Cited Recommendations & Implementation Status Annex

Referring to Designated Dangerous Individuals as "Shaheed"
Policy Advisory Opinion

Last updated February 22, 2024

1. Clarity of Policy and Narrowing its Scope

Case	Recommendation	Meta's Initial Response	Further Updates
Mention of the Taliban in News Reporting	No. 3 - Meta should **narrow the definition of "praise" in the Known Questions guidance for reviewers**, by removing the example of content that "seeks to make others think more positively about" a designated entity by attributing to them positive values or endorsing their actions.	"We are reviewing our definition of 'praise' in the Dangerous Organizations and Individuals (DOI) policy through an in-depth policy development process."	Q2 2023: "Due in part to recommendations from the Oversight Board, we conducted a thorough policy development to consider changes to our approach to "Praise" under our Dangerous Organizations and Individuals policy. We introduced changes to our Community Standards as a result of this policy development earlier this month that include clarifications to our approach to 'news reporting' which includes key examples to illustrate what content would be allowed in this context. We have also updated our Community Standards with examples and clarifications on what we consider neutral discussion of a DOI. Finally, we've included an update that clarifies what we consider to be 'condemnation' and have also included examples for this type of allowable content under our policy. We now consider this recommendation complete and will have no further updates."
Mention of the Taliban	No. 4 - Meta should **revise its Internal Implementation Standards** to make clear that the "reporting" allowance	"We are working to clarify our internal guidance on the news reporting	Q2 2023: "We have updated our guidance to add greater clarity to what constitutes 'reporting' under our Dangerous Organizations and Individuals policy. This includes examples

in News Reporting	in the Dangerous Individuals Organizations policy allows for positive statements about designated entities as part of the reporting, and how to distinguish this from prohibited "praise."	allowance under the Dangerous Organizations and Individuals policy and our definition of 'praise'."	and signals to illustrate the types of reporting context that we allow. We now consider this recommendation complete and will have no further updates."
Shared Al Jazeera Post	No. 1 - Add **criteria and illustrative examples to its Dangerous Individuals and Organizations policy** to increase understanding of the exceptions for neutral discussion, condemnation, and news reporting.	"We will add examples and language to the existing Dangerous Individuals and Organizations internal policy guidance to help clarify enforcement surrounding neutral discussion, condemnation, and news reporting in this policy area. We are also exploring ways of providing users clearer guidance on non-violating content."	Q2 2023: "Due in part to recommendations from the Oversight Board, we conducted a thorough policy development to consider changes to our approach to 'Praise' under our Dangerous Organizations and Individuals policy. We introduced changes to our Community Standards as a result of this policy development earlier this month that include clarifications to our approach to 'news reporting,' which includes key examples to illustrate what content would be allowed in this context. We have also updated our Community Standards with examples and clarifications on what we consider neutral discussion of a DOI. Finally, we've included an update that clarifies what we consider to be 'condemnation' and have also included examples for this type of allowable content under our policy. We now consider this recommendation complete and will have no further updates." Q1 2024: The December 29, 2023, edits to Meta's DOI policy which removed examples for exceptions to neutral discussion, condemnation and news reporting were added back to the DOI policy.
Öcalan's Isolation	No. 4 - Reflect in the Dangerous Organizations and Individuals "policy rationale" that **respect for human rights and freedom of expression**, in	"We will update the policy rationale of the Dangerous Organizations and Individuals section of	Q4 2021: "In December, we updated our Dangerous Organizations and Individuals policy language to clarify that we allow discussion about the human rights of designated individuals or members of designated dangerous entities

	particular open discussion about human rights violations and abuses that relate to terrorism and efforts to counter terrorism, can advance the value of safety, and that it is important for the platform to provide a space for these discussions. While safety and voice may sometimes be in tension, the policy rationale should specify in greater detail the "real-world harms" the policy seeks to prevent and disrupt when voice is suppressed.	the Community Standards with new language that makes it clear that discussion of human-rights violations and abuse, as they relate to dangerous organizations and individuals, is not a violation of our policies."	when that content does not include other praise, substantive support, or representation of designated entities or other policy violations. In that update, we also included a link for users to review our Corporate Human Rights Policy to learn more about our commitments to internationally recognized human rights. There will be no further updates on this recommendation."
Öcalan's Isolation	No. 6 - Explain in the Community Standards how users can make the **intent** behind their posts clear to Facebook and provide illustrative examples to demonstrate the line between permitted and prohibited content, including in relation to the application of the rule clarifying what "support" excludes.	"We are still assessing the trade-offs of additional transparency around our Dangerous Organizations and Individuals designations. Additionally, in response to recommendation no. 5 above, we are providing content reviewers with detailed definitions and examples of what 'support' means, outlining what type of content to leave up."	Q4 2021: "In response to Support of Abdullah Öcalan, Founder of the PKK Recommendation no. 5, illustrative examples have now been provided to reviewers to clarify the line of 'support.' We explained in our previous Quarterly Update that, because of the potential safety risks to our teams and tactical challenges to our ability to stay ahead of adversarial shifts, we determined not to publish any additional detail about the designations in this policy area. There will be no further updates on this recommendation."
Nazi Quote	No. 2 - Explain and provide examples of the application of key terms used in the Dangerous Organizations and	"Our commitment: We commit to adding language to the	Q1 2021: "We added definitions of the key terms used in the Dangerous Organizations and Individuals policy to the Community Standards. For example, we have included

	Individuals policy, including the meanings of **“praise,” “support” and “representation.”** These should align with the definitions used in Facebook’s Internal Implementation Standards. The Community Standard should provide clearer guidance to users on how to make their intent apparent when discussing individuals or organizations designated as dangerous.	Dangerous Organizations and Individuals Community Standard clearly explaining our intent requirements for this policy. We also commit to increasing transparency around definitions of ‘praise,’ ‘support’ and ‘representation’.”	definitions of ‘praise,’ ‘substantive support,’ and ‘representation’ and examples of how we apply these key terms. In addition, we created three tiers of content enforcement for different designations of severity. Tier 1, which includes terrorist, hate, and criminal organizations, results in the most extensive enforcement because we believe these entities have the most direct ties to offline harm. We also explain that our policy is designed to allow for users who clearly indicate their intent to report on, condemn, or neutrally discuss the activities of dangerous organizations and individuals.” Q1 2024: In June 2021, Meta provided information in its DOI policy addressing user intent: “Our policies are designed to allow room for these types of discussions, but we require people to clearly indicate their intent.” This was deleted in Meta’s Nov 24, 2021, update of the DOI policy, without track changes. The examples of prohibited content are still in effect.

2. Meta's Strikes System

Case	Recommendation	Meta's Initial Response	Further Updates
Former President Trump's Suspension	No. 15 - Facebook should **explain** in its Community Standards and Guidelines its strikes and penalties process for restricting profiles, pages, groups and accounts on Facebook and Instagram in a clear, comprehensive and accessible manner. These policies should provide users with sufficient information to understand when strikes are imposed (including any applicable exceptions or allowances) and how penalties are calculated.	"[W]e are publishing detailed information in our Transparency Center about our strikes and penalties. Our goal is to provide people with more information about our process for restricting profiles, pages, groups and accounts on Meta and Instagram."	Q3 2021: "In our 30-day response, we explained that we published information about our strikes system in our Transparency Center. In the post, we explain how we impose strikes and how we calculate penalties so people can better understand our processes. In providing this additional transparency, we want our users to better understand the details of our strikes and penalties processes while avoiding including certain information that malicious actors could use to circumvent our enforcement systems."
Mention of the Taliban in News Reporting	No. 2 - Meta should make its public explanation of its two-track strikes system more comprehensive and accessible, especially for "severe strikes."	"We are reviewing our strikes system to make it more comprehensive, effective and accessible."	Q4 2022: "We are constantly evaluating and pursuing work to improve our systems and policies for addressing violating content, and today announced as part of this work that we are updating our strike system. More public information about the strikes system can be found in our Transparency Center, and we will consider this recommendation complete. We will have no further updates on this recommendation."
Former President Trump's Suspension	No. 16 - Facebook should also provide users with **accessible information** on how many violations, strikes and penalties have been assessed against them, as well as the consequences that will follow future violations.	"Earlier this year, we launched 'Account Status' on Facebook, an in-product experience to help every user understand the penalties Facebook applied to their accounts."	Q3 2021: "In our 30-day response to this recommendation, we explained that earlier this year we launched 'Account Status' on the Facebook app, an in-product experience to help people understand the penalties Facebook applied to their accounts. It provides information about the penalties on a person's account (currently active penalties as well as past penalties), including why we applied the penalty. In general, if people have a restriction on their account, they

			can see their history of certain violations, warnings, and restrictions their account might have, as well as how long this information will stay in Account Status on Facebook. We have also launched Account Status on Instagram. There will be no further updates on this recommendation.”

3. Transparency

Case	Recommendation	Meta's Initial Response	Further Updates
Nazi Quote	No. 3 - **Provide a public list of the organizations and individuals** designated "dangerous" under the Dangerous Organizations and Individuals Community Standard. At a minimum, illustrative examples should be provided. This would help users to better understand the policy and conduct themselves accordingly.	"We commit to increasing transparency around our Dangerous Organizations and Individuals Policy. In the short term, we will update the Community Standard and link to all of our Newsroom content related to Dangerous Organizations and Individuals so that people can access it with one click."	Q3 2021: "In our previous update on this recommendation, we explained that sharing this information could present safety risks to our teams and pose a tactical challenge to our ability to stay ahead of adversarial shifts. Since then, we assessed how we could balance increased transparency about the individuals and organizations we designate under this policy on the one hand with the safety of our community and employees on the other. After careful consideration, we have determined not to publish any additional detail about the designations in this policy area at this time."
Former President Trump's Suspension	No. 17 - In its transparency reporting, **Facebook should include numbers of profile, page and account restrictions, including the reason and manner in which enforcement action was taken**, with information broken down by region and country.	"We agree that sharing more information about enforcement actions would be beneficial and are assessing how best to do so in a way that is consistent and comprehensive."	Q3 2023: "We are currently working on two long-term initiatives prompted by this recommendation: measuring our enforcement actions on profile, page, and account restrictions; and measuring enforcement data by location. Both of these initiatives fit into our overall vision for the Community Standards Enforcement Report (CSER). Next Expected Update: Q4 2023, publishing in Q1 2024"
Öcalan's Isolation	No. 12 - In transparency reporting, **include more comprehensive information on error rates for enforcing rules on "praise" and "support" of dangerous**	"We are continuing to assess the feasibility of measuring and reporting enforcement and error rate data by country. In addition, we will assess	Q4 2021: "Since our previous update, we have determined that we will not include enforcement data reports at the level of granularity this recommendation outlines. We are instead prioritizing the work that will enable broader, report-level changes, such as publishing enforcement data on complex objects and by location (see update to Former

	organizations and individuals, broken down by region and language.	whether we can capture data at the more granular violation type – such as 'praise' and 'support' – as a subset of the Dangerous Organizations and Individuals Community Standard."	President Trump's Suspension recommendation no. 18, above). There will be no further updates on this recommendation."
Punjabi Concern Over the RSS in India	No. 3 – **Facebook should improve its transparency reporting to increase public information on error rates by making this information viewable by country and language for each Community Standard**. The Board underscores that more detailed transparency reports will help the public spot where errors are more common, including potential specific impacts on minority groups, and alert Facebook to correct them.	"We're continuing to identify appropriate accuracy metrics to include in the Community Standards Enforcement Report, and are assessing how to report consistent, comprehensive data."	Q3 2023: "We are conducting long-term work to define our accuracy metrics, alongside our work on Breast Cancer Symptoms & Nudity recommendation no. 6. As we continue to develop the necessary measurement infrastructure and data validation protocols to report high-quality, consistent information, we are continuing to engage with the board on our more incremental roadmaps, challenges, and expansion opportunities. Next Expected Update: Q4 2024"

4. Automation

Case	Recommendation	Meta's Initial Response	Further Updates
Breast Cancer Symptoms and Nudity	No. 6 - **Expand transparency reporting to disclose data on the number of automated removal decisions per Community Standard**, and the proportion of those decisions subsequently reversed following human review.	"We need more time to evaluate the right approach to share more about our automated enforcement. Our Community Standards Enforcement Report currently includes our 'proactive rate' (the amount of violating content we find before people report it), but we agree that we can add more information to show the accuracy of our automated review systems."	Q3 2023: "Our current focus for this work is on improving what we internally call 'data readiness,' by aligning on a consistent accounting methodology across metrics. We are working to define binaries for each metric as a first step towards aggregating public-facing enforcement metrics. To do this, we are discussing complexities such as how to quantify instances of enforcement conducted by human review and automated tools (e.g., quantifying cases where a human reviewer determined that an image was violating and then a machine scaled that decision more broadly). Concurrently, we are resolving gaps in our logging infrastructure to allow us to pull those metrics once we've decided on how to report it. Next Expected Update: Q4 2024"
Colombian Police Cartoon	No. 3 - **Meta should publish the error rates for content mistakenly included in Media Matching Service banks of violating content, broken down by each content policy, in its transparency reporting.** This reporting should include information on how content enters the banks and the company's efforts to reduce errors in the process.	"While we are committed to sharing more information on our enforcement accuracy as part of previous recommendations, providing this information in the manner prescribed here would not provide a holistic and accurate picture of our content	Q3 2022: "In our initial response to this recommendation, we highlighted our commitment to gathering and sharing accuracy and precision metrics around our content moderation tools, including our Media Matching Systems. This remains a high priority for our Community Standards Enforcement Reporting (CSER) and transparency efforts. However, upon consideration, we determined that sharing metrics for individual MMS banks without sufficient context would likely create confusion and would not provide a complete picture of the accuracy of our automated enforcement actions. This is because MMS is

		moderation systems. We continue to work towards public reporting of new metrics that will provide comprehensive insights into our enforcement systems, including efforts to reduce errors in the process, but we will have no further updates on this recommendation."	one component of a larger enforcement system, and media matching does not always work in isolation. While we continue to work towards public reporting of new metrics that will provide comprehensive insights into our enforcement systems, we will have no further updates on this recommendation."
Iran Protest Slogan	No. 7 - **Meta should provide a public explanation of the automatic prioritization and closure of appeals,** including the criteria for both prioritization and closure.	"Our progress in automatic prioritization and closure of appeals is newly developed and quickly transforming. Given the nature of this work, we believe that providing ongoing updates of our implementation effort will suffice as the criteria involved are evolving. Building, testing and strengthening automatic prioritization and closure of appeals remains our priority, and we will continue to report on the implementation progress as the criteria matures."	Q1 2023: "As explained in our initial response, our current content review prioritization process across all of our products is publicly outlined in our Transparency Center. On this page, we explain that we primarily consider severity, virality, and likelihood of violation when determining which content our human review teams should prioritize for review. Since Q1 2022, we have undergone a multi-stage process to identify key drivers of trust in appeals to improve their overall effectiveness. Given this, our more granular progress in automatic prioritization and closure of appeals is newly developed and quickly transforming. When considering whether to add additional granularity to our Transparency Center page reflecting these changes, we came to the conclusion that the publication of the new system at this stage would be misleading given the fact that the criteria involved are subject to evolve, often very quickly. In the spirit of transparency, we will be sharing our full assessment of considerations with the Oversight Board. While we will have no further updates on this specific

			recommendation, we are continuing to refine our automatic prioritization and ranking of appeals throughout 2023, and will be providing further updates on the development of these new processes in our responses to Asking for Adderall® recommendation no. 2 and “Two Buttons” Meme recommendation no. 5. By continuing to publicly report on each iteration of building, testing, and strengthening automatic prioritization and closure of appeals in our Quarterly Updates, we hope to achieve the spirit of this recommendation by furthering transparency around the process without risking an inaccurate update to the Transparency Center.”