日本法曹有資格者・米国ニューヨーク州弁護士
杉山日那子

2024年5月30日

フェイスブック監督委員会御中

私は、カリフォルニア大学アーバイン校ロースクール国際司法クリニックに勤務する人権弁護士です。日本で育ち、第一言語は日本語です。**2024-027-TH-UA**に関する私の意見をお伝えします。本意見は同クリニックの見解を代表するものではありません。

本件では、自民党岸田派の所属議員の関与する政治資金収入の不記載についての岸田首相の声明を含んだスレッズの投稿に対する、岸田首相に対する「くたばれ」というハッシュタグを付した返信（以下「本投稿」）の削除の是非が問題となっています。この意見書では、第一に、本投稿の削除判断が明らかに誤りであったこと、第二に、日本で政治家を批判する表現が手厚く保護されていること、第三に、メタが取るべき再発防止策を説明します。なお、以前のコメントでもお伝えしましたが、私は、フェイスブック監督委員会が、メタによるコンテンツモデレーションのポリシーやその実施の評価にあたり適切なアプローチ―国際人権法に則りながら、問題となる表現やそのモデレーションから影響を受ける人やコミュニティの意見を考慮する参加的なアプローチ―を採用していることに感謝しています。このようなアプローチは、メタが遵守を表明する国連ビジネスと人権に関する指導原則にも合致すると考えます。

第1　本投稿の削除判断の明らかな誤り

本件では、「いじめや嫌がらせに関するコミュニティ規定」に反するというユーザーからの報告の後、メタの審査担当者が「くたばれ」という表現が死を求めるものとして「暴力と扇動のポリシー」に違反すると判断し（第1判断）、メタは本投稿を削除しました。投稿者からの不服申立てを受け、メタの別の審査担当者が改めて検討しましたが、違反を認定して削除相当と判断しました（第2判断）。その後、監督委員会が本件の審査を開始したことを受け、メタは本投稿の「くたばれ」という表現は同ポリシーに違反する危険な表現にはあたらないと判断（第3判断）して本投稿を復活しました。

改めての確認ですが、国際人権規約19条2項は表現の自由を広く手厚く保障し、同条3項は表現の自由の制約を①明確なルールに基づき、②人の権利や評判の尊重、国家の安全や公序良俗、公衆衛生や道徳の保護のいずれかのために③必要で相当である場合に限り許しています（いわゆる三部テスト）。メタは、ビジネスと人権に関する指導原則の下、ユーザーの表現の自由を侵害しないために、投稿の削除やアクセス制限の範囲を三部テストに合格するものに限定する責任を負っています。暴力を煽る投稿の拡散の防止は言うまでもなく正統な利益ですが、削除は当該利益のために必要で相当でなければならず、言い換えれば、削除が許されるのは、問題となる表現について暴力を引き起こす具体的な可能性が認められる場合に限られます（国連人

権委員会一般的意見34第35段落参照）。メタの「暴力と扇動のポリシー」は、個別のモデレーションをこの規範に適合させるために制定され運用されていると理解しています。

本投稿の「くたばれ」という表現は、メタの第3判断のとおり、「暴力と扇動のポリシー」の禁じる「Threats of violence that could lead to death (or other forms of high-sensitivity violence)」に該当しません。日本で歴史が長く一般的な辞典の一つである広辞苑によれば、「くたばれ」の原形「くたばる」の一義的意味は「体力が衰える」、「弱る」です。「『死ぬ』ことをののしっていう言葉」も意味しますが副次的とされています。言い換えれば、文脈に死や身体の損傷を連想させる要素がない限り基本的に「死」を意味せず、その命令形である「くたばれ」も同様、そのような要素がなければ、殺人または身体への加害意図の表明であるとはまず捉えられません。本投稿にそのような要素がないことは明らかですので「Threats of violence that could lead to death (or other forms of high-sensitivity violence)」に該当しません。この結論はヘイトスピーチのへの取組みを整理した国連ラバト行動計画のあてはめとしても明らかです。

つまり本件では「くたばれ」の意味が正しく理解されれば、本投稿が「暴力と扇動のポリシー」に反しないという結論を導くことは難しくなかったと思われます。第1判断と第2判断の審査担当者が判断を誤った原因は、彼らが「くたばれ」を「死ね」、「殺す」等の、文脈が修辞表現であることを示していない限り原則として相手への加害の意図を示すと考えられる言葉（また、監督委員会が2022年に審査したイランのプロテストのスローガンの事例で問題となった、一義的な意味は「death with」だが「down with」という意味で用いられることがある表現）と同じカテゴリの表現だと誤解したことにあるのではないかと推察されます。

本投稿が「Threats of violence that could lead to death (or other forms of high-sensitivity violence)」にあたるかどうかに直接関係しませんが、文脈を考えれば「くたばれ」は岸田首相の辞任を求めるという意味だと思われます。

意見聴取の対象の第二点目（暴力を煽る言葉を用いた修辞的な表現がどの程度一般的であるのか、信憑性ある害悪の告知とどう区別されるのか）について（「くたばれ」の一義的意味は暴力の呼びかけではないのでやや的外れに思いますが、その点はさて措くとして）、日本の刑法222条の規定する脅迫罪に関する裁判例には一定の参照価値があると思われます。

第2　日本における政治家に対する批判への手厚い保護

次に、監督委員会は政治家に対するインターネット上の脅迫、政治家に対する批判についての表現の自由の制約等の社会政治的背景についても意見を求めています。他国の人が日本に抱くイメージのように、協調性に一定の価値を置く文化が日本に存在することについて、私を含め日本の多くの人がおそらく同意すると思われます。しかし、私が強調したいのは、日本の憲法は表現の自由を明文で手厚く保護しており、表現規制も、個別に是正が望まれる点はありますが、全体的には表現の自由とのバランスを取る努力の上で制定・運用されています。そして何より表現の自由は市民社会により特に公共の利益に関わる事柄について積極的に行使されてきました。

中でも、公職にある者はより厳しい批判にさらされるべきだという法的規範は、明文規定や裁判例で具体化されています。例えば、刑法230条の定める名誉毀損罪には1947年に230条の1が付け加えられ、その3項においては「公務員又は公選による公務員の候補者に関する事実」については真実性の証明を条件に処罰対象から外しています。政治家に対する名誉毀損の民事事件でも、裁判所は、違法性の有無あるいは損害額の算定の場面で表現の自由に傾けた厳格な判断を下しています。またプライバシー侵害の有無について、最高裁は、小説の差し止めの可否について、プライバシーを侵害されたとする者が公的立場にあるかどうかを考慮し判断しました（平成14年9月24日 最高裁判所第三小法廷集民第207号243頁）。近時の判例でも①逮捕歴を含むツイートのツイッター社への削除命令の可否（令和4年6月24日最高裁判所第二小法廷民集第76巻5号1170頁）、②犯罪歴を含む検索結果のグーグル社への削除命令の可否（平成29年1月31日最高裁判所第三小法廷民集第71巻1号63頁）が問題となった事件においても、最高裁はプライバシーを侵害されたとする者の公的立場の有無を考慮し判断しています。

日本政府も政治家を批判する表現の要保護性を表明しています。近時の例として、2022年の刑法改正により侮辱罪の法定刑が加重されましたが、改正法の国会審議の際、政府は政治家への批判について、公正な論評など刑法35条の正当行為に該当する場合には違法性が阻却されることを明確にしています（法務省の説明Q9も参照）。

一部の国に残ってしまっているような、政治家に対する批判を特に罰するような法律は日本に存在しませんし、そのような規制を求める声は日本の市民社会には見られません。むしろ、日本の市民社会は、政治家が反対派の声を抑圧しようとする場面では、訴訟等を通じて是正するアクションを取り続けてきました。例えば、2019年の参議院選挙で、街頭演説中の当時の安倍首相に批判的な発言を投げかけた市民の複数が警察官に排除されましたが、市民たちは表現の自由侵害を主張して複数の訴訟を起こし、裁判所は排除の違法性を認めました。

なお、政治家に対するオンライン上の脅迫が一般的に増加しているという具体的な報道やレポートについて目にしたことがありません。

第3　メタが取るべき再発防止策

上記のとおり、本件は2022年のイランのプロテストのスローガンのケースと異なり、問題となった「くたばれ」という表現の意味が正しく理解されさえすれば、本投稿が「暴力と扇動のポリシー」に違反しないという判断を下すことは難しくなかったと考えられます。それにもかかわらず第1判断と第2判断の各審査担当者が判断を誤っていることを考えると、判断の誤りは偶発的というよりも、メタの日本語の投稿のモデレーションの体制が不十分である可能性、ひいては第1判断と第2判断のような質の悪い判断が他の投稿についても繰り返されている可能性を示唆しているようにも思えます。もちろん杞憂に終わることを望みますが、このようなオーバー・エンフォースメントが選挙前・中・後に繰り返されれば、監督委員会が最近のレポート「歴史的な選挙イヤーにおけるコンテンツモデレーション：業界への主な教訓」で警鐘を鳴らすとおり、候補者や政党への批判表現が不相応に鎮められ選挙結果の正統性にも影響しえます

（憲法上求められる2025年の衆議院選挙に先立って、2024年内に衆議院選挙が実施される可能性があると報道されています。政治資金収入の不記載の問題は有権者に支持政党を変更させうる大きな出来事であるとの見方もあります）。

更に、この問題をグローバルの視点から見ると、日本は2023年度世界GDPランキング第4位であり、日本の収益はメタの全体の収益のうちそれなりの割合を占めると思われます。ここからすると、日本のコンテンツモデレーションに十分なリソースを割り当てることはメタにとって比較的正当化しやすいと思われます。それにもかかわらず本件のような質の悪い判断がなされるということは、メタの全体の収益にあまり関係しない国におけるモデレーションについてより深刻な体制不備の可能性を示唆するようにも思えます。このような国でこそポリシーの不執行による悪影響が最も先鋭化するという、上記レポートにおける監督委員会の指摘に同意します。メタは、市民社会が繰り返し求めている通り、現地の言語の理解の拡充により一層本気で取り組むべきでです。このケースに関する監督委員会の判断が、メタのコミットメントと取組みを後押しするものとなることを願います。

スレッズ、フェイスブック、インスタグラムでの本件類似の誤判断を防ぐため、監督委員会はメタに対して、国連ビジネスと人権の指導原則に基づき、少なくとも以下の提言をすることが考えられます。

- 第1判断と第2判断の誤りの原因究明と特定された原因の除去（指導原則13及び19）。第1判断と第2判断の審査担当者の日本語の理解の不足に原因があった場合、日本語を第一言語とする者と同レベルの日本語の理解のある者に判断を対応させることとする。
- 「暴力と扇動のポリシー」に関するトレーニングの本件を踏まえたアップデート及び日本語の投稿を担当する審査担当者への周知（指導原則13及び19）
- 実施した再発防止策の効果測定（指導原則20）
- 実施した再発防止策とその効果のユーザーへの開示（指導原則21）

最後に、日本では今年、「特定電気通信による情報の流通によって発生する権利侵害等への対処に関する法律」が国会で可決されました。同法律は、ユーザーからの削除申請の対応を迅速化し、モデレーションの運用状況を透明化することを目的としています。同法24条は、大規模なプラットフォームに対し、十分な知識経験を持った侵害情報調査専門員の選任を求めています。メタは、ビジネスと人権に関する指導原則13と19に従い、憲法や国際人権法の規範―特に表現の自由に関するもの―への理解の深い専門員を選任ことが求められます。

杉山日那子

Hinako Sugiyama
A lawyer qualified in Japan (currently not registered) and admitted in New York State

To the Facebook Oversight Board,

I am a human rights lawyer supervising the work at the International Justice Clinic at the University of California, Irvine School of Law. I was raised in Japan and Japanese is my first language. This letter responds to the Oversight Board's call for public comments on Case **2024-027-TH-UA**. This opinion does not represent the views of the Clinic.

This case involves a Threads post containing the hashtag stating "*kutabare*," directed at Prime Minister Kishida. The post was a response to another post that included a statement from Kishida regarding unreported fundraising revenues involving members of his faction of the Liberal Democratic Party. In this comment, I will *first* discuss the clear error in the decision to delete this post, *second*, the long-standing protection of expression that is critical of politicians in Japan, and *third*, the measures Meta should take to prevent the recurrence of similar errors.

As I have previously noted, I appreciate that the Board adopts an appropriate approach in evaluating Meta's content moderation policies and enforcement – a participatory approach grounded in international human rights law and incorporating public comments from people and communities affected. I believe that such an approach aligns with the UN Guiding Principles on Business and Human Rights, to which Meta has committed to adhere.

1. Clear error in the decision to delete the post

In the present case, following a report from a user under Bullying and Harassment Community Standards, a human reviewer determined that the expression *kutabare* in the post amounts to calls for death and violated Violence and Incitement rule. The post was removed accordingly. Upon appeal from the author, another human reviewer also found the violation. As a result of the Board selecting the case, Meta determined *kutabare* did not amount to a threat that would violate the policy.

Article 19(2) of the International Covenant on Civil and Political Rights (ICCPR) broadly and robustly guarantees freedom of expression. Article 19(3) permits restrictions on freedom of expression only when they (i) are based on clear rules and (ii) are necessary and proportionate (iii) for respect of the rights or reputations of others or the protection of national security or of public order (ordre public), or of public health or morals (the so-called "three-part test"). Under the UN Guiding Principles on Business and Human Rights, Meta has the responsibility to limit the scope of content removal to those that pass the three-part test to respect users' freedom of expression. While preventing the spread of posts inciting violence is undoubtedly a legitimate interest, removal must be necessary and proportionate to that interest. In other words, in terms of violence, removal is only permissible when there is a specific and individualized threat of violence (see the UN Human Rights Committee's General Comment 34, para. 35). I understand that Meta intends to enforce the policy in a way that aligns individual moderation with this norm.

However, as Meta concluded in its final review, *kutabare* in the post would not fall under "threats of violence that could lead to death (or other forms of high-sensitivity violence)" as prohibited by the Violence and Incitement policy. According to *Kojien*, a widely used Japanese dictionary with a long history in Japan, the primary, literal meaning of *kutabaru* (the infinitive form of *kutabare*) is "to be weakened physically" or "to become weak." It can also mean to curse "death," but this is considered a secondary meaning. In other words, without contextual elements suggesting death or bodily harm, *kutabaru* would not mean "death." Likewise, the imperative form *kutabare* would not imply an intention to murder or cause bodily harm unless such elements are present. It is evident that there are no such elements in this post, so *kutabare* in the post would not violate the rule. This conclusion aligns with the application of the UN Rabat Plan of Action.

In other words, if the meaning of *kutabare* were correctly understood, it would not have been difficult to conclude that this post did not violate the rule. I suspect that the reviewers' mistake in their judgment was due to a misunderstanding of *kutabare*, categorizing it alongside expressions that suggest an intention to harm the recipient unless the context indicates rhetorical usage. Examples of such expressions include "you shall die" or "I will kill you," as well as the slogan "marg bar…" in the Iran Protest Slogan case in 2022, which literally means "death to" but is often used as political rhetoric to mean "down with."

Although it is not directly related to whether the post violates the Violence and Incitement rule, the context suggests that *kutabare* would mean demanding the resignation of Prime Minister Kishida.

Regarding the second point of the Board's inquiry (the extent to which rhetorical expressions or calls for violence are common and how easily such threats can be distinguished from credible threats), it seems somewhat off-topic to discuss here since the literal meaning of *kutabare* in the post is not a call for threats or violence. However, it may be worth noting that Japanese court precedents regarding the crime of intimidation (Article 222 of the Penal Code) might be a valuable resource.

2. Strong protection of expression that is critical to politicians

Next, the Board seeks information on the socio-political context in Japan including information about online threats of violence against politicians, and limitations on freedom of expression that is critical of politicians. Many in Japan, including myself, would agree that Japan possesses a culture valuing harmony, as the image held by people outside Japan. However, I'd like to emphasize that Japan's Constitution explicitly and robustly protects freedom of expression. Speech regulations are generally legislated and operated with efforts to balance freedom of expression. Most importantly, civil society in Japan has been actively exercising freedom of expression, especially regarding matters of public interest.

In particular, the norm that people in public office should be subject to severe criticism has been embodied in statutes and case law. For example:

- Article 230 of the Penal Code, which stipulates the crime of defamation, was amended in 1947 to add Article 230-1. Paragraph 3 carves out any statements concerning public officials or candidates for public office if the statements are verified. In civil defamation cases against politicians, courts have rigorously weighed in favor of freedom of expression in determining the legality or calculation of damages.
- Regarding privacy violation, the Supreme Court considered whether the alleged victim held a public position to decide whether to enjoin the publication of a novel. In recent cases, the Supreme Court also considered the same in a case concerning whether to order Twitter to delete certain tweets that included arrest records, and in a case regarding whether to order Google to de-index certain search results that included criminal records.

The Japanese government also acknowledges the need to protect expression critical of politicians. For example, during the deliberation at the Parliament on the amendment to the Penal Code in 2022 to increase penalties for criminal insult, the government stated that criticism of politicians would be carved out if it falls under legitimate acts defined in Article 35 of the Penal Code such as fair comment (see also Q9 on the Ministry of Justice's explanation).

There are no laws in Japan punishing criticism of politicians particularly, as seen in some countries, nor are there voices in Japanese civil society calling for such regulations. Instead, Japanese civil society has consistently taken action where politicians seek to suppress critical voices. For instance, during the 2019 House of Councillors election, several citizens who made critical remarks to then-Prime Minister Abe during street speeches were forced to stop by police officers. These citizens subsequently filed lawsuits, claiming a violation of their freedom of expression. Both the district court and high court recognized the police conduct as illegal.

3. Measures Meta Should Take to Prevent Recurrence of Similar Errors

Unlike the case of the Iran Protest Slogan in 2022, if the meaning of *kutabare* had been correctly understood, it would not have been difficult to determine that this post did not violate the Violence and Incitement rule. Nevertheless, both the first and second human reviewers made erroneous judgments, suggesting such errors were not coincidental but stemmed from a possible deficiency in Meta's moderation system for posts in the Japanese language. If that's the case, similar poor-quality judgments are likely to be repeated. If such over-enforcement continues before, during, or after elections, it can damage the legitimacy of election results, as cautioned the Board's recent report "Content Moderation in a Historic Election Year: Key Lessons for the Industry." Reportedly, a House of Representatives election might be held within 2024, ahead of the constitutionally mandated one in 2025. As media reports, the issue of slush funds could potentially influence voters to switch their support to different parties.

Furthermore, from a global perspective, Japan ranks fourth in the 2023 World GDP rankings, and it is presumed that Japan's revenue constitutes a significant portion of Meta's overall revenue. It thus seems relatively easy for Meta to justify allocating sufficient resources to content moderation in Japan. The kinds of poor-quality judgments in the present case in Japan would suggest a possibility of more serious systemic deficiencies in moderation in countries considered less lucrative. I agree with the Board's view in the above report that "these countries are where the human rights impact of non-compliance of standards can be most severe." Meta should make a more serious effort to expand local language expertise as repeatedly demanded by civil society. I hope the Board's decision and recommendations on this case will further support Meta's commitment and efforts.

I would recommend the Board make at least the following recommendations to Meta, based on the United Nations Guiding Principles on Business and Human Rights to prevent similar erroneous judgments on Threads, Facebook, and Instagram:

- Identify the root causes of errors in the reviewers' judgment and eliminate the identified causes (Guiding Principles 13 and 19). If the reviewers' lack of comprehension of the Japanese language caused the errors, the same kinds of moderation judgments should be made by individuals with a level of Japanese comprehension equivalent to Japanese native speakers.
- Update the training regarding the Violence and Incitement rule based on the lessons from the present case, and communicate to all reviewers of Japanese posts on the updated training (Guiding Principles 13 and 19).
- Measure the effectiveness of the implemented measures (Guiding Principle 20).
- Disclose to users the measures taken and their effectiveness (Guiding Principle 21).

Finally, this year in Japan, the Act on Handling of Infringement of Rights and Other Incidents Arising from Distribution of Information by Specific Telecommunications Services was enacted. This law aims to expedite the process of deletion requests from users and to make platforms' moderation practices more transparent. Article 24 of the same law requires large-scale platforms like Meta to appoint investigation specialists with sufficient knowledge and experience in content moderation. Meta should appoint specialists with a deep understanding of Constitutional and international human rights norms, especially those related to freedom of expression, in accordance with Guiding Principles 13 and 19.