusing on the intersection of geopolitics, trust and safety, and technology. Memetica, an organization that engages in open-source research on social-media trends, also provided analysis. 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.