Oversight Board

OVERTURNED

STATEMENTS ABOUT THE JAPANESE PRIME MINISTER

September 10, 2024

In the case of a user's reply to a Threads post about the Japanese Prime Minister and a tax fraud scandal, it was neither necessary nor consistent with Meta's human rights responsibilities for the content to be removed.

TYPE OF DECISION

Standard

POLICIES AND TOPICS

Elections, Freedom of expression

Violence and incitement

REGION/COUNTRIES

Japan

PLATFORM

Threads

ATTACHMENTS

Japanese Translation
Statements About the Japanese Prime Minister Decision PDF

To read the decision in Japanese, click **here**.

決定内容の全文を日本語で読むには、 **こちら**をクリックしてください。

SUMMARY

In the case of a user's reply to a Threads post about the Japanese Prime Minister and a tax fraud scandal, it was neither necessary nor consistent with Meta's human rights responsibilities for the content to be removed. This case grapples with the issue of how Meta should distinguish between figurative and actual threats of violence. The Board has repeatedly highlighted over-enforcement against figurative threats. It is concerning that Meta's Violence and Incitement policy still does not clearly distinguish literal from figurative threats. In this case, the threat against a political leader was intended as non-literal political criticism calling attention to alleged corruption, using strong language, which is not unusual on Japanese social media. It was unlikely to cause harm. Even though the two moderators involved spoke Japanese and understood the local sociopolitical context, they still removed the content in error. Therefore, Meta should provide additional guidance to its reviewers on how to evaluate language and local context, and ensure its internal guidelines are consistent with the policy rationale.

About the Case

In January 2024, a Threads post was shared that shows a news article about the Japanese Prime Minister Fumio Kishida and his response to fundraising irregularities involving his party. The post's

caption criticizes the Prime Minister for tax evasion. A user replied publicly to that post, calling for an explanation to be given to Japan's legislative body followed by the word "hah," and referring to the Prime Minister as a tax evader by using the phrase "死ね," which translates as "drop dead/die" in English. The phrase is included in several hashtags and the user's reply also includes derogatory language about a person who wears glasses.

The user's reply to the Threads post did not receive any likes and was reported once under Meta's Bullying and Harassment rules. Three weeks later, a human reviewer determined the content broke the Violence and Incitement rules instead. When the user appealed, another human reviewer decided once more that the content was violating. The user then appealed to the Board. After the Board selected the case, Meta decided its original decision was wrong and restored the user's reply to Threads.

Around the time of the original Threads post and the user's reply, Japanese politicians from the Liberal Democratic Party had been charged with underreporting fundraising incomes, although this did not include Prime Minister Kishida. Since 2022, when former Prime Minister Shinzo Abe was assassinated, there has been some concern about political violence in Japan.

Fumio Kishida recently announced he will not seek re-election as leader of Japan's Liberal Democratic Party on September 27, 2024 and is to step down as Prime Minister.

Key Findings

The Board finds that the phrase "drop dead/die" (translated from the original "死ね") was not a credible threat and did not break the Violence and Incitement rule that prohibits "threats of violence that could lead to death." Experts confirmed the phrase is broadly used in a figurative sense as a statement of dislike and disapproval. The content also points to this figurative use, with inclusion of the word "hah" expressing amusement or irony.

However, Meta's Violence and Incitement rule that prohibits calls for death using the phrase "death to" against high-risk persons is not clear enough. Meta's policy rationale suggests that context matters when evaluating threats but, as has been noted by the Board in a previous case, Meta's at-scale human reviewers are not empowered to assess the intent or credibility of a threat, so if a post includes threatening statements like "death to" and a target (i.e., "a call for violence against a target"), it is removed. Repeating a **2022 recommendation**, the Board calls on Meta to include an explanation in the policy's public language that rhetorical threats using the phrase "death to" are generally allowed, except when directed at high-risk individuals, and to provide criteria on when threatening statements directed at heads of state are permitted to protect rhetorical political speech.

It is also confusing how this policy differs in its treatment of “public figures” and “high-risk persons.” Currently, medium severity violence threats against public figures are only removed when “credible,” compared with content removal “regardless of credibility” for other individuals. More confusingly still, there is another line in this policy that gives “additional protections” to high-risk persons. Internal guidance on this to reviewers, which is not available publicly, specifically indicates that “death to” content against such high-risk people should be removed. When asked by the Board, Meta said its policy offers greater protection for users’ speech involving medium severity threats at public figures because people often use hyperbolic language to express their disdain, without intending any violence. However, threats of high-severity violence, including death calls against high-risk persons, carry a greater risk of potential offline harm. In this case, Meta identified the Japanese Prime Minister as falling into both categories. The Board has real concerns about the policy’s definitions of “public figures” and “high-risk persons” not being clear enough to users, especially when the two categories interact.

In response to the Board’s previous recommendations, Meta has completed some policy work to strike a better balance between violent speech and political expression, but it has not yet publicly clarified who “high-risk persons” are. The Board believes providing a general definition with illustrative examples in the Community Standards would allow users to understand that this protection is based on the person’s occupation, political activity or public service. The Board offered such a list in the 2022 **Iran Protest Slogan** case.

The Oversight Board’s Decision

The Board overturns Meta’s original decision to take down the content.

The Board recommends that Meta:

- Update the Violence and Incitement policy to provide a general definition for “high-risk persons” clarifying that they encompass people such as political leaders, who may be at higher risk of assassination or other violence – and provide illustrative examples.
- Update internal guidelines for at-scale reviewers on calls for death using the phrase “death to” directed against “high-risk persons,” specifically to allow posts expressing disdain or disagreement through non-serious and casual ways of threatening violence. Local context and language should be taken into consideration.
- Hyperlink to its Bullying and Harassment definition of public figures in the Violence and Incitement policy, and other relevant Community Standards, where such figures are referenced.

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

FULL CASE DECISION

1. Case Description and Background

In January 2024, a user replied publicly to a Threads post containing a screenshot of a news article. The article included a statement by Prime Minister Fumio Kishida about unreported fundraising revenues involving members of his Liberal Democratic Party. In the statement, Kishida said the amount "remained intact and was not a slush fund." The main Threads post included an image of the Prime Minister and a caption criticizing him for tax evasion. The user's response to the post calls for an explanation to be given to Japan's legislative body and includes the interjection "hah." It also includes several hashtags using the phrase "死ね" (transliterated as "shi-ne" and translated as "drop dead/die") to refer to the Prime Minister as a tax evader as well as derogatory language for a person who wears glasses, such as #dietaxevasionglasses and #diefilthshitglasses (translated from Japanese). All the content is in Japanese.

Both the post and reply were made around the time of the Prime Minister's parliamentary statement addressing his party's alleged underreporting of this revenue. Fumio Kishida, who has served as Japan's Prime Minister since October 2021, recently **announced** he will not seek re-election in the Liberal Democratic Party's leadership election, to be held on September 27, 2024.

The user's reply did not receive any likes or responses. It was reported once under the **Bullying and Harassment** policy for "calls for death" towards a public figure. Due to a backlog, a human moderator reviewed the content approximately three weeks later, determining that it violated Meta's **Violence and Incitement** policy and removing it from Threads. The user then appealed to Meta. A second human reviewer also found that the content violated the Violence and Incitement policy. Finally, the user appealed to the Board. After the Board selected this case, Meta determined that its original decision to remove the content was an error and restored it on Threads.

The Oversight Board considered the following context in coming to its decision.

When the reply to the Threads post was disseminated in January 2024, prosecutors had recently indicted Japanese politicians belonging to the Liberal Democratic Party for **underreporting** fundraising incomes. Prime Minister Kishida himself was not indicted.

Research commissioned by the Board identified a general sentiment of disapproval and criticism on Threads towards the Prime Minister in relation to the tax fraud allegations, with other posts containing the phrase "死ね" (drop dead/die). Experts consulted by the Board noted that people in Japan use social media frequently to post political criticism. In the past, online message boards

have served as anonymous platforms to express social discontent without fear of consequences (see also public comments, PC-29594 and PC-29589).

According to experts consulted by the Board, in recent decades, political violence in Japan has been rare. For this reason, the nation was shocked in 2022 when Prime Minister Shinzo Abe was **assassinated** while campaigning. Concerns about political violence rose in April 2023 when a man **used a pipe bomb** during a campaign speech by Prime Minister Kishida, wounding two bystanders but not harming the Prime Minister.

According to linguistic experts consulted by the Board, the phrases used in the post are offensive and widely used to convey severe disapproval or frustration. While the phrase "死ね" (drop dead/die) may in some instances be used literally as a threat, it is generally used figuratively to express anger without being a genuine threat (see also public comment by Ayako Hatano, PC-29588).

In 2017, the UN Special Rapporteur on Freedom of Expression **voiced** concerns about freedom of expression in Japan. These concerns related to the use of direct and indirect pressure by government officials on media, the limited capacity to debate historical events and the increased restrictions on information access based on assertions of national security.

In its 2024 Global Expression **Report**, Article 19 placed Japan at 30 out of 161 countries. Freedom House **classified** Japan as "Free" in its 2023 Freedom on the Net evaluation, but raised concerns about government intervention in the online media ecosystem, the lack of independent regulatory bodies and the lack of clear definitions in the recent legislative **amendments** regulating online insults. However, the organization's Freedom in the World **report** gave the country 96 out of 100 points for political and civil liberties. Japan also consistently receives high ratings in democracy and rule of law indices. In 2023, the World Justice Project's Rule of Law Index **ranked** Japan 14 of 142 countries.

2. User Submissions

In their statement to the Board, the user who posted the reply claimed they were merely criticizing the Liberal Democratic Party government for its alleged acts of condoning and abetting tax evasion. They said that Meta's removal of their post contributed to the obstruction of freedom of speech in Japan by prohibiting criticism of a public figure.

3. Meta's Content Policies and Submissions

I. Meta's Content Policies

The Board's analysis was informed by Meta's commitment to voice, which the company describes as "paramount," and its value of safety. Meta assessed the content under its **Violence and Incitement** and **Bullying and Harassment** policies and initially removed it under the Violence and Incitement Policy. After the Board identified this case for review, the company determined the content did not violate either policy.

Violence and Incitement Community Standard

The policy rationale for the **Violence and Incitement Community Standard** explains that Meta intends to "prevent potential offline harm that may be related to content on [its] platforms" while acknowledging that "people commonly express disdain or disagreement by threatening or calling for violence in non-serious and casual ways." It acknowledges: "Context matters, so [Meta] consider[s] various factors such as condemnation or awareness raising of violent threats, … or the public visibility and vulnerability of the target of the threats."

This policy provides for universal protection for everyone against "threats of violence that could lead to death (or other forms of high-severity violence)." Threats include "statements or visuals representing an intention, aspiration or call for violence against a target."

Before April 2024, the policy prohibited "threats that could lead to serious injury (mid-severity violence) and admission of past violence" towards certain people and groups, including high-risk persons. In April 2024, Meta updated this policy to provide universal protection against such threats for everyone regardless of credibility, except for threats "against public figures," which the policy requires to be "credible." The only mention of "high-risk persons" left in the current version of the policy relates to low-severity threats that still allows for "[a]dditional protections for Private Adults, All Children, high-risk persons and persons or groups based on their protected characteristics ..."

The public-facing language of the policy does not define the term "high-risk persons." However, Meta's internal guidelines to reviewers contain a list of high-risk persons that includes heads of state; former heads of state; candidates and former candidates for head of state; candidates in national and supranational elections for up to 30 days after election if not elected; people with a history of assassination attempts; activists and journalists (see **Iran Protest Slogan** decision).

Bullying and Harassment Community Standard

Meta's **Bullying and Harassment Community Standard** prohibits various forms of abuse directed against individuals, including "making threats" and "distinguishes between public figures and private individuals" to "allow discussion, which often includes critical commentary of people who are

featured in the news or who have a large public audience." This policy prohibits "severe" attacks on public figures, as well as certain attacks where the public figure is "purposefully exposed," defined as "directly tagg[ing] [a public figure] in the post or comment."

The policy defines a "public figure" to include "state and national level government officials, political candidates for those offices, people with over one million fans or followers on social media and people who receive substantial news coverage."

II. Meta's Submissions

Meta informed the Board that the term "死ね" (drop dead/die) in the hashtags did not violate its policies in this case. Meta regards this use as a political statement that contains figurative speech, rather than a credible call for death. The company explained that it often cannot distinguish at-scale between statements containing credible death threats and figurative language intended to make a political point, which is why it initially removed the content.

Meta told the Board that Prime Minister Kishida is considered a public figure under the company's Violence and Incitement and Bullying and Harassment policies, while the user replying to the post was not considered a public figure. Meta also informed the Board that Prime Minister Kishida is considered a "high-risk person" under the Violence and Incitement Community Standard.

Violence and Incitement Community Standard

Under its **Violence and Incitement** policy, Meta prohibits: "Threats of violence that could lead to death (or other forms of high-severity violence)." In its non-public guidelines to human reviewers, Meta notes that it removes calls for death of a high-risk person if those calls use the words "death to." Meta told the Board that the concept of a high-risk person is limited to this policy and includes political leaders, who may be at higher risk of assassination or other violence.

Meta acknowledged that it is challenging to maintain a distinction between the phrases "death to" and "die" in every case, particularly when the meaning of the phrases may overlap in the original language. In this case, the content uses the phrases "die" and not "death to" in the hashtags, #dietaxevasionglasses and #diefilthshitglasses (translated from Japanese). In addition, Meta noted that even if it treated "die" and "death to" similarly (as a call for death), the company would not remove this content on escalation because it would be a non-literal threat that does not violate the spirit of the policy. The spirit of the policy allowance permits content when a strict interpretation of a policy produces an outcome that is at odds with that policy's intent (see **Sri Lanka Pharmaceuticals** decision). Meta deemed the threat to be non-literal because the other words of

the hashtags and of the reply itself are about political accountability through hearings before Japan's legislature. As such, the call to have a political leader be held to account before a legislative body indicated that the death threat was figurative rather than literal. For these reasons, Meta determined that the content does not violate the Violence and Incitement policy.

Bullying and Harassment Community Standard

Meta informed the Board that the content did not violate its **Bullying and Harassment** policy because the content did not "purposefully expose" a public figure. The user did not tag or reply to a comment by Prime Minister Kishida and did not post the content on the Prime Minister's page. Meta therefore determined the content did not purposefully expose Prime Minister Kishida and would not violate the Bullying and Harassment policy even if the threat was literal.

The Board asked Meta 19 questions in writing. Questions related to Meta's enforcement practices and resources in Japan, the training provided for at-scale human reviewers and how it incorporates local context, the process for escalating at-scale policy lines, the feasibility to enforce the policy prohibiting death threats against high-risk persons only on escalation, Meta's review backlog on Threads and automated detection capacities. Meta answered 17 questions in full and two questions in part. The company partially answered the questions related to the review backlog and governmental requests to take down content in Japan.

4. Public Comments

The Oversight Board received 20 public comments that met **the terms for submission**: 13 from Asia Pacific and Oceania, three from the United States and Canada, three from Europe and one from Central and South Asia. To read public comments submitted with consent to publish, click **here**.

The submissions covered the following themes: the sociopolitical context in Japan; online threats of violence against politicians and limitations on freedom of expression; the use of rhetorical threats or calls for violence in Japanese political discourse; the linguistic context of the phrase "drop dead/die"; and Meta's choice not to recommend political content on Threads for pages not followed by users.

5. Oversight Board Analysis

The Board examined whether this content should be removed by analyzing Meta's content policies, human rights responsibilities and values. The Board also assessed the implications of this case for Meta's broader approach to content governance.

5.1 Compliance with Meta's Content Policies

I. Content Rules

Violence and Incitement Community Standard

The Board finds that the content in this case does not violate the **Violence and Incitement** policy prohibiting "threats of violence that could lead to death (or other forms of high-severity violence)." The phrase "死ね" (drop dead/die) was used in a non-literal way and was not a credible threat.

Linguistic experts consulted by the Board explained that although sometimes this phrase can be used to threaten someone's life literally, it is broadly used in a figurative sense as a statement of dislike and disapproval. The experts found that the use of the term in this content fell into the figurative category. Data experts, who examined the incidence of the phrase on Threads and other platforms, concluded that the term is commonly used figuratively or ironically. This includes examples of users reporting they are "dying" of pain or wishing other users to "die" because of a comment that those users made.

The reply itself also suggests that the phrase was meant figuratively. The user's reply to the Threads post called for the head of the National Tax Agency to appear before the national legislative body and explain the fraud allegations. The reply also included the interjection "hah." In the Board's view, the word "hah," which usually expresses amusement or irony, suggests a non-literal meaning of the term "死ね" (drop dead/die). Similarly, the Board agrees with Meta's assessment that the user's proposed remedy – that Kishida be held to account by the country's legislative body – suggests the content was political criticism, rather than a literal call for death.

The Board acknowledges that recent events in Japan would lead to sensitivities about any call for the death of a political leader. The assassination of Prime Minister Abe in 2022 and the use of a pipe bomb near Prime Minister Kishida in 2023 underscore the critical importance of taking credible death threats seriously. In this case, however, the call for death was simply not credible.

Bullying and Harassment Community Standard

The Board finds that the content in this case does not violate the Bullying and Harassment policy. The Board agrees with Meta that while Prime Minister Kishida meets the policy criteria for public figures, he was not "purposefully exposed" by the content. The user did not post the reply directly to Prime Minister Kishida's page and did not tag him, thus the content did not directly address the Prime Minister.

5.2 Compliance with Meta's Human Rights Responsibilities

The Board finds that removing the content from the platform was not consistent with Meta's human rights responsibilities.

Freedom of Expression (Article 19 ICCPR)

Article 19 of the **International Covenant on Civil and Political Rights** (ICCPR) provides "particularly high" protection for "public debate concerning public figures in the political domain and public institutions," (**General Comment No. 34**, para. 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 human rights responsibilities in line with the **UN Guiding Principles on Business and Human Rights**, which Meta itself has committed to in its **Corporate Human Rights Policy**. The Board does this both in relation to the individual content decision under review and what this says about Meta's broader approach to content governance. As the UN Special Rapporteur on freedom of expression has stated, although "companies do not have the obligations of governments, their impact is of a sort that requires them to assess the same kind of questions about protecting their users' right to freedom of expression," (**A/74/486**, para. 41). The Board has recognized the importance of political speech against a head of state, even when it is offensive, as such leaders are legitimately subject to criticism and political opposition (see **Iran Protest Slogan** and **Colombia Protests** decisions; General Comment No. 34, at paras 11 and 38).

I. Legality (Clarity and Accessibility of the Rules)

The principle of legality requires rules limiting expression to be accessible and clear, formulated with sufficient precision to enable an individual to regulate their conduct accordingly (General Comment No. 34, para. 25). Additionally, these rules "may not confer unfettered discretion for the restriction of freedom of expression on those charged with [their] execution" and must "provide sufficient guidance to those charged with their execution to enable them to ascertain what sorts of expression are properly restricted and what sorts are not," (Ibid.). The UN Special Rapporteur on freedom of expression has stated that when applied to private actors' governance of online speech, rules should be clear and specific (**A/HRC/38/35**, para. 46). People using Meta's platforms should be able to access and understand the rules, and content reviewers should have clear guidance regarding their enforcement.

The Board finds that the prohibition on calls for death using the phrase “death to” against high-risk persons is not sufficiently clear and accessible to users.

The policy rationale allows for context when evaluating the credibility of a threat, such as content posted for condemnation, awareness raising and non-serious or casual threats. However, the policy rationale does not specify how non-literal statements are to be distinguished from credible threats. As noted by the Board in the **Iran Protest Slogan** case, at-scale human reviewers follow specific guidelines based on signals or criteria such as calls to death against a target. They are not empowered to assess the intent or the credibility of a threat, so if a post includes threatening statements like “death to” or “drop dead” (as in this case) and a target, it is removed. The Board therefore reiterates recommendation no. 1 from the **Iran Protest Slogan** case, which states that Meta should include an explanation in the public-facing language of the Violence and Incitement policy that rhetorical threats using the phrase “death to” are generally allowed, except when directed at high-risk individuals, and provide criteria for when threatening statements directed at heads of state are permitted to protect clearly rhetorical political speech.

The policy is also not sufficiently clear about its treatment of “public figures” and “high-risk persons.” The policy currently gives less protection to public figures, noting that threats of medium severity violence towards public figures are removed only when “credible,” while such threats are removed “regardless of credibility” for other figures. In contrast, the policy gives more protection to high-risk persons, through a policy line citing “additional protections” for such groups. As noted above, internal guidance also offers more protection to high-risk persons, calling for removal of “death to” when directed at high-risk persons. In response to the Board’s question, Meta explained that the policy offers greater protection for speech containing medium severity threats directed at public figures because people frequently express disdain or disagreement with adult public figures using hyperbolic language but often do not intend to incite violence. In contrast, threats of high severity violence carry a greater risk of potential offline harm, including death calls against high-risk persons. In this case, Meta deemed Prime Minister Kishida to fall into both categories. The Board is concerned that the Violence and Incitement policy definitions of “public figures” and “high-risk persons” do not provide sufficient clarity for users to understand either category, much less what happens when the two categories interact.

In the **Iran Protest Slogan** case, the Board recommended that Meta amend the Violence and Incitement Community Standard to include an illustrative list of high-risk persons, explaining the category may include heads of state. Since the publication of that decision, Meta initiated a policy development **process** to strike a better balance between violent speech and political expression. Nevertheless, the company has not yet publicly clarified who is a high-risk person. During its work

on this case, the Board held a briefing session with Meta where the company explained that publishing its internal definition of high-risk persons could lead some users to circumvent existing policies and enforcement guidelines.

The Board acknowledges Meta's concern that publishing detailed guidelines could allow certain users to evade established enforcement rules. However, the Board believes that Meta should not take an all-or-nothing approach. Instead, Meta should publish a general definition of high-risk persons and an illustrative list of examples. Such an approach would allow users to understand that the protection of these persons is based on their occupation, political activity, public service or other risk-related activity. The Board believes that such an approach would not impede enforcement efficiency. Indeed, the Board has already offered such a list with Meta's agreement in the Iran Protest Slogan case, noting: "In addition to heads of state, other examples of high-risk persons include: former heads of state; candidates and former candidates for head of state; candidates in national and supranational elections for up to 30 days after election if not elected; people with a history of assassination attempts; activists and journalists." Given that these examples are already in the public domain, they should be reflected in the Community Standard itself.

Building on the Board's findings in the Iran Protest Slogan case and the updates that Meta has already implemented to the Violence and Incitement policy, the Board recommends that Meta provide a general definition for high-risk persons clarifying that high-risk persons encompass people, like political leaders, who may be at higher risk of assassination or other violence and provide illustrative examples, such as those discussed in the Iran Protest Slogan case.

II. Legitimate Aim

Any restriction on freedom of expression should also pursue one or more of the legitimate aims listed in the ICCPR. The Violence and Incitement Community Standard aims to "prevent potential offline harm" by removing content that poses "a genuine risk of physical harm or direct threats to public safety." This policy serves the legitimate aim of protecting the right to life and the right to security of person (Article 6, ICCPR; Article 9 ICCPR).

III. Necessity and Proportionality

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

The Board finds that Meta's original decision to remove the content under its **Violence and Incitement** policy was not necessary, as it was not the least intrusive measure to protect the safety of Prime Minister Kishida. This analysis is the crux of this case, as it again grapples with the challenging issue of how Meta should distinguish between rhetorical and actual threats. The Board has repeatedly expressed concern about over-enforcement against figurative threats in the **Iran Protest Slogan**, **Iranian Woman Confronted on Street** and **Reporting on Pakistani Parliament Speech** cases. These cases might be distinguished from the case in question because they concerned a slogan, a coordinated protest movement, or an impending election. Yet the core issue is the same, the restriction of political speech due to a non-credible threat of violence. The Board believes that Meta should enable such discussions and ensure that users can express their political views, including dislike or disapproval of politicians' actions and behavior, without creating unnecessary barriers.

However, the Board is concerned that Meta's Violence and Incitement policy still does not clearly distinguish literal and figurative threats. This problem is further emphasized by the fact that the content in this case was marked as violating by two human moderators who, according to Meta, spoke Japanese and were familiar with the local sociopolitical context.

The six factors described in the **Rabat Plan of Action** (context, speaker, intent, content of the speech, extent of the speech and likelihood of imminent harm) provide valuable guidance in assessing the credibility of the threats. Although the Rabat framework was created to assess incitement to national, racial or religious hatred, the six-factor test is useful for evaluating incitement to violence generally (see, for example, **Iran Protest Slogan** and **Call for Women's Protest in Cuba** decisions). Given Meta's original assumption that removing the content was necessary to protect the safety of Prime Minister Kishida, the Board used the six factors to assess the credibility of the alleged threat in this case.

The content was posted during the 2023 tax fraud scandal involving Prime Minister Kishida's party. Experts consulted by the Board explained that while political criticism online has increased in Japan, there is no clear link between online threats and recent violence against Japanese politicians. The user had fewer than 1, 000 followers and was not a public figure, while the content received no views or likes, reflecting the low interest in the reply. The user's intent appeared to be political criticism by calling attention to political corruption, using strong language, which is not unusual on Japanese social media (see public comments PC-29589 and PC-29594), and was unlikely to cause imminent harm.

The Board acknowledges that assessing the credibility of threats of violence is a context-specific, very difficult exercise, especially when enforcing against content on a global scale. The Board also understands that Meta can conduct a more accurate evaluation of credibility of threats on-escalation. The Board considered recommending that Meta enforce the policy prohibiting threats using the phrase “death to” against high-risk persons on-escalation only. Escalation-only policies require additional context to enforce against content, with decisions made by subject matter experts as opposed to the at-scale human moderators who initially review the content. The Board understands that the number of Meta’s subject matter experts is significantly lower than the number of at-scale human reviewers, thus the former’s capacity is limited. As such, escalation-only enforcement of this policy may lead to a significant amount of content not being reviewed due to lower expert capacity. Moreover, escalation-only rules can only be enforced if brought to Meta’s attention by some other means, for example, Trusted Partners or content with significant press coverage (see **Sudan’s Rapid Support Forces Video Captive** decision). This means that Meta will be able to review threats of death using the phrase “death to” only when flagged through certain channels.

The Board ultimately determined that this would likely result in under-enforcement and more death threats remaining on Meta’s platforms. Moreover, because Meta could not provide validated data about the prevalence of such content on its platforms, the Board could not assess the magnitude of such under-enforcement.

The Board therefore is of the view that to effectively protect political speech, Meta should provide additional guidance to its reviewers to evaluate language and local context, ensuring the guidelines it issues for moderators are consistent with the underlying policy rationale. In its previous cases on similar issues (see Iran Protest Slogan, Iranian Woman Confronted on Street and Reporting on Pakistani Parliament Speech), the Board explored policy and enforcement solutions, often time-sensitive and narrowly tailored to the specific context, including elections, crises and conflicts. This has allowed Meta to adjust its enforcement practices and account for specific context by using mechanisms such as the Crisis Policy Protocol (CPP) and Integrity Product Operations Center (IPOC).

In this case, Meta informed the Board that it did not establish any special enforcement measures. Meta stated that a single incident such as the assassination of former Prime Minister Shinzo Abe, while tragic, is generally not sufficient to trigger such mechanisms, unless there are additional signals of wider risk or instability. Instead, Meta designated the assassination under its “Violating Violent Event” protocol, limited to content related to that instance of violence only. In these circumstances Meta can only rely on its general policy and enforcement practices. Therefore,

developing a scalable solution to distinguish credible from figurative threats is the only way to effectively protect political expression.

Moreover, if Meta chooses to continue enforcing this policy at-scale, the accuracy of its automated systems will continue to be impacted by the quality of training data provided by human moderators. The Board reiterates its findings from the **Iranian Woman Confronted on Street** decision, that when human moderators remove figurative statements based on the rigid enforcement of a rule, that mistake is likely to be reproduced and amplified through automation, leading to over-enforcement.

Based on the Board's findings that calls for death require a context-driven assessment of the probability that a threat will result in real-world harm, this could require more nuanced enforcement guidelines to at-scale human reviewers than those currently available. Meta's internal guidelines instruct reviewers to remove calls for death using the specific phrase "death to" when directed against high-risk individuals. These guidelines do not reflect the **Violence and Incitement** policy rationale, which states that "context matters" and that it accounts for non-serious and casual ways of threatening or calling for violence to express disdain or disagreement. The Board, therefore, finds that Meta should update its internal guidelines and specific instructions to reviewers to explicitly allow for consideration of local context and language, and to account for "non-serious and casual ways" of threatening or calling for violence to express such disdain or disagreement.

Finally, the Board is also concerned about Meta's ability to handle context-sensitive content on Threads. Meta informed the Board that the review of the content in this case was delayed for about three weeks due to a backlog. Meta explained that at the time of enforcement, Threads content moderation relied exclusively on human reviewers for Threads reports, whereas the company typically uses multiple techniques to prevent backlogs from accumulating, such as automatic closure of reports. Automatic closing of reports after 48 hours means that, unless there are any mechanisms to keep them open, the reports will be closed without review, leaving users without an effective remedy.

6. The Oversight Board's Decision

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

7. Recommendations

Content Policy

1. Meta should update the Violence and Incitement policy to provide a general definition for “high-risk persons” clarifying that high-risk persons encompass people, like political leaders, who may be at higher risk of assassination or other violence and provide illustrative examples.

The Board will consider this recommendation implemented when the public-facing language of the Violence and Incitement policy reflects the proposed change.

Enforcement

2. Meta should update its internal guidelines to at-scale reviewers about calls for death using the specific phrase “death to” when directed against high-risk persons. This update should allow posts that, in the local context and language, express disdain or disagreement through non-serious and casual ways of threatening violence.

The Board will consider this recommendation implemented when Meta shares relevant data on the reduction of false positive identification of content containing calls for death using the specific phrase “death to” when directed against high-risk persons.

Content Policy

3. Meta should hyperlink to its Bullying and Harassment definition of public figures in the Violence and Incitement policy, and in any other Community Standards where public figures are referenced, to allow users to distinguish it from high-risk persons.

The Board will consider this recommendation implemented when the public-facing language of the Violence and Incitement policy, and of Meta’s Community Standards more broadly, reflects the proposed change.

***Procedural Note:**

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

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

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

****Translation Note:**

The translation process for the announcement of the case concerning Statements About the Japanese Prime Minister led to the use of a phrase in the Japanese version of the announcement that had a similar meaning, but differed from the original phrase used. Instead of the original phrase “死ね” (shi-ne), the announcement used the term “くたばれ” (kutabare) as the translation of “drop dead.”

We understand that this may have caused confusion. Please be assured that the Board’s deliberation and decision were based on the original wording of “死ね” (shi-ne) and that we are committed to ensuring accuracy in our translation processes.

Return to Case Decisions and Policy Advisory Opinions

STRATEGIC PRIORITIES

TRACK YOUR APPEAL

SUBSCRIBE

GET IN TOUCH

CAREERS

PRESS REQUESTS

RESOURCES

FAQS

PRIVACY NOTICE

TERMS

COOKIES