Political dispute ahead of Turkish elections

2023-007-FB-UA, 2023-008-FB-UA, 2023-009-IG-UA

Case summary

The Oversight Board has overturned Meta's original decisions to remove the posts of three Turkish media organizations, all containing a similar video of a politician confronting another in public, using the term "İngiliz uşağı," which translates as "servant of the British." The Board finds that the term is not hate speech under Meta's policies. Furthermore, Meta's failure to qualify the content as permissible "reporting," or to apply the public newsworthiness allowance, made it difficult for the outlets to freely report on issues of public interest. The Board recommends that Meta make public an exception for permissible reporting on slurs.

About the cases

For these decisions, the Board considers three posts – two on Facebook, one on Instagram – from three different Turkish media organizations, all independently owned. They contain a similar video featuring a former Member of Parliament (MP) of the ruling party confronting a member of the main opposition party in the aftermath of the Turkish earthquakes in February 2023. In the run-up to the Turkish elections, the earthquakes were expected to significantly impact voting patterns.

The video shows Istanbul's Mayor Ekrem İmamoğlu, a key opposition figure, visiting one of the most heavily impacted cities when he is confronted by a former MP, who shouts that he is "showing off," calls him a "servant of the British," and tells him to return to "his own" city. Both the public and expert commentators confirm the phrase "İngiliz uşağı" is understood by Turkish speakers to mean "a person who acts for the interests and benefits" of Britain or the West in general.

Meta removed all three posts for violating its Hate Speech policy rule against slurs. Although several of Meta's mistake-prevention systems had been engaged, including cross-check,

which led to the posts in each case undergoing several rounds of human review, this did not result in the content being restored.

In total, the posts were viewed across the three accounts more than 1,100,000 times before being removed.

While the three users were notified they had violated the Hate Speech Community Standard, they were not told the specific rule they had broken. Additionally, feature limits to the accounts of two of the media organizations were applied, which prevented one from being able to create new content for 24 hours, and another losing its ability to livestream video for three days.

After the Board identified the cases, Meta decided that its original decisions were wrong because the term "İngiliz uşağı" should not have been on its slur lists, and it restored the content. Separately, Meta had been conducting an annual audit of its slur lists for Turkey ahead of the elections, which led to the term "İngiliz uşağı" being removed in April 2023.

Key findings

The role of the media in reporting information across the digital ecosystem is critical. The Board concludes that removing the three posts was an unnecessary and disproportionate restriction on the rights of individuals in the Turkish media organizations and on access to information for their audience. Furthermore, Meta's measures in these cases made it difficult for two of the three organizations to freely share their reporting for the duration of the feature limits on their accounts. This had real impact since the earthquakes and run-up to the elections made access to independent local news especially important.

The Board finds that the term "İngiliz uşağı" is not hate speech under Meta's policies because it does not attack people on the basis of "a protected characteristic." The public confrontation in the videos involves politicians from competing political parties. Since the term used has historically functioned as political criticism in Türkiye (Turkey), it is political speech on a matter of significant public interest in the context of elections.

Even if Meta had designated the term correctly as a slur, the content should nevertheless have been allowed because of its public interest value. The Board is concerned the three

posts were not escalated for an assessment under the newsworthiness allowance by Meta's Core Policy Team.

Meta's policies also allow people to share hate speech and slurs to raise awareness of them, provided the user's intent is clear. In responses to these cases, Meta has explained that in order to "qualify as reporting that is awareness raising, it is not enough to restate that someone else used hate speech or a slur. Instead, we [Meta] need specific additional context." None of the media organizations in these cases would have qualified because the content was shared with a neutral caption, which would not have been considered sufficient context. The politician's use of the term in the video was not the main story being told, so a caption focused on explaining or condemning it would not have made sense. Rather, the main news story was the disagreement between politicians in the context of the earthquake response.

Finally, the Board finds that Meta should make public that reporting on hate speech is permitted, ideally in a standalone exception that distinguishes journalistic "reporting" from "raising awareness." Meta's internal guidance seems to permit broader exceptions than those communicated publicly to users at present. This information would be especially important to help media organizations to report on incidents during which a slur has been used by third parties in a matter of public interest, including when it is not the main point of the news story. The framing of this information should recognize that media outlets and others engaged in journalism may not always state intent for "raising awareness," in order to impartially report on current events.

The Oversight Board's decision

The Oversight Board overturns Meta's original decisions to remove three posts.

The Board recommends that Meta:

1. Revise the Hate Speech Community Standard to explicitly protect journalistic reporting on slurs when such reporting, in particular in the context of elections, does not create an atmosphere of exclusion and/or intimidation. This exception should be made public, be separate from the "raising awareness" exception, and make clear to users, especially in the media, how such content should be contextualized. There

should also be appropriate training to moderators, especially outside of English languages, to ensure respect for journalism.

2. Ensure the Hate Speech Community Standard has clearer explanations of each exception, with illustrative examples, to ensure greater clarity about when slurs can be used.

3. Expedite audits of its slur lists in countries with elections for the remainder of 2023 and early 2024, with the goal of identifying and removing terms mistakenly added to those lists.

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

Full case decision

1. Decision summary

The Board overturns Meta's original decisions to remove the posts of three Turkish media organizations – BirGün Gazetesi, Bolu Gündem, and Komedya Haber – which all contained a similar video. The videos all featured Ms. Nursel Reyhanlıoğlu, a former Member of Parliament of President Erdoğan's Justice and Development Party (AKP party), referring to Istanbul Mayor Ekrem İmamoğlu, a member of the largest opposition party in Türkiye (Turkey), as an "İngiliz uşağı," translated as "servant of the British." In all three cases, Meta removed the video for violating its Hate Speech Community Standard, which prohibits "slurs that are used to attack people on the basis of their protected characteristics." After the Board identified these cases, Meta reversed each of its decisions to remove the posts, deciding the term "İngiliz uşağı" should not be on its internal slur list.

2. Case description and background

On February 6, 2023, a series of powerful earthquakes struck southern Türkiye (Turkey) near the northern border of Syria. The disaster killed over 50,000 people in Türkiye (Turkey) alone,

injured more than 100,000, and triggered the displacement of three million people in the provinces most affected by the tremors. On February 8, 2023, Istanbul Municipality Mayor Ekrem İmamoğlu, a member of the main opposition party, the Republican People's Party (CHP), visited Kahramanmaraş, one of the cities impacted by the disaster. During his visit, a former Member of Parliament (MP) from the ruling Justice and Development Party (AKP), Nursel Reyhanlıoğlu, confronted him. In the recorded confrontation, former MP Reyhanlıoğlu shouted at Mayor İmamoğlu that he was "showing off" with his visit, calling him a "British servant" (Turkish: İngiliz uşağı), and that he should "get out" and return to "his own" Istanbul.

Public comments and experts the Board consulted confirmed that the phrase "İngiliz uşağı" is understood by Turkish speakers to mean "a person who acts for the interests and benefits of the British nation or government officials or the West in general." External experts underlined that implying that someone is betraying their own country by serving the interests of foreign powers can be a serious and damaging accusation as it questions a person's loyalty and commitment to their own country, particularly in a political context.

The three media organizations in these cases do not have ties to the Turkish government and are independently owned. External experts noted that BirGün Gazetesi has had the most contentious relationship with the government. One of its columnists, Turkish-Armenian journalist Hrant Dink, was assassinated in 2007 and the paper has also been subject repeatedly to criminal prosecution.

In the immediate aftermath of the February earthquakes, there was significant attention on the presidential and parliamentary elections due to take place in May. Meta announced in an April 2023 blog post that it was ready to combat "misinformation" and "false news" in the upcoming Turkish election. Experts the Board consulted described how election observers had expected the earthquakes to impact voting patterns. One of the main points of criticism centered on the government's legislation to provide amnesty to construction companies for erecting buildings that failed to meet safety codes, a law that Reyhanlıoğlu supported as an MP in 2018. Public criticism of disaster management agency Afet ve Acil Durum Yönetimi Başkanlığı (AFAD) for failures in its earthquake response became an election issue.

In the first case that the Board accepted on appeal, the Turkish news site page Bolu Gündem posted the video of the confrontation to its Facebook page. Users reported the post, and it was queued for moderator review. At the time of review, Meta had enabled a mistake prevention

system known as Dynamic Multi Review, which allows for jobs to be assessed by multiple reviewers in order to get a majority outcome. Two out of three reviewers found that the content violated Meta's Hate Speech policy, and one reviewer found it did not. Due to the Early Response Secondary Review (ERSR) protocol, which is a form of cross-check, the content was escalated for secondary review rather than being immediately removed (The various mistake prevention systems engaged in these cases are further explained in Section 8.1). During this secondary review, two reviewers found that the content violated the Hate Speech policy, and it was removed. Meta applied a strike and 24-hour feature limit to this case's content creator's account (and not to the page), which prevented the user from creating new content on the platform (including any pages they administer) and creating or joining Facebook messenger rooms. Before being removed, the post was viewed more than one million times.

In the second case, the Turkish media outlet BirGün Gazetesi posted a longer video including the same confrontation as the other two shorter videos as a live stream on its Facebook page. After the live stream ended, it became a permanent post on the page. Distinct from the other two videos, it included further footage of Mayor İmamoğlu and CHP leader and presidential candidate Kemal Kılıçdaroğlu speaking to two members of the public. In the conversation that followed the confrontation, the two members of the public requested more aid to rescue people trapped under the rubble and expressed frustration at the government's emergency response. A user reported the Facebook post for violating Meta's policies. At the time of review, Meta had enabled Dynamic Multi-Review (see section 8.1), and two out of three reviewers found the content violated the Hate Speech policy, while one reviewer found it did not. The content was sent for additional review due to the General Secondary Review (GSR) ranker, which is another cross-check protocol running alongside ERSR. The GSR algorithm ranks content for additional review based on criteria such as topic sensitivity, enforcement severity, false-positive probability, predicted reach, and entity sensitivity (see further explanation of this protocol in Section 8.1 and cross-check policy advisory opinion, para 42). A reviewer in Meta's regional market team determined the post violated the Hate Speech policy and it was removed. Meta applied a standard strike to both the content creator's profile and the Facebook page, but it did not apply any feature limits (such as restricting the ability to post) because the number of strikes did not reach the necessary threshold. Before being removed, the post was viewed more than 60,000 times.

In the third case, a digital media outlet called Komedya Haber posted the video to Instagram. A classifier designed to identify the "most viral and potentially violating content" detected the

content as potentially violating the Hate Speech policy, lining it up for moderator review. The reviewer found that the content violated the Hate Speech policy. Later, a user reported the content for violating Meta's policies. At the time of review, Meta had enabled Dynamic Multi-Review, and two reviewers found that the content violated the Hate Speech policy. The GSR ranker prioritized the content for additional review, so the post was sent to another moderator. Based on information from Meta in the cross-check policy advisory opinion, the GSR review is conducted by either an employee or a contractor on Meta's Regional Market Team (cross-check policy advisory opinion, page 21). Through GSR, a reviewer assessed the content as violating the Hate Speech policy and it was removed. Meta applied a standard strike resulting in a three-day feature limit preventing the Instagram account from using live video. Before being removed, the post was viewed more than 40,000 times.

After each post was removed, all of the three users were notified that they violated Meta's Hate Speech Community Standard, but not the specific rule within that policy they had broken. The notifications the two Facebook users received stated that hate speech includes "attacks on people because of their race, ethnicity, religion, caste, physical or mental ability, gender, or sexual orientation" and lists several examples, but do not mention slurs. The Instagram user received a shorter notification, stating that the content was removed "because it goes against our [Instagram] Community Guidelines, on hate speech or symbols."

Though Meta applied a 24-hour feature limit on the content creator who posted the content to Bolu Gündem's Facebook page, the user notification did not alert the user to the restriction. On Instagram, Komedya Haber received a notification that its Instagram account was temporarily restricted from creating live videos. All three users then appealed Meta's decision to remove the content, and Meta's reviewers again concluded that each post violated the Hate Speech policy. Each user was notified that the content had been reviewed once more but that the content violated Facebook's Community Standards or Instagram's Community Guidelines. The appeal messages did not tell them which policy was violated.

As a result of the Board selecting these three appeals, Meta identified that all three of its original decisions were wrong, and restored the content on each account on March 28, 2023, reversing the applicable strikes. By this point, the feature limits applied to two of the cases had already expired. Meta explained to the Board that the phrase was not used as a slur and therefore the three posts did not violate the Hate Speech policy. Between January and April 2023, Meta was conducting an annual audit of its slurs list for the Turkish market, which

eventually led to the phrase “İngiliz uşağı” being removed from the list. This took place in parallel to the Board selecting these three cases, which, in line with regular process, led to Meta reviewing its original decisions. Through that review, the company also determined “İngiliz uşağı” was not a slur.

3. Oversight Board authority and scope

The Board has authority to review Meta’s decision following an appeal from the person whose content was removed (Charter Article 2, Section 1; Bylaws Article 3, Section 1). The Board may uphold or overturn Meta’s decision (Charter Article 3, Section 5), and this decision is binding on the company (Charter Article 4). Meta must also assess the feasibility of applying its decision in respect of identical content with parallel context (Charter Article 4). The Board’s decisions may include non-binding recommendations that Meta must respond to (Charter Article 3, Section 4; Article 4). The Board monitors implementation of recommendations Meta has committed to act on, and may follow-up on any prior recommendation in its case decisions.

When the Board selects cases like these, in which Meta subsequently acknowledges that it made an error, the Board reviews the original decisions to increase understanding of the content moderation process and to make recommendations to reduce errors and increase fairness for people who use Facebook and Instagram. The Board further notes that Meta’s reversal of its original decisions in each of these cases was partly based on a change to its internal guidance after each post was made, removing the phrase “İngiliz uşağı” from its non-public slur list in April 2023. The Board understands that at the time of the company’s original decisions, Meta’s at-scale reviewers applied the policy and internal guidance that were in force at the time.

When the Board identifies cases in which the appeals give rise to similar or overlapping issues, including related to content policies or their enforcement, or Meta’s human rights responsibilities, they may be joined and assigned to a panel to deliberate the appeals together. A binding decision will be made in respect of each post.

4. Sources of authority and guidance

The following standards and precedents informed the Board’s analysis in this case:

I. Oversight Board decisions

The most relevant previous decisions of the Oversight Board include:

- Armenians in Azerbaijan case (2020-003-FB-UA)
- Depiction of Zwarte Piet case (2021-002-FB-UA)
- Colombia protests case (2021-010-FB-UA)
- South Africa slurs case (2021-011-FB-UA)
- Reclaiming Arabic words case (2022-003-IG-UA)
- Mention of the Taliban in news reporting case (2022-005-FB-UA)
- Iran protest slogan case (2022-013-FB-UA)

II. Meta's content policies

The Instagram Community Guidelines state that content containing hate speech will be removed. Under the heading "Respect other members of the Instagram community," the guidelines state that it is "never OK to encourage violence or attack anyone based on their race, ethnicity, national origin, sex, gender, gender identity, sexual orientation, religious affiliation, disabilities, or diseases." The Instagram Community Guidelines do not mention any specific rule on slurs, but the words "hate speech" link to the Facebook Community Standard on Hate Speech.

In the rationale for its Hate Speech policy, Meta prohibits "the usage of slurs that are used to attack people on the basis of their protected characteristics." Protected characteristics in Meta's policy include, for example, national origin, religious affiliation, race, and ethnicity. At the time each of the three posts were created, removed and appealed to the Board, and when Meta reversed its original decisions in all three cases, "slurs" were defined as "words that are inherently offensive and used as insulting labels for the above characteristics." Following a policy update on May 25, 2023, Meta now defines "slurs" as "words that inherently create an atmosphere of exclusion and intimidation against people on the basis of a protected characteristic, often because these words are tied to historical discrimination, oppression, and violence" and adds that "they do this even when targeting someone who is not a member of the [protected characteristic] group that the slur inherently targets."

The policy rationale also outlines several exceptions that allow the use of a slur "to condemn it or raise awareness" or to be used "self-referentially or in an empowering way." However, Meta may still remove the content "if the intention is unclear." The May 25 revisions to the Hate Speech policy did not alter this language.

In addition to the exceptions set out in the Hate Speech policy, the newsworthiness allowance allows "content that may violate [the] Facebook Community Standards or Instagram Community Guidelines, if it's newsworthy and if keeping it visible is in the public interest." Meta only grants newsworthiness allowances "after conducting a thorough review that weighs the public interest against the risk of harm" and looks to "international human rights standards, as reflected in [its] Corporate Human Rights Policy, to help make these judgments." Meta states it assesses whether content raises "an imminent threat to public health or safety, or gives voice to perspectives currently being debated as part of a political process." This assessment takes into account country circumstances such as whether an election or conflict is under way, whether there is a free press, and whether Meta's products are banned. Meta states there is "no presumption that content is inherently in the public interest solely on the basis of the speaker's identity, for example their identity as a politician."

The Board's analysis was informed by the Meta's commitment to "Voice," which the company describes as "paramount", and its values of "Safety," "Privacy" and "Dignity."

III. Meta's human rights responsibilities

The UN Guiding Principles on Business and Human Rights (UNGPs), endorsed by the UN Human Rights Council in 2011, establish a voluntary framework for the human rights responsibilities of private businesses. In 2021, Meta announced its Corporate Human Rights Policy, in which it reaffirmed its commitment to respecting human rights in accordance with the UNGPs.

The Board's analysis of Meta's human rights responsibilities in this case was informed by the following international standards:

- The rights to freedom of opinion and expression: Articles 19 and 20, International Covenant on Civil and Political Rights (ICCPR), General Comment No. 34, Human

Rights Committee, 2011; UN Special Rapporteur on freedom of opinion and expression, reports: A/HRC/38/35 (2018), A/74/486 (2019); Joint Declaration on Media Freedom and Democracy, UN and regional mandates on freedom of expression (2023).

- The rights to participation in public affairs and to vote: Article 25, ICCPR.
- The rights to equality and non-discrimination: Article 2 and 26, ICCPR.
- The right to protection of the law against unlawful attacks on honour and reputation: Article 17, ICCPR.

5. User submissions

All three media outlets separately appealed Meta's removal decisions to the Board. In its appeal to the Board, Bolu Gündem pointed out that it paid a news agency for the video and that other news organizations had shared the video on Facebook without it being removed. BirGün Gazetesi emphasized the public's right to receive information, while Komedya Haber's appeal contested that the content included hate speech.

6. Meta's submissions

Meta removed all three posts under its Hate Speech Community Standard, because the phrase "İngiliz uşağı" was, from the time the videos were posted to when they were reinstated, a designated slur in Meta's Turkish market, translated as "servant of the British."

At the time the three posts were removed, the Community Standard defined slurs consistent with the public-facing policy as "words that are inherently offensive and used as insulting labels for [...] protected characteristics" including national origin. Meta shared with the Board that, following an internal Policy Forum, it decided to move away from the concept of "inherently offensive" as its basis for describing slurs towards "a research-based definition focused on the word's connection to historical discrimination, oppression, and violence against protected characteristic groups." Meta has shared with the Board that this definitional change did not impact operational guidance to reviewers on how to implement the policy. The only change that would have impacted the outcome of these cases was the removal of "İngiliz uşağı" from the slur list.

Meta explained that its policies "allow people to share hate speech and slurs to condemn, to raise awareness, self-referentially, or in an empowering way. However, the user's intent must be clear. In order to qualify as reporting that is awareness raising, it is not enough to restate that someone else used hate speech or a slur. Instead, we [Meta] need specific additional context." In response to the Board's questions, Meta clarified that it allows slurs in a "reporting" context only when shared to raise awareness about the use of the slur with "specific additional context" and that "a neutral caption is not enough." Meta explained that it didn't apply this exception in these posts because the videos did not include clear awareness-raising or condemning context.

In its response to the Board's questions, Meta stated that the newsworthiness allowance was not necessary to apply in these cases because the content did not contain a violating slur. However, at the time of the original removals, Meta did consider the phrase to be a slur. For that scenario, Meta added that it would find that the public interest value of the content in the context of an election to outweigh any risk of harm, so it would also have restored the content. For the Board's assessment of newsworthiness, see Section 8.1.

In November 2022, Meta staff identified the need to update the Turkish slur list as part of the company's preparations for the May 2023 presidential and parliamentary elections in Türkiye (Turkey). The annual audit of the country's market slur list began in January 2023 and the company's regional team submitted its proposed changes in mid-March 2023. In its audit, Meta decided that the phrase "İngiliz uşağı" did not constitute a slur and removed it, effective April 12, 2023, two weeks after the content was restored in all three cases. At the same time, Meta removed from its slur list other terms that combined the use of "uşak" (servant) with specific nationalities. In response to the Board's questions, Meta stated it does not have documentation on when and why the phrase was originally designated as a slur, but it now recognizes it does not attack people based on a protected characteristic. The company also added that "İngiliz uşağı" was still on the slur list for the Turkish market at the time the three posts in this case were reviewed and therefore moderators acted in accordance with internal guidance by removing the content.

Meta audits its slur lists through a process led by regional market teams "with the goal of de-designating any slurs that should not be on the lists" in January each year. Meta used a new auditing process that was trialled in the 2023 annual audit of the Turkish market slur list. The new process involves two steps: first, a qualitative analysis to determine the history and use

of the term; and second, a quantitative analysis, to determine key data questions such as how much of the sample falls within policy exceptions. Meta explained that because "İngiliz uşağı" did not qualitatively meet its "slurs" definition (step one), it was removed from the list without progressing to a quantitative analysis of its use (step two).

The Board asked Meta 23 questions in writing. The questions addressed issues related to the criteria and processes for slur designation; the internal guidance on slurs and application of policy exceptions; how mistake prevention systems operated differently in the reviews of the three posts, and evaluation of account level enforcements resulting from each content decision. Of the 23 questions, 22 were answered and one partially. The partial response was about when and why the phrase "İngiliz uşağı" was designated as a slur, with the company explaining that it lacked documentation. Meta also provided the Board with an oral briefing on the changes to its slurs definition and designation process.

7. Public comments

The Oversight Board received 11 public comments relevant to these three cases. One of the comments was submitted from Central and South Asia; nine from Europe; and one from the United States and Canada. The submissions covered the following themes: the importance of a contextual approach to moderating slurs; proper user notice of the reasons for content removals; the effects of erroneous removals of content on news outlets; the relevance of newsworthiness allowance to the content; and calls for a public list of slur examples.

To read public comments submitted for this case, please click here.

8. Oversight Board analysis

The Board examined whether to uphold or overturn Meta's original decisions in these three cases by analyzing Meta's content policies, human rights responsibilities and values. Taking these decisions together also provides the Board with a greater opportunity to assess their implications for Meta's broader approach to content governance, particularly in the context of elections.

8.1 Compliance with Meta's content policies

I. Content rules

Hate Speech

The Board finds that the term "İngiliz uşağı" in these three cases is not hate speech under Meta's Community Standards. Whether assessed against the definition of slurs prior to or following the May 25, 2023 policy changes, the term "İngiliz uşağı" does not attack individuals on the basis of a protected characteristic. The removal of content containing this term in all three cases is inconsistent with the rationale of the Hate Speech policy, as it does not attack people on the basis of a protected characteristic.

The term "İngiliz uşağı" has a long history functioning as political criticism in Türkiye (Turkey). According to experts consulted by the Board, the use of the phrase preceded the founding of modern Türkiye (Turkey), when the term was used to criticize leaders in the Ottoman Empire for serving the interests of Britain, and the term is not discriminatory in nature. The confrontation in these three cases involves politicians from competing political parties. The AKP, MP Reyhanlıoğlu's party, has faced criticism and public anger over the government's handling of the earthquake response and its legislation granting amnesty to building developers for constructing buildings that did not adhere to earthquake safety codes. She directed the slur at Mayor İmamoğlu, a key figure of the CHP, the country's largest opposition party. The tense relationship between the AKP and CHP leading up to the election, including the importance of the earthquake as an electoral topic, played out publicly during Mayor İmamoğlu and CHP presidential candidate Kemal Kılıçdaroğlu's visit to Kahramanmaraş. The content in each of the three cases is therefore political speech on a matter of significant public interest in the electoral context.

As Meta has explained, in order "to qualify as reporting that is awareness raising, it is not enough to restate that someone else used hate speech or a slur. In other words, a neutral caption is not enough." If the content had included a slur, none of the media organizations would have qualified as "discussing" or "reporting" hate speech because the content was shared with a neutral caption in all three cases. In the Board's view, even if this slur was appropriately designated on the list, the content in all three cases should nevertheless have been protected as "reporting." currently framed, if the content had included a slur, none of

the media organizations would have qualified as “discussing” or “reporting” hate speech because the content was shared with a neutral caption in all three cases. In the Board’s view, even if this slur was appropriately designated on the list, the content in all three cases should nevertheless have been protected as “reporting.”

The Board finds that the phrase “İngiliz uşağı” should not have been added to Meta’s confidential slur list, as it is not a form of hate speech. In other contexts, accusations of being a “foreign agent” may amount to a credible threat to individuals’ safety, but these can be addressed under other policies (for example, under Violence and Incitement). Even in those situations, Meta should distinguish threats from a speaker in an influential position from media reporting on those threats. Given the facts of these cases and the internal guidance in place at the time, content reviewers, who are moderating content at scale, acted in accordance with that guidance to remove content containing terms on Meta’s slur lists. At the time, that list included “İngiliz uşağı.” The reason for the errors in these cases was the policy decision to add the term to the slur list and the inappropriately narrow and confidential guidance on how reviewers should apply the “raising awareness” exception to posts “reporting” on slur usage.

Newsworthiness allowance

The Board expresses its concern that, at a time when Meta’s internal policies categorized “İngiliz uşağı” as a violating slur, the three posts were not escalated for a newsworthiness allowance assessment by Meta’s Core Policy Team (previously known within the company as the “Content Policy Team”).

Turkish freedom of expression organization İfade Özgürlüğü Derneği (İFÖD) argued in its public comment that because of its public interest value, the content in all three cases should have qualified for a newsworthiness allowance. If the content contained a slur properly designated in accordance with Meta’s Hate Speech policy, the Board would agree. The three posts concern reporting on speech by one (former) politician, targeting a current politician, in a way that is within the boundaries of (even offensive) criticism that a politician should be expected to tolerate, including insulting epithets. That assessment could be different, for example, if a term was used in its particular context as a discriminatory slur. The video emerged at a moment of significant political and social importance after a series of devastating earthquakes had struck Türkiye (Turkey). The earthquakes, as well as discussions

related to the government response and preparation for them, were important topics for President Erdoğan and CHP Presidential Candidate Kemal Kılıçdaroğlu in the campaign period prior to the May 2023 elections. In the aftermath of the earthquakes, the Turkish government also temporarily restricted access to Twitter and other social media sites as criticism of the government's earthquake response spread. Since this footage was in the public interest and its removal would not reduce any risk of harm, Meta should have allowed the term to be used for public interest reporting, even if it had properly qualified as a slur. The Board has previously insisted that Meta leave up content containing discriminatory slurs when the content otherwise related to significant moments in a country's history (see Colombia protests case).

II. Systemic challenges for enforcement and error prevention

Slur list designation and audit processes

Meta could not provide the Board with information on when or why it originally designated "İngiliz uşağı" as a slur because of insufficient documentation, a concern it seeks to address with its new slur designation and audit processes.

Under the previous auditing process, the company's regional teams with the support of policy and operations experts would conduct qualitative and quantitative analysis on the language and culture of the related region or market to create slur lists. This process would include reviewing the word's associated meaning, its prevalence in Meta's platforms, and its local and colloquial usage. Meta had required collecting and assessing at least 50 pieces of content containing that term in this process. However, Meta noted in its recent Policy Forum that the previous slur designation process had a number of issues, including indexing on offensiveness, lack of documentation, and subjective criteria; and as Meta noted to the Board, this was "inconsistently applied" with removal criteria not fixed or weighted.

When Meta was trialing its new designation process in 2023 for the Turkish market, "İngiliz uşağı" was removed from the slur list. By coincidence, that audit was ongoing at the time the Board selected these three cases. The term had been on Meta's slur list since at least 2021. According to Meta, the new process intends to better quantify alternative meanings and usages of a term for removing a slur designation, a process that focuses on better accounting

for the changing meanings of words over time.

These governance changes are generally positive, and if effectively implemented should reduce over-enforcement of the slurs policy. However, the new process would be enhanced if it specifically aimed to identify terms that were incorrectly added to the slur list. Meta should also ensure it updates and makes more comprehensive its explanation of slurs designation and auditing in the Transparency Center, aligning this with its new definition of slurs and its revised approach to slur lists audits.

Mistake prevention measures and escalation challenges

Reviewing these three cases together allowed the Board to assess how a variety of Meta's mistake-prevention systems worked with respect to similar content and revisit a broader systematic challenge it has also noted in prior decisions. The Board is concerned that while various mistake-prevention systems were engaged in the review of each post, it appears they did not operate consistently for the benefit of media organizations or their audiences. In addition, the measures did not empower reviewers to escalate any of the three posts for further contextual review. Such escalations could have either led to the content being left up (e.g., for newsworthiness), and/or the error in adding this term to the slur list being identified earlier, outside of the annual audit.

Cross-check was engaged in all three cases, but operated differently in the decision for each post. Only Bolu Gündem was listed as a media organization for the purpose of Early Response Secondary Review (ERSR), whereas BirGün Gazetesi and Komedya Haber were not. ERSR is the entity-based form of cross-check, for which any post from a listed entity receives additional review if marked for removal (see cross-check policy advisory opinion, paras 27-28). Of the three media organizations in these cases, only Bolu Gündem had a partner manager. According to Meta, a media organization must have a "partner manager" to be eligible for ERSR. According to Meta, partner managers "act as the link between external organizations and individuals who use Meta's platforms and services" and they help account holders "optimize their presence and maximize the value they generate from Meta's platforms and services." The Board notes that the posts from BirGün Gazetesi and Komedya Haber both received cross-check review under General Secondary Review (GSR), which prioritizes content based on the "cross-check ranker." Nevertheless, the Board is concerned that local or smaller media entities are not systematically included as ERSR listed entities as

they do not have a partner manager. This reinforces concerns the Board expressed in its policy advisory opinion on cross-check about the program's lack of transparency, and lack of objective criteria for inclusion in ERSR. Entities engaged in public interest journalism ought to have access to clear information on how their accounts can benefit from cross-check protection; if having a partner manager is a necessary condition for inclusion, there should be clear instructions on applying for a partner manager. In addition, the Board is concerned that the fact that all three posts were reviewed through cross-check did not lead to closer consideration of whether a policy exception should have applied, and/or an escalation to be made for a newsworthiness assessment.

Moreover, Dynamic Multi-Review (DMR) was also "turned on" for the applicable review queue at the time the three posts were sent for initial moderator review. For the purpose of DMR, automation identified all three posts for multiple moderator reviews prior to removal, for the accuracy of human review and to mitigate the risk of incorrect decisions based on several factors such as virality and number of views. Out of a total of eight reviews across the three similar posts, which all preceded the additional cross-check reviews, only two reviewers (of one post each) determined those posts did not violate the Hate Speech policy. The Board is concerned that reviewers are not prompted when automation is identifying a higher risk of enforcement error, as this might encourage them to examine the content more closely, either to consider potentially applicable policy exceptions and/or to escalate the content for closer contextual analysis.

Meta's current mistake-prevention measures, in both DMR and cross-check, appear to be almost entirely geared towards ensuring moderators enforce the policies in line with internal guidance. They do not contain, it seems, additional mechanisms for reviewers to identify when strict adherence to Meta's internal guidance is leading to the wrong decision, because the policy itself is wrong (as Meta later admitted was the case with respect to all three of its initial decisions). While automation correctly identified that the posts in all three cases were at risk of false-positive enforcement, the additional reviews by moderators did not lead to escalations for applying a newsworthiness allowance. Given the challenges of false positives in at-scale review, escalations should be more systematic and frequent for content relating to public interest debates, in particular in the context of elections. The fact that Meta applied its resource-intensive mistake-prevention systems to these cases, but still reached incorrect outcomes in all three, shows that they require further review. Meta previously dismissed similar concerns the Board raised about escalation pathways for newsworthiness

assessments in the "Colombia protests" decision (see Meta's response to "Colombia protests" recommendation no. 3) as it felt the work it was already doing was sufficient. The Board finds that these three cases demonstrate this issue requires re-examination.

8.2 Compliance with Meta's human rights responsibilities

The Board finds that Meta's decision to remove the content in all three cases was inconsistent with Meta's human rights responsibilities.

Freedom of expression (Article 19 ICCPR)

Article 19 of the ICCPR provides for broad protection of expression, including the "freedom to seek, receive, and impart information and ideas of all kinds." The scope of the protection includes expression that "may be regarded as deeply offensive" (General Comment 34, para. 11). The protection of expression is also "particularly high" when public debate concerns "figures in the public and political domain" (General Comment 34, para. 34). The role of the media in reporting information across the digital ecosystem is critical. The Human Rights Committee has stressed that a "free, uncensored and unhindered press or other media is essential" with press or other media being able to "comment on public issues without censorship or restraint and to inform public opinion" (General Comment 34, para. 13).

The expression at issue in each of these three cases deserves "particularly high" protection because the political dispute came during a significant political debate concerning the government's earthquake response in the lead up to presidential and parliamentary elections in Türkiye (Turkey). Public anger and criticism after the earthquakes came as President Erdoğan and CHP presidential candidate Kemal Kılıçdaroğlu were campaigning in the months before the May 2023 presidential and parliamentary elections. In the Joint Declaration on Media Freedom and Democracy, UN and regional freedom of expression mandate holders advise that "large online platforms should privilege independent quality media and public interest content on their services in order to facilitate democratic discourse" and "swiftly and adequately remedy wrongful removals of independent quality media and public interest content, including through expedited human review" (Recommendations for social media platforms, page 8).

Where 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, both in relation to the individual content decision under review and what this says about Meta's broader approach to content governance. As the UN Special Rapporteur on freedom of expression has stated, although "companies do not have the obligations of Governments, their impact is of a sort that requires them to assess the same kind of questions about protecting their users' right to freedom of expression" (A/74/486, para. 41).

I. Legality (clarity and accessibility of the rules)

The principle of legality requires rules that limit expression to be clear and publicly accessible (General Comment No. 34, para. 25). The Human Rights Committee has further noted that rules "may not confer unfettered discretion for the restriction of freedom of expression on those charged with [their] execution" (*Ibid.*). In the context of online speech, the UN Special Rapporteur on freedom of expression has stated that rules should be specific and clear (A/HRC/38/35, para. 46).

Meta's hate speech prohibition on slurs is not sufficiently clear to users. Meta's slurs definition prior to May 25, 2023, focused on offensiveness, which was excessively subjective and much broader than Meta's definition of hate speech as framed in the policy rationale. Prior to the changes, Facebook and Instagram users were likely to have different interpretations of what "offensive" meant, creating confusion that may include circumstances where there is an attack on a protected characteristic, others where there is not. The May 25 changes have clarified Meta's policy position to some extent, moving away from the vague concept of "offense."

The notifications in each of the three cases did not inform the respective users that the posts were removed because of slur usage, only that the content was removed for violating Meta's Hate Speech policy. In its Q2 2022 update on the Oversight Board, Meta stated they are "planning on assessing the feasibility of further increasing the depth by adding additional granularity to which aspect of the policy has been violated at scale (e.g., violating the slurs prohibition within the Hate Speech Community Standard)." Meta noted in this report that its

review systems are most accurate at the policy level and accordingly prioritize "correct, broader messaging" over "specific, yet inaccurate messaging." For example, Meta has greater confidence it can accurately inform users they have violated the Hate Speech policy, but has less confidence it can accurately inform users the specific rule within that policy (e.g., prohibition on slurs) they have violated. In Meta's response to the "South Africa slurs" case recommendation, however, the company said it is "building new capabilities to provide more detailed notifications" which is now offered in English, with testing in Arabic, Spanish, and Portuguese notifications on Facebook. This would not have benefited the users in these cases because the Board understands the notifications the users received were in Turkish. The Board urges Meta to provide this level of detail for non-English users.

Meta's list of exceptions to the prohibition on slurs, and hate speech more broadly, could be explained more clearly to users and content reviewers. Though the Board has reservations with requiring clear statements of intent as a requirement to benefit from exceptions, to the extent intent should be a necessary consideration, Meta needs to more clearly specify to users how they can demonstrate intent for each of the policy exceptions listed. In addition, internal guidance for reviewers seems to permit broader exceptions than those communicated publicly to users, creating accessibility and clarity concerns. Meta's policy guidance states that "reporting" is permitted under the Hate Speech policy when it is raising awareness. The Board has previously criticized Meta's public-facing Hate Speech policy for failing to explain rules that are contained in internal guidance to reviewers (see, e.g., Two buttons meme case). Meta should make public that reporting on hate speech is permitted, ideally in a standalone exception that distinguishes journalistic "reporting" from "raising awareness". This information is particularly important to aid media organizations and others who wish to report on incidents during which a slur has been used by third parties in a matter of public interest, including when the slur is incidental to or not the main point of the news story, in ways that do not create an atmosphere of exclusion and/or intimidation. It should be framed in such a way that recognizes that media outlets and others engaged in journalism, in order to impartially report on current events, may not always state intent for "awareness raising" and that this may need to be inferred from other contextual cues.

II. Legitimate aim

Any restriction on expression should pursue one of the legitimate aims listed in the ICCPR, which include the "rights of others." In several decisions, the Board has found that Meta's

Hate Speech policy, including the slurs prohibition, pursues the legitimate aim of protecting the rights of others, namely not to be discriminated against (see, for example, "Armenians in Azerbaijan" decision).

The Board notes that Meta's May 25 update to its slurs definition has made clearer this aim. Prior references to slurs as "inherently offensive" may have been read to imply a right of individuals to protection from offensive speech *per se.* This would not be a legitimate aim, as no right to be protected from offensive speech exists under international human rights law. Meta's new definition, substituting the concept of offensiveness for a more objective definition for terms that "inherently create an atmosphere of exclusion and intimidation against people on the basis of a protected characteristic" more closely aligns with the legitimate aim of protecting the rights of others.

III. Necessity and proportionality

The principle of necessity and proportionality provides that 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; [and] they must be proportionate to the interest to be protected" (General Comment 34, para. 34). The Board finds that it was not necessary to remove the content in these three cases. When combined with systemic failures to apply relevant exceptions, Meta's internal list of slurs can amount to a near-absolute ban, raising both necessity and proportionality concerns in the context of journalistic reporting.

In relation to necessity, the inclusion of "İngiliz uşağı" on the slurs list was not necessary to protect people from hate speech because it is not used to attack persons on the basis of a protected characteristic. Meta's slurs list also appears to include terms that do not meet the company's own definition of slurs, prior to or following the May 25 policy revisions. The Board has been given full access to slurs lists last updated in the first quarter of 2023, and there are many terms listed, across markets, that are questionable in terms of whether they are hate speech or would be better understood as offensive insults that are not discriminatory in nature. Some board members have also expressed concern that the list is under-inclusive of many hate speech terms one would expect to see on such lists but are not there; whereas for some markets or languages the list of designated terms run for several pages, for other markets the lists are much shorter.

At the time of Meta's original decisions in these cases, Meta's prior definition of slurs, which then hinged on the concept of offensiveness, was overbroad and led to disproportionate restrictions on expression when three media organizations reported on events of political importance involving slur usage by public figures. Even if the phrase had been properly designated as a slur, when reporting about events that included its use by third parties in ways that would not incite violence or discrimination, the content should have been qualified as permissible "reporting." Meta's undisclosed policy guidance on how the reporting of slurs must be accompanied with additional context to be considered "awareness raising" interfered with each of the news outlets' editorial discretion and attempts to inform the Turkish public. The media entities in these three cases shared the video without the additional context that would indicate an intent to condemn or raise awareness (see above for the Board's analysis of Meta's exceptions in section 8.1).

In the "Mention of the Taliban in news reporting" decision, the Board examined the challenges of requiring clear user intent "even where contextual clues make clear the post is, in fact, reporting". While that case concerned the Dangerous Organizations and Individuals policy (where there is a public exception for reporting on designated entities), the observations on intent there apply to these cases on hate speech too. It is often considered good practice in journalism to report facts neutrally or impartially, without value judgment, a practice that is in tension with Meta's qualifications for reporting requiring clear intent to condemn or raise awareness. These cases bring an additional facet to that critique. While Meta's "raising awareness" exception addresses *reporting on slur usage*, that narrow application is underinclusive of circumstances, such as those in these cases, in which slur use was largely incidental to the main topic being reported. In these cases, removing the three posts was an unnecessary and disproportionate restriction on the freedom of expression of the rights of the individuals in the media and on the access to information rights of their audience.

The Board is concerned about Meta mechanically enforcing its hate speech policy on slurs and failing to account for when a public figure is present and the target of criticism. The Human Rights Committee has observed that public officials are "legitimately subject to criticism and political opposition" (General Comment 34, para. 38). The Board has raised this concern before in its "Colombia protests" decision. In that case, the Board said context should be carefully considered, not only the political context where a slur is used, but also if a

slur is used as part of criticism of political leaders. The Board's "Iran protests slogan" decision addressed hypothetical threats against political leaders, emphasizing the importance of protecting rhetorical political speech while also ensuring all people, including public figures, are protected from credible threats. Criticism of public figures can take a variety of forms, even forms that include offensive language, but Meta's current enforcement approach does not give the space necessary to thoughtfully balance these competing factors under either the undisclosed rules for reporting, or the parallel and more generally applicable newsworthiness allowance. A policy that can better accommodate news reporting would allow for more thoughtful assessment of context during at-scale review, without requiring escalation.

As the Board stressed above (Section 8.1: mistake prevention measures and escalation challenges), and in its "Colombia protests" decision, potentially newsworthy posts that merit closer contextual assessment appear not to be escalated to Meta's policy team as systematically or frequently as they should be. Whereas Meta presents the newsworthiness allowance as somewhat of a fail-safe for protecting public interest expression, Meta's own transparency reporting reveals the allowance was only applied 68 times in the year from June 2021 – May 2022. As the Board previously noted in its "Colombia protests" decision, the "newsworthiness exception should not be construed as a broad permission for hate speech to remain up." However, there needs to be stronger mechanisms to protect public interest expression, which can too easily be wrongly removed.

In two of the cases, Meta's strikes and penalty systems compounded necessity and proportionality concerns, with the wrongful removals resulting in further limitations on user expression and media freedom. These measures made it more difficult for both media organizations to freely share their reporting for the duration of those feature limits. Because of the chilling effect of likely future, even more grave sanctions, this had a real impact at a time when the earthquakes and pre-electoral period made access to independent local news particularly important.

The Board also encourages Meta to experiment with proactive in-house procedures to avoid false positives and less intrusive means of regulating the use of slurs, besides the removal of content that can result in strikes and feature limits. Given that freedom of expression, reflected in Meta's paramount value of "voice," is the rule and Meta's prohibition on slurs the exception, Meta's internal guidance to moderators should establish a presumption that

journalistic reporting (including citizen journalism) should not be removed. While the Board emphasized in the Colombia Protests decision that the "newsworthiness exception should not be construed as a broad permission for hate speech to remain up," Meta's internal rules should encourage the full consideration of the specific circumstances, to ensure that public interest reporting, which is not hate speech, is not incorrectly removed. The Board also recalls its decision in the Wampum Belt case, in which it emphasized the importance of Meta assessing content as a whole, rather than making assessments based on isolated parts of the content.

In addition, revising user notifications to include behavior nudges, for example to inform users when their posts appear to contain prohibited slurs, and inviting them to edit their posts, may increase compliance with the company's policies. Additional resources for media organizations are needed to understand how they should report on stories that include slur usage in ways that will not lead to content removal. Advice to users on how to edit broadcast video to obscure slur usage while still allowing current events to be reported on may also reduce the number of media organizations that find their accounts restricted as a result of reporting on public interest issues.

9. Oversight Board decision

The Oversight Board overturns Meta's original decisions to take down the content in each of these three cases.

10. Recommendations

Content policy

1. To ensure media organizations can more freely report on topics of public interest, Meta should revise the Hate Speech Community Standard to explicitly protect journalistic reporting on slurs, when such reporting, in particular in electoral contexts, does not create an atmosphere of exclusion and/or intimidation. This exception should be made public, and be separate from the "raising awareness" and "condemning" exceptions. There should be appropriate training to moderators,

especially outside of English languages, to ensure respect for journalism, including local media. The reporting exception should make clear to users, in particular those in the media, how such content should be contextualized, and internal guidance for reviewers should be consistent with this.

The Board will consider this recommendation implemented when the Community Standards are updated, and internal guidelines for Meta's human reviewers are updated to reflect these changes.

2. To ensure greater clarity of when slur use is permitted, Meta should ensure the Hate Speech Community Standard has clearer explanations of each exception with illustrative examples. Situational examples can be provided in the abstract, to avoid repeating hate speech terms.

   The Board will consider this implemented when Meta restructures its Hate Speech Community Standard and adds illustrative examples.

Enforcement

3. To ensure fewer errors in the enforcement of its Hate Speech policy, Meta should expedite audits of its slur lists in countries with elections in the second half of 2023 and early 2024, with the goal of identifying and removing terms mistakenly added to the company's slur lists.

   The Board will consider this implemented when Meta provides an updated list of designated slurs following the audit, and a list of terms de-designated, per market, following the new audits.

***Procedural note:**

The Oversight Board's decisions 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 case decision, 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 over 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.