Policy advisory opinion 2022-01, Removal of COVID-19 misinformation

April 2023

I. Executive summary

In July 2022, the Oversight Board accepted a request from Meta to assess whether it should continue to remove certain categories of COVID-19 misinformation, or whether a less restrictive approach would better align with its values and human rights responsibilities. This policy advisory opinion is the Board's response to that request.

The Board has conducted extensive investigation and public consultation. Given Meta's insistence that it takes a single, global approach to COVID-19 misinformation, the Board concludes that, as long as the World Health Organization (WHO) continues to declare COVID-19 an international public health emergency, Meta should maintain its current policy. That means it should continue to remove COVID-19 misinformation that is likely to directly contribute to the risk of imminent and significant physical harm. However, the Board finds that Meta should begin a process to reassess each of the 80 claims it currently removes, engaging a broader set of stakeholders. It should also prepare measures for when the WHO declaration is lifted, to protect freedom of expression and other human rights in these new circumstances. The Board strongly recommends that Meta publish information on government requests to remove COVID-19 content, take action to support independent research of its platforms, examine the link between its platforms' architecture and misinformation, and promote understanding around COVID-19 misinformation globally.

Background

In early 2020, as the COVID-19 pandemic took hold, Meta began removing several claims from Facebook and Instagram that the company identified as COVID-19 misinformation. The list of COVID-19-related claims that the company removes has developed over the course of the pandemic. Today, Meta removes about 80 distinct COVID-19 misinformation claims under its "Misinformation about health during public health emergencies" policy, a subsection of the Misinformation Community Standard created in response to the Board's recommendations in the "Claimed COVID cure" decision. This policy advisory opinion focuses exclusively on Meta's actions during the COVID-19 pandemic under the "Misinformation about health during public health emergencies" policy. It does not address actions Meta has taken during the COVID-19 pandemic under other policies.

Under the "Misinformation about health during public health emergencies" policy, Meta removes "misinformation during public health emergencies when public health authorities conclude that the information is false and likely to directly contribute to the risk of imminent

physical harm." Meta relied exclusively on public health authorities to determine whether that standard had been met. The 80 claims it currently removes include, for example, denying the existence of COVID-19 and asserting that COVID-19 vaccines cause magnetism. Meta removed 27 million pieces of COVID-19 misinformation from Facebook and Instagram between March 2020 and July 2022, 1.3 million of which were restored through appeal. COVID-19 misinformation that does not meet the standard for removal can be fact-checked, labeled or demoted. Fact-checkers rate content (for example, as "false," or "missing context"). Meta then labels it as such and links to a fact-checker article on the subject. The company also demotes content labeled by fact-checkers, meaning it appears less frequently and prominently in users' feeds, depending on a number of factors. Meta also applies "neutral labels" to COVID-19-related content. These labels contain statements such as, "some unapproved COVID-19 treatments may cause serious harm," and direct people to Meta's COVID-19 information center, which provides information on prevention measures, vaccines and resources from public health authorities.

In its request to the Board, Meta asked whether it should continue to remove certain COVID-19 misinformation. Alternatively, the company said it could stop removing it and instead demote content, send it to third-party fact-checkers, or label it. Meta insists on taking a single, global approach to COVID-19 misinformation, rather than varying its approach by country or region. According to the company, taking a localized approach to the policy at scale would lead to lack of clarity for users and poor enforcement, and it lacks the capacity to adopt such an approach. In considering the request, the Board held extensive public consultations. These included a series of virtual roundtables with participants from around the world, convened in partnership with civil society, through which the Board heard from a wide range of experts and stakeholders.

Key findings and recommendations

The Board finds that continuing to remove COVID-19 misinformation that is "likely to directly contribute to the risk of imminent physical harm" during a global public health emergency is consistent with Meta's values and human rights responsibilities. The Board initially explored whether it would be preferable for Meta to take a localized approach to COVID-19 misinformation at scale. However, Meta insisted that this was not feasible without significantly undermining clarity and fairness for users and significantly increasing errors in enforcing its policy. Meta's concerns may be warranted. However, by ruling out this option, Meta has frustrated the Board's efforts to reconcile competing viewpoints from stakeholders and Board Members on how to best address harmful COVID-19 misinformation, while respecting human rights, especially the right to freedom of expression. The 18 recommendations in this policy advisory opinion, most of which are summarized below, work within this constraint.

The Board recommends that Meta:

Continue to remove false content about COVID-19 that is "likely to directly contribute to the risk of imminent physical harm during the ongoing global public health emergency," while beginning a transparent and inclusive review and reassessment of the 80 claims it currently removes. A public health emergency presents a serious and direct danger to health. Given Meta's insistence on a single, global approach to COVID-19 misinformation, the Board finds Meta is justified in responding with its current exceptional measures of removing false information likely to directly contribute to the risk of imminent physical harm, as determined by public health authorities. Meta has not, among other things, returned to the relevant public health authorities to ask them to re-evaluate the claims it removes. Nor has it conducted broader stakeholder and expert consultations to re-evaluate the individual claims or the overall policy. As Meta has not yet engaged in a due diligence process to change its policy (which is Meta's responsibility in the first instance), the Board is not in a position to recommend a policy change that could disproportionately affect the most vulnerable.

However, now that we are no longer in the early stages of the crisis, to meet its human rights responsibilities, Meta should regularly assess whether the threshold for removal, set out in its policies, continues to be met. It should therefore begin a transparent process to regularly review the 80 claims subject to removal, consulting with a broad range of stakeholders. Only when stakeholders provide clear evidence of the potential of a claim to cause imminent physical harm is it justifiable to include it on the list of claims subject to removal. Meta should share with the public the outcomes of these periodic reviews.

Explore localizing its approach. Meta needs to plan what to do when the WHO stops classifying COVID-19 a global health emergency, but local public health authorities continue to designate it a public health emergency. The Board recommends initiating a risk assessment process to identify the measures it will take in this scenario. These should address misinformation likely to directly contribute to significant and imminent real-life harm, without compromising freedom of expression globally. The risk assessment should include assessing whether it is feasible to localize enforcement of its policies.

Assess the impact of its platforms' architecture. Experts raised concerns that the architecture of Meta's platforms amplifies harmful health misinformation. Given these claims, the Board recommends that Meta assess the human rights impact of its design choices. The company should commission a human rights impact assessment on how its newsfeed, recommendation algorithms, and other features amplify harmful health misinformation and its impacts.

Increase transparency around government requests. At the height of the pandemic, concerns were raised about Meta reviewing COVID-19-related content at the behest of

governments. This is particularly problematic where governments make requests to crack down on peaceful protesters or human rights defenders, to control conversations about the origins of the pandemic, and to silence those criticizing or questioning government responses to the public health crisis. The United Nations has raised concerns that some governments have used the pandemic as a pretext to erode the tenets of democracy. Meta should be transparent and report regularly on state actor requests to review content under the "Misinformation about health during public health emergencies" policy.

Support independent research and promote understanding of COVID-19 misinformation. The Board heard from experts that wider efforts to understand COVID-19 misinformation, and the effectiveness of Meta's response to it, are frustrated by lack of access to the company's data and research. A lack of data has also created challenges for the Board when assessing the merits of this policy advisory opinion request. The Board recognizes that in comparison with other social media companies, Meta has taken significant steps to share data with external researchers, many of whom told the Board of the importance of Meta tools such as CrowdTangle, and Facebook Open Research and Transparency (FORT). At the same time, researchers have also complained about the difficulty of accessing tools such as FORT. Meta should continue to make these tools available, while improving their accessibility, and allow external researchers to access data that is not public. The Board also recommends that the company conduct research and publish data on its efforts to enforce its COVID-19 policies, and that it publish the findings of the "neutral labels" research shared with the Board. Finally, it recommends that Meta take steps to expand access to the company's data to Global Majority, also referred to as the Global South, researchers and universities, and that it support digital literacy programs across the world.

Increase fairness, clarity and consistency around the removal of COVID-19 misinformation. To meet its human rights responsibilities, Meta must also make sure its rules are clear to users. To this end, the company should explain how each category of the COVID-19 claims it removes directly contributes to the risk of imminent physical harm. It should also explain the basis of its assessment that a claim is false and create a record of any changes made to the list of claims it removes. To support the consistent application of its rules across languages and regions, the company should translate its internal guidance for content moderators into the languages in which it operates. Meta should also protect users' right to a remedy by expanding users' ability to appeal fact-checker labels, and by ensuring such appeals are not reviewed by the fact-checker who made the initial decision.

II. Request from Meta

1. This policy advisory opinion request concerns Meta's policy on COVID-19 misinformation, as outlined in the company's policy on ***Misinformation about health during public health emergencies*.** This policy is part of the Misinformation Community

Standard. Meta sent the Board a request for this policy advisory opinion in July 2022, asking the Board whether it should continue removing certain content on COVID-19 under this policy or whether a less restrictive approach would better align with the company's values and human rights responsibilities.

2. Meta explained in its request that under the ***Misinformation about health during public health emergencies*** policy, it removes misinformation where it is "likely to directly contribute to the risk of imminent physical harm" during a public health emergency. This policy is explained in more detail in Section III, below. In determining whether misinformation meets this standard, the company partners with public health authorities with knowledge and expertise to assess the falsity of specific claims and whether they are likely to directly contribute to the risk of imminent physical harm. This assessment is made at the global level. Meta made clear to the Board that a scaled, localized approach to enforcing the policy was not feasible without significantly undermining clarity and fairness for users, and significantly increasing errors in enforcing the policy. For misinformation that does not meet this high threshold, Meta uses other measures such as demotions, labeling or fact-checking.

3. In developing its approach to COVID-19 misinformation from January 2020 to February 2021 (see Section III, below), Meta consulted more than 180 experts from different disciplines, including public health and infectious diseases experts, experts in national security and public safety, dis- and misinformation researchers, fact-checkers, and freedom of expression and digital rights experts. Meta also consulted with regional stakeholders. Meta noted differing and at times conflicting recommendations and perspectives from experts from different disciplines and stakeholders from different regions of the world, as explained further below. Meta stated that in adopting the current policy, it determined that risks associated with certain misinformation were significant enough to make removal necessary given the global public health emergency created by the COVID-19 pandemic. In its request to the Board, Meta explained that the landscape surrounding COVID-19 has changed since Meta decided to remove certain COVID-19 misinformation over two years ago. Those changes have prompted the company to consider whether removing such misinformation is still necessary. Meta noted three changes for the Board's consideration.

4. First, according to Meta, there is a change in the COVID-19 information ecosystem. At the beginning of the pandemic, the lack of authoritative guidance created an information vacuum that encouraged the spread of misinformation. Today, the company noted, much has changed. People have access to reliable information about the virus and public health authorities that can inform and influence the behavior of those at risk.

5. Second, Meta noted that due to vaccines and the evolution of disease variants, COVID-19 is less deadly than it was in early 2020. With the development and distribution of

effective vaccines, there is now a means to prevent and reduce the severity of COVID-19 symptoms. Additionally, according to public health experts, the current variants cause less severe disease than prior variants, and therapeutic treatments are evolving rapidly.

6. Third, Meta stated that public health authorities are actively evaluating whether COVID-19 has evolved to a less severe state and "some public health authorities are noting that certain regions of the world have begun transitioning to a less severe state of the pandemic." (Meta policy advisory opinion request, page 14) Meta cited statements made by Dr. Anthony Fauci in the United States, the European Commission, and the Public Health Minister of Thailand, in support of this observation.

7. However, Meta acknowledged that the course of the pandemic has been different across the globe. In its request to the Board, the company wrote:

> *While vaccines, medical treatment, and authoritative guidance are increasingly available in high-income countries, experts predict that access will lag for people in low-income countries with less developed healthcare systems. The most significant variation right now is between developed nations, (…) and less developed nations. (…) Eighty percent of people in high-income countries have received at least one dose of the vaccine, as opposed to only 13 percent of people in low-income countries. Low-income countries are also more likely to have health care systems with less capacity, less robust economies, and lower trust in government guidance, all of which will add challenges to vaccinating people and treating those that contract COVID-19.*
>
> (Meta policy advisory opinion request, page 3)

Meta's questions to the Board

8. In view of the above, Meta presented the following policy options to the Board for its consideration. Section III, below, provides more detail on these measures:

   i. Continue removing certain COVID-19 misinformation: This option would mean continuing with Meta's current approach of removing false content "that is likely to directly contribute to the risk of imminent physical harm during a public health emergency." Meta states that under this option the company would eventually stop removing misinformation when it no longer poses a risk of imminent physical harm and requests the Board's guidance on how the company should make this determination.

   ii. Temporary emergency reduction measures: Under this option, Meta would stop removing COVID-19 misinformation and instead reduce its distribution. This would be a temporary measure, and the company requests the Board's guidance as to when it should stop using it if adopted.

iii. Third-party fact-checking: Under this option, content currently subject to removal would be sent to independent third-party fact-checkers for evaluation. In its request to the Board, Meta notes that "the number of fact-checkers available to rate content will always be limited. If Meta were to implement this option, fact-checkers would not be able to look at all COVID-19 content on our platforms, and some of it would not be checked for accuracy, demoted, and labeled." (Meta policy advisory opinion request, page 16)

iv. Labels: Under this option, Meta would add labels to content which would not obstruct users from seeing the content but would provide direct links to authoritative information. Meta considers this a temporary measure and seeks the Board's guidance on what factors the company should consider in deciding to stop using these labels.

9. Meta stated that each of these options has advantages and disadvantages, particularly in terms of scalability, accuracy, and in terms of the amount of content affected. The company strongly urged that the policy should be appropriate for all regions, while being consistent and workable globally. Regarding country-specific policies, Meta stated that: "For technical reasons, we strongly recommend maintaining global policies regarding COVID-19, as opposed to country or region-specific policies." In its request to the Board, Meta said: "Enforcing policies at the country level can lead to both over-enforcement when one set of market reviewers covers multiple countries, and under-enforcement because content can spread across countries and regions." (Meta policy advisory opinion request, page 17, footnote 37.) (For more on scaled, localized enforcement, see paragraph 61 below.)

III. Meta's policy on *Misinformation about health during public health emergencies*

10. Prompted by the COVID-19 pandemic and a surge of misinformation observed on its platforms, Meta began removing COVID-19 misinformation in January of 2020. Its policy, and the types of claims subject to removal, evolved over the following two-years and culminated in the current section on ***Misinformation about health during public health emergencies*** of the Misinformation Community Standard, and the accompanying Help Center page.

11. The Misinformation Community Standard begins with a policy rationale that explains Meta's approach to misinformation, including the use of removals, fact-checks, demotion and labels as enforcement measures. According to the company, it only removes misinformation when "it is likely to directly contribute to the risk of imminent physical harm." The policy then identifies four types of misinformation subject to removal: (1) physical harm or violence; (2) ***Harmful Health Misinformation;*** (3) voter or census interference; and (4) manipulated media (emphasis added).

12. The *Harmful Health Misinformation* section of the Misinformation Community Standard has three sub-categories: (i) misinformation about vaccines; (ii) ***misinformation about health during public health emergencies;*** and (iii) promoting or advocating for harmful miracle cures for health issues (emphasis added). While other community standards potentially apply to COVID-19 claims, this policy advisory opinion focuses exclusively on the ***Misinformation about health during public health emergencies*** section of the Misinformation policy, as it applies to the COVID-19 pandemic. It should not be read to address other policies.

Removal of COVID-19 misinformation claims

13. The policy on ***Misinformation about health during public health emergencies*** states:

> *We remove misinformation during public health emergencies when public health authorities conclude that the information is false and likely to directly contribute to the risk of imminent physical harm, including by contributing to the risk of individuals getting or spreading a harmful disease or refusing an associated vaccine. We identify public health emergencies in partnership with global and local health authorities. This currently includes false claims related to COVID-19 that are verified by expert health authorities, about the existence or severity of the virus, how to cure or prevent it, how the virus is transmitted or who is immune, and false claims which discourage good health practices related to COVID-19 (such as getting tested, social distancing, wearing a face mask, and getting a vaccine for COVID-19). Click here for a complete set of rules regarding what misinformation we do not allow regarding COVID-19 and vaccines.*

14. Under this policy, Meta removes misinformation when three elements are satisfied: 1) there is a public health emergency; 2) the claim is false; and 3) the claim is likely to directly contribute to the risk of imminent physical harm. Applying this standard to the COVID-19 pandemic, Meta relied on the conclusions of public health authorities to identify about 80 claims that are subject to removal. These claims, which are periodically updated, and common questions on how the policy is enforced, are available through this Help Center page.

15. As stated in the policy, these 80 claims are classified in the following five categories: (1) claims about the existence or severity of COVID-19, this includes claims that deny the existence of COVID-19 or undermine its severity, for example claims that no one has died from COVID-19; (2) COVID-19 transmission and immunity, which includes claims that COVID-19 is transmitted by 5G technologies; (3) guaranteed cures or prevention methods for COVID-19, such as those that affirm that topical creams can cure or prevent the infection of coronavirus; (4) discouraging good health practices, for example claims that face masks contain harmful nano worms or that the COVID-19 vaccines alter

people's DNA or cause magnetism; (5) access to essential health services, which includes claims that hospitals kill patients in order get more money, or in order to sell people's organs.

16. Meta explained in its request that it relies on the following to assess whether a public health emergency exists: (1) whether the World Health Organization (WHO) declared a public health emergency; (2) whether the WHO designated a disease as communicable, deadly or high risk; or (3) in the event a WHO risk assessment is unavailable, Meta defers to local health authorities' designation of a public health emergency for a given country. As Meta explained, the WHO advised the company that, during a declared emergency, "there is a high risk of irreversible physical harm to individuals when the risk of exposure, rate of transmission, association between exposure and risk, and morbidity and mortality rates are unusually high." (Meta policy advisory opinion request, page 6) The WHO declared the outbreak of COVID-19 a "Public Health Emergency of International Concern" on January 30, 2020.

17. According to its policy, in the context of public health emergencies, Meta relies on expert public health authorities to determine falsity, and whether a false claim is "likely to directly contribute to a risk of imminent physical harm". Meta informed the Board that it relies on public health experts, such as the WHO and the U.S. Centers for Disease Control and Prevention (CDC), in assessing the claims currently subject to removal. In the past, Meta has also consulted with, for example, country heads of UNICEF and the National Health Ministry of Pakistan (Meta policy advisory opinion request, page 6). Based on samples of correspondence between Meta and a public health authority that the Board has reviewed, Meta identifies claims, in part, through the monitoring of its platforms and submits them for public health bodies to assess, rather than public health authorities themselves identifying or defining the claims that should be removed. In an example of an exchange with one public health authority providing this assessment, Meta identified a claim (that the COVID-19 vaccines cause heart attacks) and asked the public health authority whether it is false and whether it could lead to vaccine refusal. The public health authority responded with sources and analysis in support of the conclusion that "there is no evidence that COVID-19 vaccines cause heart attacks." It noted reported cases of myocarditis (inflammation of the heart muscle) after vaccination, especially in male adolescents and young adults. The public health authority also concluded that "unfounded fear about vaccines causing heart attacks (myocardial infarction) or other heart conditions could lead to vaccine refusal."

18. Meta's misinformation policy includes exceptions, whereby the company exempts certain types of speech from removal. For example, statements focused on politics or political decisions, such as "COVID mandates don't work" or "vaccine companies just want to line their wallets," are permitted. Content shared with explicit humor and satire for example "Only Brad Pitt's blood can cure COVID [two laughing crying emojis]," do not violate the policy. Claims that express a personal anecdote or experience are

permitted where: the content shares a personal experience of a specific person; the identity of the person is explicitly mentioned in the post; and the content does not make a categorical claim or include a call to action. Content that is speculating, questioning, or expressing uncertainty, for instance, "Do vaccines cause autism?" is also permitted.

19. Meta relies on its regular at-scale enforcement system of automated tools (or classifiers), content moderators, and internal escalation teams to remove the claims listed under the misinformation policy. The company informed the Board that it has trained classifiers in 19 languages to identify likely violations of this policy. Content moderators enforcing the misinformation policy are provided with internal guidelines, including guidance on how to identify content that should remain on the platform as humor, satire, or a personal anecdote, as explained on the Help Center page. When a piece of content is removed, the user has the option to "disagree with the decision." Review is not guaranteed but some decisions are reviewed on appeal when there is capacity. Meta informed the Board that between March 2020 and July 2022, the company removed 27 million pieces of content for COVID-19 misinformation across Facebook and Instagram. Of the 27 million removals, 1.3 million were restored through the appeals it was able to assess.

Third-party fact-checking and demotion of misinformation

20. Content that does not fall under the list of claims subject to removal but that may constitute COVID-19 misinformation is subject to third party fact-checking. Machine learning technology detects posts that are likely to be misinformation and sends it to third party fact-checkers. Meta partners with independent fact-checking organizations to review and label content. The company does not evaluate fact-checker performance. It relies on the International Fact Checking Network (IFCN) to evaluate organizations and ensure quality. Fact-checking organizations are assessed by IFCN for compliance with a Code of Principles which includes a commitment to non-partisanship and fairness, transparency of sources, transparency on funding, and an open and honest corrections policy.

21. Content sent to the fact-checking queue may temporarily appear lower in users' feeds, especially if it is going viral, before it is reviewed. Meta uses a ranking algorithm to prioritize content for fact-checkers, and viral content is prioritized in the fact-checkers' queue. In response to the Board's question, Meta informed the Board that an overwhelming majority of content in the queue for fact-checking is never reviewed by fact-checkers. According to Meta, most of the content in the queue is not false, a claim for which Meta did not provide supporting evidence. According to Meta, fact-checking partners prioritize provably false claims or clear hoaxes that have no basis in fact and that are timely, trending, and consequential. Fact-checkers can also find content to review on their own initiative, outside the queue provided by Meta.

22. Posts and ads from politicians are not eligible for fact-checking. Meta defines "politicians" as "candidates running for office, current office holders—and by extension, many of their cabinet appointees—along with political parties and their leaders." According to Meta, this policy is in place because "political speech [in mature democracies with a free press] is the most scrutinized speech there is [and] limiting political speech [...] would leave people less informed about what their elected officials are saying and leave politicians less accountable for their words."

23. Fact-checkers can rate content as "false," "altered," "partially false" or "missing context." Meta will label the content accordingly and provide a link to the fact-checker's article on the topic. To read the article, the user must exit Facebook and go to another page, this requires the use of additional data. This means incurring additional costs for some users, in countries where Meta's platforms are zero-rated (meaning users do not pay data or other charges for mobile access to Meta's apps). Content labeled "false" and "altered" is covered by a warning screen that obscures the content, requiring the user to click through to see it. Content labeled "partly false" and "missing context" is not obscured. Meta told the Board that in the 90-day period leading up to December 9, 2022, for posts that were labeled "false" or "altered," and were covered by a screen obscuring the content, 10% of Facebook users and 43% of Instagram users uncovered the post to view the content." During the same period, Meta reported for posts that received a "false," "altered," "partly false" or "missing context" label, on average, 3% of Facebook users and 19% of Instagram users clicked on the "See Why" prompt. This takes the user to a separate screen that provides information on why the content was rated and links to the fact-checker's article.

24. When a piece of content is labeled by fact-checkers, Meta will rank it lower in user feeds. The demotion strength applied to COVID-19 misinformation labeled "false," "altered," and "partly false" is higher than that applied to content labeled "missing context." According to Meta, the effect of demotion on where a piece of content will appear on an individual user's feed will vary. The content a user sees is personalized to them with the aim, according to Meta, of showing the user content that is most likely to be of interest to them. A demotion of a piece of content would result in a reduction in its ranking score. It would not mean that the content would have a specific number or percentage of fewer views. How likely a particular user is to see a piece of content that has been demoted will depend on how high the ranking score is for that piece of content relative to the other pieces of content in the user's feed. As a result, demotions impact a piece of content in different ways depending on the user and their content inventory. When content is shared in a group or by a user with a large number of followers, its ranking score for those following and regularly engaging with the group or page will likely be higher than, or similar to, the ranking score of the other content in the feed. This means demotion may not have the intended effect.

25. Where a fact-checker determines that content created by a page or group violates the company's misinformation policy, page managers or group administrators can appeal. Facebook profiles do not have the option to appeal a fact-check label. Group violations can be appealed through the Facebook app on IOS and Android or on the web browser. Page violations cannot be appealed through a web browser. According to the Facebook Help Center, "this in-product appeals feature is only available to group admins and page managers in some countries right now. Anyone else can still send appeals to fact-checkers through email."

26. In its request to the Board, Meta explained that the number of fact-checkers available to rate content will always be limited. In asking the Board to evaluate the different approaches available to address COVID-19 misinformation, Meta explained that if the company were to rely exclusively or predominantly on fact-checking, fact-checkers would not be able to look at all COVID-19 content on the platforms. Some misinformation would therefore not be checked for accuracy, demoted, and labeled.

Applying labels

27. Meta also applies two types of what it calls, "neutral inform treatments" (NITs): neutral treatment labels; and Facts about X informed treatments, or FAXITs. These are applied directly by Meta and do not involve fact-checkers. FAXITs provide a tailored statement on the content shared in the post before directing the user to the information center. Meta has two FAXIT labels: (1) "COVID-19 vaccines go through many tests for safety and effectiveness and are then monitored closely"; and (2) "Some unapproved COVID-19 treatments may cause serious harm." A label is placed on any content a classifier identifies as discussing COVID-19. The content can be true or false and the label makes no statement about the content specifically. All NITs direct users to Meta's COVID-19 information center.

28. During the discussions concerning this policy advisory opinion request, Meta informed the Board that it would be scaling back its NITs effective December 19, 2022. The decision was based on a global experiment that Meta's Product and Integrity teams conducted on the effectiveness of NITs to limit COVID-19 misinformation prevalence on the platform. The test involved a control group that continued to see the COVID-19 NITs without any limitations. Three additional test groups were also set up: first, a group that could see one of each COVID-19 NIT every three days; second, a group that could see one of each COVID-19 NIT every 30 days; and third, a group that saw no labels at all. According to the company, the second group (the group that could see each one of the NITs every 30 days) had the highest average click-through rate to authoritative information among all groups, including the control group. The average time spent viewing the NITs was also the highest for this group. Moreover, there was "no statistically significant regression" in COVID-19 misinformation prevalence between the control group and the test groups. Based on the results of the experiment, Meta

informed the Board that it had capped the number of COVID-19 NITs that users could see on their platforms to one of each type of COVID-19 label every 30 days, beginning December 19, 2022. Thereafter, Meta informed the Board that the company has stopped using all COVID-19 NITs, so that users can be exposed to a reduced volume of labels. This is to ensure NITs are also effective in other public health crises.

Penalties

29. Meta applies account-level and group-level penalties that impact the spread of misinformation. A profile, page or a group that posts content that is removed or is labeled "false" or "altered" under this policy will receive a strike and be removed from recommendations, will not be able to monetize, and pop-ups will begin to appear to visitors to the page or group informing them that this page has shared misinformation, once a strike threshold is reached. According to the Help Center page, pages, groups and Instagram accounts may also be removed "if they have shared content that violates the COVID-19 and vaccines policies and are also dedicated to sharing other vaccine discouraging information on the platform."

IV. External engagement

30. Over the course of developing this policy advisory opinion, the Oversight Board engaged both with stakeholders and with Meta in several ways.

Public comments

31. The Oversight Board received 181 public comments in August 2022 related to this policy advisory opinion. Four of the comments were from Latin America and the Caribbean, five from Central and South Asia, eight from Asia Pacific and Oceania, 81 from Europe and 83 from United States and Canada. The Board did not receive any public comments from the Middle East and North Africa or from Sub-Saharan Africa.

32. The issues covered by the submissions included:

- A submission from the Khazanah Research Institute (PC-10703), a policy research institute in Malaysia, highlighted the differing levels of access to reliable health information in different countries and the diverse levels of risk from leaving misinformation unmoderated. If there must be a global approach, the submission recommends that Meta err on the side of caution and continue to remove harmful COVID-19 misinformation. The submission also commented on the lack of clear definition of "imminent physical harm" and the importance of understanding context, continuous monitoring, and transparency on enforcement to ensure the use of this standard can be effectively assessed.

- The American Civil Liberties Union's submission (PC-10759) raised the concern that the difficulty in distinguishing, at scale, between fact and fiction, and between opinion, experience and assertion of fact, means Meta will stifle speech that should be permitted.

- The submission from US non-profit organization Asian Americans Advancing Justice (PC-10751) noted the "scapegoating" of Asian Americans as responsible for bringing the virus to the US.

- A submission from Professor Simon Wood of the University of Edinburgh (PC-10713) highlighted concerns that fact-checkers have inadequate technical knowledge to effectively fact-check complex scientific papers and evidence.

- The submission from Media Matters for America (PC-10758) called attention to the impact of Meta's cross-check system in undermining efforts to address misinformation. Because celebrities, politicians, journalists and other prominent users were "afforded slower or more lenient enforcement" for content violations, misinformation was allowed to remain on the platform.

- Several submissions noted Meta's responsibility to address the risks to public safety given its reach and the role of its systems in amplifying misinformation. Concerns were raised about the adequacy of labels and demotions in addressing the risk of harm. For example, the submission from the senior vice president of the Center of Internet and International Studies (PC-10673) highlighted concerns that labels are insufficient to address misinformation disseminated by politicians and prominent influencers. This is because, "simply labelling the posting is insufficient for the potential risk. Permitting the publication of false information that increase the chances of death or serious illness is an abdication of responsibility."

- Several submissions recommended relying on labels and demotions instead of removing misinformation. One submission from Assistant Professor Saiph Savage of Northeastern University and Universidad Nacional Autónoma de México (UNAM) (PC-10519) called attention to the impact of the removals policy and the associated penalties on indigenous communities and voices, noting that indigenous communities "have expressed themselves differently about best practices to treat COVID-19 due to their religious beliefs."

33. To read public comments submitted for this policy advisory opinion, please click here.

Regional stakeholder roundtables

34. The Board engaged in further stakeholder consultations through a series of regional stakeholder roundtables. The Board, in partnership with civil society organizations, convened six roundtable discussions with stakeholders from North America, Latin

America, Africa, Asia and Europe. Through these roundtables, the Board spoke with approximately 100 individuals representing fact-checking organizations, public health bodies and experts, misinformation researchers, digital literacy and communication experts and human rights advocates. These engagements were held under the Chatham House Rule to ensure frank discussion and to protect participants.

35. The following themes and problems emerged through these consultations:

- Issues common across regions include: lack of data and the challenge of measuring both the scale of misinformation in countries, and the impacts of Meta's existing policies, given the lack of access to Meta's data and internal research; the significant amount of misinformation about the severity of the virus, home remedies or alternative treatments (including promotion of bleach), the connection of pandemic measures to 5G technology and anti-vaccine misinformation; concerns that the removal policy may lead to overenforcement that can stifle speech; the need to address Facebook's underlying architecture that promotes misinformation; major spreaders of misinformation having financial and/or political motives for promoting this content; COVID-19 misinformation not only exacerbating the public health crisis, but undermining people's trust in institutions, scientific communication and scientific and medical treatments such as vaccines.

- Issues identified by stakeholders in Latin America include: fact-checking can be effective but misinformation is often labelled after the content has reached the target audience; concerns raised that fact-checking is not scalable and that there is significantly less coverage in languages other than English; organized harassment against scientists who discredit misinformation; individual fact-checkers have been threatened and harassed, some have fled countries or regions for fear of their physical safety; fact-checking organizations have been sued for fact-checking and have had to defend those suits, draining their already limited resources; politicians and prominent figures are not subject to fact-checking; public health experts and health care personnel do not have the knowledge or capacity to effectively counter misinformation campaigns, especially those promoted by notorious influencers or actors with political or economic interests, such as the promotion of inefficient alternative medication for COVID-19.

- Issues identified by stakeholders in North America include: medical professionals reported the significant strain put on health care providers to address misinformation, requiring considerable time and resources and leading to burn out; the need for better and fairer appeals mechanisms for users who have content removed; the concern that much of the research examining prevalence of misinformation and the effectiveness of various interventions focuses on the U.S. and Western Europe; concerns raised that fact-checking is not scalable and has significantly less coverage in non-English languages; the need for a more effective account penalties; concerns that

inconsistencies in the removal of misinformation could reaffirm existing conspiracy theories; difficulties of content moderation for video formats.

- Issues identified by stakeholders in Asia include: misinformation was widespread at the start of the pandemic and has been spread through social media as well as traditional media; misinformation narratives in multiple countries targeting minorities or vulnerable populations, such as migrant workers or religious minorities, as spreaders of the virus; fact-checking cannot move at the pace of misinformation and often comes too late; governments have used the threats from misinformation to target media agencies and shut down dissenting voices; fact-checked articles are not as accessible or compelling as content that contains misinformation; greater coordination among fact-checking organizations in the region would be useful as the same or similar narratives spread from one country to another; fact-checking organizations do not have the resources to provide fact-check articles in the diversity of languages spoken in most countries.

- Issues identified by stakeholders in Africa include: religious leaders in Africa were prominent spreaders of misinformation; governments and opposition parties have been prominent spreaders of COVID-19 misinformation; distrust in government and public institutions has created fertile ground for misinformation to take hold; public health authorities have faced significant challenges in their public communications and there is a great deal of variation among countries in terms of effectiveness of public health authorities and availability of resources; fact-checking is not available in all languages and this limits its effectiveness.

- Issues identified by stakeholders in Europe include: the effect of misinformation will vary from country to country, depending on levels of digital literacy, trust in government and public health institutions, and an open media environment; some raised the concern that removal of misinformation will stifle open debate on matters of public concern; some argue that the most vulnerable populations are the most affected by misinformation, such as the immunosuppressed, children, those with low digital literacy, low access to diverse media, and people that lack access to an adequate health care system; the deterrent effects of misinformation on people's decision-making process; others mentioned the need to adopt measures that allow people to openly discuss public health measures such as the use of face masks or social distancing; some pointed out that misinformation removed from one platform easily moves to another, there needs to be a coordinated approach; the need to design a preventive, rather than responsive, approach to disinformation; the need to promote good and accurate information on Meta's platforms.

Meta engagement

36. After Meta submitted its request for this policy advisory opinion, between July and December 2022, the Board submitted 50 written questions to the company. These were addressed by Meta either in writing or orally in three question and answer sessions. Forty questions were answered in full and 10 were partially answered. The partial responses had to do with requests to breakdown data by region and language, internal research on the effectiveness of the various enforcement measures, and how the company weighed competing considerations and expert advice in devising its policy. The Board submitted questions about: Meta's internal data or research on prevalence of COVID-19 misinformation on Meta's platforms globally and broken down by country and language; the process and measures used to enforce the removal policy and data on number of removals; public reporting on enforcement data for the misinformation community standard; the role of public health authorities and experts in developing and enforcing the removal policy; any research on the effectiveness and impact of removal of COVID-19 misinformation; any research on the effectiveness and impact of other measures, including fact-checking, neutral labels and demotions; the role of third-party fact-checking organizations; the penalties applied for COVID-19 misinformation; the internal guidelines implementing the COVID-19 misinformation policy; whether the company had evaluated the effect of the cross-check program on the effectiveness of the COVID-19 misinformation policy; the stakeholder engagement process and experts consulted during the development of this policy and how Meta evaluated the input received; the feasibility of adopting a bifurcated enforcement of the policy where removal would be enforced in some countries but not others; and additional alternative measures, including investments in digital literacy.

V. Framework for Board analysis and assessment of the policy on *Misinformation about health during public health emergencies*

37. In its request, Meta asked the Board whether it should continue removing certain content on COVID-19 under the policy on ***Misinformation about health during public health emergencies*** or whether a less restrictive approach would better align with the company's values and human rights responsibilities. To answer this question, the Board analyzes whether the policy is consistent with Meta's values and human rights commitments, including whether it is necessary and proportionate at the present time, or if the company should adopt a less restrictive approach to address misinformation on COVID-19. The recommendations derived from this analysis are set forth in the final section of this policy advisory opinion.

38. In outlining the strategy that Meta undertook in developing its policy on COVID-19 misinformation, Meta called attention to the significant divergence in assessment of risk and mitigation measures among experts in different disciplines and stakeholders in different regions. The Board, in its own research and stakeholder engagement, also heard opposing and conflicting positions on the risks involved in leaving COVID-19

misinformation on the platforms and the effectiveness of various measures in addressing the risks.

39. There is no single agreed solution to the problem of COVID-19 misinformation and the risks it presents to human rights, life and health, particularly those of the most vulnerable. A more regionally grounded approach would better establish the necessary link between misinformation and imminent physical harm. However, the Board had to consider Meta's representations about the current limitations of its existing systems and how COVID-19 misinformation narratives move across the world. Board Members responded differently to these concerns and this policy advisory opinion reconciles, to the extent possible, the various perspectives on the Board. It is the result of a careful compromise around the difficult need to account for different approaches to COVID-19 misinformation across the globe amid a public health emergency, taking into account Meta's purported technical limitations. It may not, therefore, represent the personal views of each Board Member.

Meta's values

40. Meta's values are outlined in the introduction to Facebook's Community Standards where the value of "Voice" is described as "paramount." Meta limits "Voice" in service of four other values, two of which are relevant here: "Safety" and "Dignity." To protect the value of "Safety" Meta "remove[s] content that could contribute to a risk of harm to the physical security of persons." The value of "Dignity" states that "all people are equal in dignity and rights," and users are expected to "respect the dignity of others and not harass or degrade others."

41. The Board finds Meta's policy on ***Misinformation about health during public health emergencies*** complies with Meta's values of "Voice," "Safety," and "Dignity." During a public health emergency, the risk of harm is significant, and the value of "Voice" may be limited to serve the value of "Safety" for health misinformation "likely to directly contribute to risk of imminent physical harm." Imminent harms from COVID-19 misinformation will fall disproportionately on the most vulnerable, including the immunocompromised and people with other underlying conditions, persons with disabilities, poor communities and the elderly as well as health care workers.

Meta's human rights responsibilities

42. On 16 March 2021, Meta announced its Corporate Human Rights Policy, where it outlines its commitment to respecting rights in accordance with the UN Guiding Principles on Business and Human Rights (UNGPs). The UNGPs, endorsed by the UN Human Rights Council in 2011, establish a voluntary framework for the human rights responsibilities of private businesses.

43. According to Principle 12 of the UNGPs, the responsibility of business enterprises to respect human rights refers to internationally recognized human rights, understood, at a minimum, as those referred to in the International Bill of Human Rights. This is comprised of the Universal Declaration on Human Rights, the International Covenant on Civil and Political Rights (ICCPR) and the International Covenant on Economic, Social, and Cultural Rights (ICESCR). This responsibility means that companies "should avoid infringing on the human rights of others and should address adverse human rights impacts with which they are involved" (Principle 11). Companies are expected to: "(a) Avoid causing or contributing to adverse human rights impacts through their own activities, and address such impacts when they occur; (b) Seek to prevent or mitigate adverse human rights impacts that are directly linked to their operations, products or services by their business relationships, even if they have not contributed to those impacts" (Principle 13).

44. Principle 17 further states that in order to "identify, prevent, mitigate, and account for how they address their adverse human rights impacts," companies should "carry out human rights due diligence." This process should include assessing actual and potential human rights impacts, integrating and acting upon the findings, and tracking responses, and communicating how impacts are addressed. Responsibility to undertake human rights due diligence is ongoing, recognizing that human rights risks may change over time as the business enterprise's operations and operating context evolve. Finally, Principle 20 establishes that business enterprises should track the effectiveness of their response, based on appropriate qualitative and quantitative indicators, and draw on feedback from both internal and external sources, including affected stakeholders.

45. The Board's analysis in this policy advisory opinion was informed by the following human rights standards:

- The right to freedom of expression, as protected under Article 19, para. 2 of the ICCPR. This article provides broad protection for freedom of expression through any media, and regardless of frontiers. The right to freedom of expression extends to the right to seek, receive, and impart information of all kinds.

- The right to life (Art. 6, ICCPR): every human being has the inherent right to life.

- The right to health (Art. 2 and 12, ICESCR): the right of everyone to enjoy the highest attainable standard of physical and mental health. Art. 12(2) states that the realization of this right includes the "creation of conditions which would assure to all medical service and medical attention in the event of sickness." It encompasses "underlying determinants to health," such as access to health-related education and information as well as "participation of the population in all health-related decision-making at the community, national and international levels." (General Comment No. 14, ICESCR, para

11.) Information accessibility includes the right to seek, receive and impart information and ideas concerning health issues. Respecting the right to health means safeguarding legitimate discussion of public health issues.

- The right to enjoy the benefits of scientific progress and its applications (Art. 15(1)(b), ICESCR).

- The right of non-discrimination (Art. 26, ICCPR): Art. 26 prohibits discrimination and guarantees to all people equal and effective protection against discrimination on grounds of any protected characteristic.

- The right to an effective remedy (Art. 2, ICCPR).

46. The UN Special Rapporteur on the right to freedom of expression has emphasized the importance of the right to freedom of expression in the context of the COVID-19 pandemic, noting that "promoting access to information bolsters the promotion of health, life, autonomy and good governance" and "warned against viewpoint discrimination." (A/HRC/44/49, para. 2, 52.) Misinformation during a public health emergency can significantly impact peoples' rights to access reliable information and health guidance and resources, essential for the protection of the right to health and the right to life. As the UN Special Rapporteur noted, "lies and propaganda deprive individuals of autonomy, of the capacity to think critically, or trust in themselves and in sources of information, and of the right to engage in the kind of debate that improves social conditions." (A/HRC/44/49, para 60.) The Special Rapporteur also noted how "false information is amplified by algorithms and business models that are designed to promote sensational content that keep users engaged on platforms" and called on companies to "respond to these concerns, going beyond improving content moderation to reviewing their business models." (A/HRC/47/25, para. 16, 95.)

47. Article 19 allows the right to freedom of expression to be restricted under certain narrow and limited conditions, known as the three-part test of legality (clarity), legitimacy, and necessity, which also includes an assessment of proportionality. The UN Special Rapporteur on freedom of opinion and expression has suggested that Article 19, para. 3 of the ICCPR provides a useful framework to guide platforms' content moderation practices and that companies should tie their content policies to human rights principles (A/HRC/38/35, para. 10-11, A/74/486, para. 58). The Board has acknowledged that while the ICCPR does not create obligations for Meta as it does for states, Meta has committed to respecting human rights as set out in the UNGPs. (A/74/486, para. 47- 48). Therefore, when the company's policies differ from the high standard States must meet to justify restrictions on speech, Meta must provide a reasoned explanation of the policy difference, consistent with the human rights standards they have committed to respect (para. 47- 48).

Legality (clarity and accessibility of the rules)

48. Any restriction on freedom of expression should be accessible and clear enough, in scope, meaning and effect, to provide guidance to users and content reviewers as to what content is permitted on the platform and what is not. Lack of clarity or precision can lead to inconsistent and arbitrary enforcement of the rules. (A/HRC/47/25, para 40).

49. In its "Claimed COVID cure" decision [2020-006-FB-FBR], the Board recommended that Meta "set out a clear and accessible Community Standard on health misinformation, consolidating and clarifying existing rules in one place (including defining key terms such as misinformation). This rule-making should be accompanied with 'detailed hypotheticals that illustrates the nuances of interpretation and application of [these] rules to provide further clarity for users.'" In response to the Board's recommendation, Meta created the Misinformation Community Standard. It also published an article in its Help Center that provides a list of claims subject to removal as well as common questions on how the policy is enforced, including how the company approaches humor, satire, and personal anecdotes under the policy. The Board commends the company for taking these steps.

50. The claims removed under this policy are on a spectrum in terms of how broad or specific they are. For example, several claims currently subject to removal are narrowly defined by Meta (e.g. "claims that COVID-19 social distancing orders are really just a way to install 5G wireless communication technology infrastructure"), while others are worded more broadly (e.g. "claims that social/physical distancing does not help prevent the spread of COVID-19"). The Board has not analyzed whether the restriction in each of these claims is sufficiently clear, as the responsibility to ensure precision and clarity belongs in the first instance to Meta. The Board notes that Meta should possess information about which claims have systemically resulted in under and over enforcement problems, which may indicate relevant vagueness issues. Additionally, the Board notes that the specific claims subject to removal under the COVID-19 Misinformation policy are provided in a Help Center page. The page does not have a change log, which would allow users to see when a claim has been added, removed or edited.

51. To better align the ***Misinformation about health during public health emergencies*** policy with legality standards, the Board issues recommendations 1, 2, 3, 4, and 11 explained in detail in section VI, below.

Legitimate aim

52. Restrictions on freedom of expression must pursue a legitimate aim, which includes the protection of the rights of others and public health, among other aims. The Human Rights Committee has interpreted the term "rights" to include human rights as recognized in the ICCPR and more generally in international human rights law (General Comment 34, para. 28).

53. Meta's policy on ***Misinformation about health during public health emergencies*** is directed towards the legitimate aims of protecting public health during a health crisis, as well as protecting individuals' right to access information, right to life, right to health, right to enjoy the benefits of scientific progress and its applications, and the right to non-discrimination.

Necessity and proportionality

Overview

54. Any restrictions on freedom of expression "must be appropriate to achieve their protective function; they must be the least intrusive instrument amongst those which might achieve their protective function; they must be proportionate to the interest to be protected" (General Comment 34, para. 34).

55. For reasons explained below, the Board finds that the policy authorizing Meta to remove COVID-19 misinformation "likely to directly contribute to the risk of imminent physical harm" during a public health emergency is necessary and proportionate. It is therefore, in principle, compatible with the company's values and its human rights responsibilities. By declaring a public health emergency, the United Nations' public health authority establishes that there is an extraordinary event, which constitutes a risk to public health or life of human populations, through the international spread of a disease that presents "a serious and direct danger" (WHO International Health Regulation 2005). Given the WHO's declaration of a global public health emergency with respect to COVID-19, a disease whose consequences were highly uncertain, volatile and lethal, the Board finds that Meta's response was proportionate. The Oversight Board understands that under such emergency circumstances, certain harmful health misinformation, especially when distributed on a large scale or by prominent influencers, can lead to serious public health harms and adversely affect the rights of individuals on and off Meta's platforms. The Board is cognizant that during the most intensive phase of a public health emergency, it may not be possible to conduct robust prior consultations with numerous experts on individual misinformation claims. In assessing the proportionality of Meta's approach, the Board also considered the company's position that a localized approach to COVID-19 misinformation was not feasible.

56. However, as COVID-19 circumstances change, the necessity and proportionality calculus necessarily changes as well. The Board recognizes that the impact of COVID-19 varies across the world. This depends on the spread of the virus, a country's health systems, and the quality of civic space that allows people to receive and share information about COVID-19, among other things. Though the WHO's emergency COVID-19 declaration remains in effect (and it was reiterated in January 2023), in many parts of the world COVID-19 cases have abated and emergency measures have been

scaled back dramatically. This contributes to the difficulty of implementing a global approach that satisfies the proportionality test. As outlined in recommendation 1 below, Meta should begin a transparent and inclusive process to determine whether any of the 80 claims subject to removal are no longer false or "likely to directly contribute to the risk of imminent physical harm." The process should include a mechanism for dissenting views to be heard and considered. These should include diverging views within the scientific community, experts on freedom of expression, and of those who have expertise on how misinformation circulates online and its impacts. The Board also calls on Meta in recommendation 4 below, to begin a process to identify human rights risks that may persist in some countries, and to prepare for a more localized approach to mitigating those risks once the global health emergency has ended.

Stakeholder Input

57. Stakeholders in multiple regions around the world spoke to the Board about politicians, religious leaders, influencers and medical authorities who promoted misinformation with great visibility, with whom fact-checkers, scientific experts, and public health authorities were unable to keep up. Stakeholders in each region also spoke of the impact of misinformation on people turning to alternative treatments or on their willingness to get vaccinated. They noted that misinformation impacted individuals' willingness to follow public health guidance or to adopt preventive measures. This type of misinformation was highlighted by stakeholders as frustrating preventive measures and risk management, which in turn affects the general population, disproportionally impacting vulnerable groups such as the immunocompromised, persons with disabilities, persons with pre-existing conditions, the elderly and poor and marginalized communities. (For more on COVID-19 data see the WHO Dashboard.) Robust studies on the impact of online misinformation show an increased disregard for public health guidance and reduced likelihood of accepting future diagnostic tests or a vaccination. The Board takes note of other reported harms that have resulted from COVID-19 misinformation, including undermining trust in scientific and public health authorities. This frustrates the effective implementation of public health measures for COVID-19 and for other public health crises. Other reported harms of COVID-19 misinformation include direct attacks, harassment and strategic lawsuits against fact-checking organizations and individual fact-checkers.

58. Experts also noted that after Meta began removing COVID-19 misinformation, the overall quantity of misinformation on the platform was substantially reduced and argued that without those measures, misinformation will rise again and anti-vaccination content on social platforms such as Facebook will dominate discourse. Those experts noted that the lack of transparency and access to Meta's data or internal research frustrates efforts to find clear evidence of the effectiveness of measures to address misinformation, including removal. However, stakeholders across the globe,

argued to the Board that as long as widespread loss of human life and the risk to health of countless people is ongoing, the company must continue urgent measures and accept that any mistakes must be on the side of saving lives at risk, especially those of the most vulnerable people. While the availability of reliable scientific information on COVID-19 has significantly improved since the start of the pandemic, access to that information varies between countries and communities and the scale of false and misleading information has made accessing and evaluating existing scientific information difficult for people across the world, undermining the benefits of access. In this regard, for example, a submission from the Khazanah Research Institute (PC-10703), a policy research institute in Malaysia, highlighted the differing levels of access to reliable health information in different countries and the diverse levels of risk from leaving misinformation unmoderated. The same position was supported by other experts and stakeholders from different parts of the world, especially from countries with lower income levels. If there must be a global approach, the Khazanah Research Institute submission recommends, Meta should err on the side of caution and continue to remove harmful COVID-19 misinformation.

59. As Meta recognized in its request to the Board, the course of the pandemic has and will continue to vary across the world. There are important variations in vaccination rates, health care system capacity and resources, and trust in authoritative guidance. These contribute to the disproportionate effect of the virus on most vulnerable people in different countries. Although vaccines have been developed and are readily available in the US and other countries across the globe, this does not reflect global trends. In Meta's words: "Eighty percent of people in high-income countries have received at least one dose of the vaccine, as opposed to only 13 percent of people in low-income countries. Low-income countries are also more likely to have health care systems with less capacity, less robust economies, and lower trust in government guidance, all of which will add challenges to vaccinating people and treating those that contract COVID-19." (Meta policy advisory opinion request, page 15, July 2022). To cite just a few cases that show the significant difference in vaccination rates, as of February 2023, under 20% of the population in Iraq has completed the primary series, and less than 1% have had a booster. In Bulgaria, around 30% of the population has completed the primary series. That number is 13% in Syria and under 5% in Papua New Guinea and Haiti. Several experts consulted by the Board warned of the danger of relying on information and data that focuses overwhelmingly on Western countries for a global policy and approach. These experts also noted the narrow geographic perspective of most empirical studies on misinformation and disinformation.

60. In January 2023, the WHO noted that while "the world is in a better position than it was during the peak of the Omicron transmission one year ago, more than 170,000 COVID-19 related deaths have been reported globally within the last eight weeks" and health systems "are currently struggling with COVID-19 and caring for patients with influenza

and respiratory syncytial virus (RSV), health workforce shortages, and fatigued health workers." The WHO also highlighted that "the COVID-19 response remains hobbled in too many countries unable to provide [vaccines, therapeutics, and diagnostic tools] to the populations most in need, older people and health workers." The WHO Committee noted "[v]accine hesitancy and the continuing spread of misinformation continue to be extra hurdles to the implementation of crucial public health interventions."

Meta's insistence on a global approach

61. Meta acknowledges that the course of the pandemic has been different across the globe, with the most significant variation being between "developed" and "less-developed" nations. In seeking the Board's guidance, the company all but ruled out a localized approach, stating that applying such enforcement measures "would create significant transparency and fairness concerns, result in a poor user experience, and be operationally infeasible." According to Meta, adopting regional or country specific enforcement measures at scale would lead to lack of clarity for users on what policies and penalties will apply to a piece of content, given how users and information travel across borders. This approach would require an even more complex and lengthy policy "outlining where and under what circumstances different claims are either removed, demoted, or subject to other enforcement." According to the company, it does not currently have capacity to adopt a localized approach and developing one would require extensive resources and time, making this approach infeasible for the immediate future. It argued that, "[e]nforcing policies at the country level can lead to both over-enforcement when one set of market reviewers covers multiple countries, and under-enforcement because content can spread across countries and regions." Given this, Meta stated that the proposed policy should be appropriate for all regions, "while being consistent and workable globally."

Analysis

62. In reaching its decision on the proportionality issue, the Board considered various factors, including: (i) potential human rights harms in an ongoing public health emergency; (ii) the burdens on freedom of expression; (iii) the relevant Community Standard's requirement that content subject to removal must be considered both false and likely to directly contribute to imminent and significant physical harm; (iv) the platform's architecture which, according to some experts, could contribute to the amplification of harmful content (see recommendation 10 on the need to carry out a Human Rights Impact Assessment on platform design choices); (v) the significant concerns raised about the scalability and effectiveness of content moderation measures short of removals (as explained in the following paragraphs about fact-checking, demotions, and labelling); and (vi) Meta's representation that a scaled, localized approach was not feasible in implementing its policy.

63. Given Meta's insistence on a global approach, and while COVID-19 continues to be designated a "Public Health Emergency of International Concern" by the WHO, the Board cannot recommend a change to how Meta implements its global policy without additional due diligence, and an assessment by the company of the impact of its policies and various enforcement tools. Recommending a change in these circumstances could disproportionately affect the most vulnerable across the globe. This includes people who are elderly, immunocompromised, and have preexisting conditions, as well as poor and marginalized communities with fewer resources, more fragile civic spaces, no other reliable sources of information, and poor health systems or lack of access to health services. As noted above, the Board is cognizant that during the most intensive phase of a public health crisis, Meta used exceptional measures. The Board understands that the company needed to take exceptional measures during a declared public health emergency, such as in this case, by removing entire categories of misinformation based on an assessment provided solely by a public health authority with the purpose of avoiding likely imminent physical harm. The Board finds the measures were proportional given the unique circumstances of the pandemic.

64. However, such exceptional measures must be temporary, strictly tailored to the exigencies of the circumstances and identified publicly. As circumstances change so does the necessity and proportionality analysis. Given the evolving nature of pandemics, Meta must now undertake a more robust consultation process as soon as practicable to try to ensure that the automated removal of specific claims does not stifle debate on matters of public interest or lead to undue governmental influence on Meta's content moderation. The consultation process should draw on the expertise of a more diverse set of stakeholders, including dissenting voices, (as set forth in recommendation 1, below). The Board notes that Principle 17 of the UNGPs states that to "identify, prevent, mitigate and account for how they address their adverse human rights impacts," business enterprises must carry out an ongoing human rights due diligence, which includes, "assessing actual and potential human rights impacts, integrating and acting upon the findings, tracking responses, and communicating how impacts are addressed." In addition, as stated by Principle 20 of UNGPs, the company should track the effectiveness of its response, "based on appropriate qualitative and quantitative indicators and draw on feedback from both internal and external sources, including affected stakeholders."

65. As noted above, in reaching its conclusion, the Board considered whether less intrusive measures short of content removals could address the scale of misinformation and protect public health during a public health emergency, as well as the rights of people on and off the platform. First, while adding fact-check labels to content provides a means of correcting information without removing it, several stakeholders, as well as information provided by Meta, show the limited capacity of this tool to address the speed and scale of likely harmful health misinformation during a public health emergency. Meta informed the Board that fact-checkers are unable to review an

overwhelming majority of the content in their queue. Meta also stated that it would be unable to scale-up the fact-checking program, as these are third-party organizations not controlled or owned by Meta. Additionally, limitation built into the program makes this measure less effective. Meta does not permit its fact-checkers to review content shared by politicians, which includes candidates running for office, current office holders and their appointees, and political parties and their leaders. As has been widely reported and confirmed by stakeholders from each region, these kinds of users have been prominent spreaders of misinformation. Verification by fact-checkers takes longer than automated removal at scale, which can be a determining factor when dealing with harmful misinformation in the context of a public health crisis. This measure additionally refers the user to an article that is usually outside the platform (and therefore less accessible to people who do not have the resources to consume additional data). The language of these articles is often particularly technical and sometimes complex, as opposed to the short, emotive messages through which misinformation is spread. A submission from Professor Simon Wood of the University of Edinburgh (PC-10713) highlighted the concern that fact-checkers often have insufficient technical knowledge to effectively fact-check complex scientific papers and evidence.

66. Second, while demotions impact where in a user's feed the piece of content will appear, the individualized nature of each user's feed means that the impact of this measure on virality or reach of a piece of content is difficult to determine. The ranking score of a piece of content aims to show users content that they "may be most interested in," and content shared in a group or by a page a user follows is likely to be ranked highly. As a result, it is unclear whether a demotion would effectively address the reach of the content shared by users with significant following or content shared in a group. Demotion is likely to have the least impact for users who follow multiple accounts, pages or groups that regularly share COVID-19 misinformation, given the overall inventory of content in their newsfeed. It also appears that the company does not have data on how many users are less likely to access demoted content, even if that content was demoted significantly. Demotions alone are not accompanied by strikes and penalties. Finally, as users are not able to appeal demotions of their content, this option would raise significant concerns about treating users fairly.

67. Third, according to the company's internal research, there is no evidence that neutral labels are effective at reaching users at scale and informing their knowledge or attitudes. Meta applies NITs (or neutral labels) through an automated system that detects a COVID-19 topic in a post. These labels provide a link to a COVID-19 Information Center, which provides authoritative information about COVID-19. According to Meta, the company's preliminary research on these labels showed that the "click through rate" (the rate at which users click the label to see the authoritative information) decreases the more NITs a user sees. Meta further informed the Board that the company has stopped using COVID-19 NITs. According to Meta, these labels have no

detectable effect on users' likelihood to read, create or re-share fact-checked misinformation or discouraging vaccine content. Finally, the company reported that initial research showed that these labels may have no effect on user knowledge and vaccine attitudes.

68. In sum, the Board concludes that, given Meta's insistence on a global approach to COVID-19 misinformation and the continued WHO emergency declaration, Meta should continue to apply its policy on COVID-19 misinformation that is likely to directly contribute to the risk of imminent physical harm. At the same time, it should begin to undertake a robust and inclusive due diligence review process of the claims currently being removed. To better align the ***Misinformation about health during public health emergencies policy*** with necessity and proportionality standards, the Board issues recommendations 1, 4, 5, 9, 10, 12, 13, 14, 15, and 18 explained in detail in section VI below.

VI. Recommendations

Recommendations on content policy

69. **Recommendation 1**: *Given the World Health Organization's declaration that COVID-19 constitutes a global health emergency and Meta's insistence on a global approach, Meta should continue its existing approach of removing globally false content about COVID-19 that is "likely to directly contribute to the risk of imminent physical harm." At the same time, it should begin a transparent and inclusive process for robust and periodic reassessment of each of the 80 claims subject to removal to ensure that: (1) each of the specific claims about COVID-19 that is subject to removal is false and "likely to directly contribute to the risk of imminent physical harm"; and (2) Meta's human rights commitments are properly implemented (e.g., the legality and necessity principles). Based on this process of reassessment, Meta should determine whether any claims are no longer false or no longer "likely to directly contribute to the risk of imminent physical harm." Should Meta find that any claims are no longer false or no longer "likely to directly contribute to the risk of imminent physical harm," such claims should no longer be subject to removal under this policy. The Board will consider this recommendation implemented when Meta announces a reassessment process and announces any changes to the 80 claims on the Help Center page.*

70. The sub-parts below outline the Board's recommendations on best practices for carrying out the reassessment of the claims subject to removal under the ***Misinformation about health during public health emergencies*** policy. Each of the sub-recommendations will be considered as separate recommendations from recommendation 1, meaning the Board will review Meta's actions to implement the recommendations separately.

Recommendation 1A: Broader expert and stakeholder consultation

71. *The company must put a process in place, as soon as feasible, to consider a broader set of perspectives in evaluating whether the removal of each claim is needed by the exigencies of the situation. The experts and organizations consulted should include public health experts, immunologists, virologists, infectious disease researchers, misinformation and disinformation researchers, tech policy experts, human rights organizations, fact-checkers, and freedom of expression experts. The Board will consider this recommendation implemented when Meta publishes information about its processes for consultation with a diverse set of experts on its policy on* ***Misinformation about health during public health emergencies****, as well as information about the impact of those conversations on its policy.*

72. As outlined above, the Board accepts that the company needed to take exceptional measures during a declared public health emergency, in this case, by removing entire categories of misinformation on the basis of an assessment provided by a single public health authority. The Board is cognizant that during a public health emergency it is not possible to immediately conduct robust prior consultations with numerous experts on individual claims. However, as soon as feasible, a broader group of experts and stakeholders must be consulted given the continuously evolving information about the novel pandemic and the various views surrounding the best approach to addressing pandemic-related misinformation. As the company stated, it had to change its position on at least two claims previously subject to removal, one on the origins of the virus and the other on the mortality rate of COVID-19. Broader consultation and greater transparency on that input is essential for better decision-making and to guard against unjustified censorship.

73. The Board asked Meta whether the claims that are included in its "do not post" list (because they are considered false and "likely to directly contribute to the risk of imminent physical harm") had been reevaluated to take account of the impact of the three changes outlined in the company's request. Meta informed the Board that it has no information to support the conclusion that the current claims being removed are no longer false or are no longer likely to directly contribute to the risk of imminent harm. However, the company has not returned to the relevant public health authorities to ask them to re-evaluate the claims. Nor has the company conducted stakeholder or expert consultation to re-evaluate the individual claims or the overall policy. According to Meta, the company elected to turn to the Board with a policy advisory opinion request, instead of undertaking external stakeholder engagement on changing its policy, in order not to delay the request. The Board commends Meta for seeking external input on the policy developed during a global emergency, and for recognizing the need for re-assessment. However, the company's responsibility to respect human rights does not stop there. Putting a process in place to evaluate whether the continued removal of

each claim is necessary would ensure that the company is conducting relevant due diligence, as per the UNGPs.

Recommendation 1B: Timing of review

74. *Meta should establish the timing for this review (e.g., every three or six months) and make this public to ensure notice and input. The Board will consider this recommendation implemented when Meta publishes the minutes of its review meeting publicly, in a similar fashion to its publication of its public policy forum minutes in its Transparency Center.*

Recommendation 1C: Procedures for collecting public input

75. *Meta should articulate a clear process for regular review, including means for interested individuals and organizations to challenge an assessment of a specific claim (e.g., by providing a link on the Help Center page for public comments, and virtual consultations). The Board will consider this recommendation implemented when Meta creates a mechanism for public feedback and shares information on the impact of that feedback on its internal processes with the Board.*

Recommendation 1D: Guidance on type of information to be considered and evaluated

76. *Meta's review of the claims should include the latest research on the spread and impact of such online health misinformation. This should include internal research on the relative effectiveness of various measures available to Meta, including removals, fact-checking, demotions, and neutral labels. The company should consider the status of the pandemic in all regions in which it operates, especially those in which its platforms constitute a primary source of information and where there are less digitally literate communities, weaker civic spaces, a lack of reliable sources of information, and fragile health care systems. Meta should also evaluate the effectiveness of its enforcement of these claims. Meta should gather, if it doesn't already possess, information about which claims have systemically resulted in under and over enforcement problems. This information should inform whether a claim should continue to be removed or should be addressed through other measures. The Board will consider this recommendation implemented when Meta shares data on its policy enforcement review and publishes this information publicly.*

Recommendation 1E: Guidance on providing transparency on decision-making

77. *In order to provide transparency on the types of experts consulted, their input, the internal and external research considered and how the information impacted the outcome of the analysis, Meta should provide to the Board a summary of the basis for its decision on each claim. The summary should specifically include the basis for the company's decision for continuing to remove a claim. Meta should also disclose what role, if any, government personnel or entities played in its decision-making. If the company decides to cease removing a specific claim, the company should explain the basis of that decision*

(including: (a) what input led the company to determine that the claim is no longer false; (b) what input, from what source, led the company to determine the claim no longer directly contributes to the risk of imminent physical harm, and whether that assessment holds in countries with lowest vaccination rates and under-resourced public health infrastructure; (c) did the company determine that its enforcement system led to over-enforcement on the specific claim; (d) did the company determine that the claim is no longer prevalent on the platform.) The Board will consider this recommendation implemented when Meta shares the assessment of its policy evaluation process. This information should align with the reasons listed publicly in the Help Center post for any changes made to the policy, as outlined in the first paragraph of this recommendation.

78. **Recommendation 2:** *Meta should immediately provide a clear explanation of the reasons why each category of removable claims is "likely to directly contribute to the risk of imminent physical harm." The Board will consider this recommendation implemented when Meta amends the Help Center page to provide this explanation.*

79. Currently, the Help Center page provides an example of the link between a specific claim and why and how it contributes to risk of imminent physical harm by "increasing the likelihood of exposure to or transmission of the virus, or having adverse effects on the public health system's ability to cope with the pandemic." The same page then identifies five categories of false information that, according to Meta, satisfy the "likelihood to contribute to imminent physical harm" standard. The Help Center page does not, however, systematically explain how each category of removable claims satisfies the established standard. Meta should explicitly explain how each category of claims is likely to directly contribute to the risk of imminent physical harm and the sources of information the company relied on to reach that conclusion.

80. **Recommendation 3:** *Meta should clarify its* ***Misinformation about health during public health emergencies*** *policy by explaining that the requirement that information be "false" refers to false information according to the best available evidence at the time the policy was most recently re-evaluated. The Board will consider this recommendation implemented when Meta clarifies the policy in the relevant Help Center page.*

81. At least twice, Meta has had to amend the claims subject to removal when known information changed, or evolution of the disease made a claim inaccurate or incomplete. Mistakes may be made, new data or research may challenge the existing consensus, or the definition of a claim may need to be refined. Given this reality, and to make clear that Meta understands that it has a responsibility to continually reevaluate the assessment that specific claims satisfy the broader standard in its policy, Meta should clarify the policy to make clear the assessment is based on best available evidence at the time and may evolve.

Recommendations on enforcement

82. **Recommendation 4**: *Meta should immediately initiate a risk assessment process to identify the necessary and proportionate measures that it should take, consistent with this policy decision and the other recommendations made in this policy advisory opinion, when the WHO lifts the global health emergency for COVID-19, but other local public health authorities continue to designate COVID-19 as a public health emergency. This process should aim to adopt measures addressing harmful misinformation likely to contribute to significant and imminent real-life harm, without compromising the general right to freedom of expression globally. The risk assessment should include: (1) a robust evaluation of the design decisions and various policy and implementation alternatives; (2) their respective impacts on freedom of expression, the right to health and to life and other human rights; and (3) a feasibility assessment of a localized enforcement approach. The Board will consider this recommendation implemented when Meta publicly communicates its plans for how it will conduct the risk assessment and describes the assessment process for detecting and mitigating risks and updates the Help Center page with this information.*

83. **Recommendation 5:** *Meta should translate internal implementation guidelines into the working languages of the company's platforms. The Board will consider this recommendation implemented when Meta translates its internal implementation guidelines and updates the Board in this regard.*

84. Content moderators have access to detailed internal implementation guidelines which provide additional information on how to identify violating content and content that should remain on the platform under one of the established exceptions (e.g. humor, satire, personal anecdote, opinion). In order to ensure consistent enforcement in distinct parts of the world, Meta needs to make sure these guidelines are provided to and are accessible to moderators in the language in which they are operating.

85. The Board has previously recommended that Meta translate its internal implementation guidance provided to moderators into the language in which they review content (see, "Reclaiming Arabic words," [2022-003-IG-UA]; and "Myanmar bot," [2021-007-FB-UA] case decisions). In its response to the Board, Meta stated that "[having] one set of internal policy guidelines in the language in which all of our content reviewers are fluent... is the best way to ensure standardized global enforcement of our rapidly evolving policies... Because this guidance rapidly evolves (it is constantly being updated with new clarifications, definitions, and language including market-specific slurs) relying on translations could lead to irregular lags and unreliable interpretations."

86. Since Meta provided the above explanation, an independent assessment of Meta's enforcement of its policies in Israel and Palestine identified lack of language capability of content moderators as one of the causes of over-enforcement of Meta's policies in Arabic (see, the Business for Social Responsibility's "Human Rights Due Diligence of

Meta's Impacts in Israel and Palestine in May 2021"). Given this finding, and considering the complexity of the internal guidelines and the nuanced interpretation they provide to content moderators, the Board believes that the danger of over or under-enforcement of the ***Misinformation about health during public health emergencies*** policy is real. Meta should mitigate these risks to ensure that the application of its policy is consistent across languages and regions.

87. **Recommendation 6:** *User appeals for a fact-check label should be reviewed by a different fact-checker than the one who made the first assessment. To ensure fairness and promote access to a remedy for users that have their content fact-checked, Meta should amend its process to ensure a different fact-checker that has not already made the assessment on the given claim, can evaluate the decision to impose a label. The Board will consider this recommendation implemented when Meta provides a mechanism to users to appeal to a different fact-checker, and when it updates its fact-checking policies with this new appeals mechanism.*

88. **Recommendation 7:** *Meta should allow profiles (not only pages and groups) that have content labeled by third party fact-checkers enforcing Meta's misinformation policy, to appeal the label to another fact-checker through the in-product appeals feature. The Board will consider this recommendation implemented when Meta rolls out the appeal feature to profiles in all markets and demonstrates that users are able to appeal fact-check labels through enforcement data.*

89. User appeals are a key feature for error correction and for ensuring users' right to access to a remedy. Fact-checkers review content that varies considerably in its complexity, technical content and context. Some errors are inevitable. One public comment raised the concern that fact-checkers do not have the scientific and technical knowledge to fact-check complicated scientific articles shared on the platform. Fact-check labels carry consequences for users. When a fact-checker applies a label to content, the label can lead to a strike, if the content is labeled "false" or "altered." An accumulation of strikes will lead to feature-limits and demotion of content shared by that profile. Implementing this recommendation would allow users to notify fact-checkers when they believe a mistake has been made and to share additional information to facilitate review.

90. **Recommendation 8:** *Meta should increase its investments in digital literacy programs across the world, prioritizing countries with low media freedom indicators (e.g. Freedom of the Press score by Freedom House) and high social media penetration. These investments should include tailored literacy trainings. The Board will consider this recommendation implemented when Meta publishes an article on its increased investments, specifying the amount invested, the nature of the programs and the countries impacted, and information it has about the impacts of such programs.*

91. Meta informed the Board, in response to a question, that over the last three years the company has invested over seven million US dollars "to help people improve their media literacy skills and proactively reduce the amount of misinformation that gets shared." According to the sources provided by Meta, these investments were mostly focused on the United States. Meta has partnered with organizations in other countries to deliver social media campaigns or advertisements focused on media literacy.

92. Studies evaluating the impact of Meta's investments in media literacy programs in the United States (one program in partnership with PEN America and the other with the Poynter Institute) found significant improvements in participants' ability to evaluate online information. For example, participants' ability to detect COVID-19 misinformation improved from pre-intervention average of 53% to post-intervention average of 82%. A media literacy program for seniors resulted in a 22% improvement in participants' ability to accurately judge headlines as true or false after taking the course.

93. **Recommendation 9:** *For single accounts and networks of Meta entities that repeatedly violate the misinformation policy, Meta should conduct or share existing research on the effects of its newly publicized penalty system, including any data about how this system is designed to prevent these violations. This research should include analysis of accounts amplifying or coordinating health misinformation campaigns. The assessment should evaluate the effectiveness of the demonetization penalties that Meta currently uses, in addressing the financial motivations/benefits of sharing harmful and false or misleading information. The Board will consider this recommendation implemented when Meta shares the outcome of this research with the Board and reports a summary of the results on the Transparency Center.*

Recommendations on transparency

94. **Recommendation 10:** *Meta should commission a human rights impact assessment of how Meta's newsfeed, recommendation algorithms, and other design features amplify harmful health misinformation and its impacts. This assessment should provide information on the key factors in the feed-ranking algorithm that contribute to the amplification of harmful health misinformation, what types of misinformation can be amplified by Meta's algorithms, and which groups are most susceptible to this type of misinformation (and whether they are particularly targeted by Meta's design choices). The assessment should also make public any prior research Meta has conducted that evaluates the effects of its algorithms and design choices in amplifying health misinformation. The Board will consider this recommendation implemented when Meta publishes the human rights impact assessment, which contains such analysis.*

95. The UN Special Rapporteur on freedom of expression described social media platforms' responses to COVID-19 related misinformation, including Meta's removal of

misinformation and third-party fact-checking program, as "generally positive" but "insufficient" to address the challenges posed by disinformation. The Special Rapporteur highlighted the need to undertake a "serious review of the business model that underpins much of the drivers of disinformation and misinformation." (A/HRC/47/25, paras 65-67.)

96. The Board is concerned that Meta has not conducted a human rights impact assessment on how its platforms' design features and current measures impact public health and human rights, such as the rights to life, health, access to information, and expression of ideas and views about the pandemic and related public health measures. Meta should make sure it has access to all the information required to properly assess potential human rights impacts. Given the disparities in access to adequate and accessible information, essential vaccines, medicines and treatments, and the resourcing of content moderation around the world, a human rights impact assessment is crucial to assess the risks arising from the spread of misinformation on COVID-19 capable of causing imminent physical harm in Meta's products globally.

97. **Recommendation 11:** *Meta should add a change log to the Help Center page providing the complete list of claims subject to removal under the company's* ***Misinformation about health during public health emergencies*** *policy. The Board will consider this recommendation implemented when a change log is added to the Help Center page.*

98. The Community Standards provide a change log to alert users to changes in the policies being enforced. However, the Help Center page that contains the specific claims subject to removal under the ***Misinformation about health during public health emergencies*** policy does not provide a change log or any means for users to determine when the list of claims has been updated or amended. Any additions or changes to the claims subject to removal are therefore difficult to track.

99. Meta informed the Board that between March 2020 and October 2022, various claims have been added to the list of claims removed under the ***Misinformation about health during public health emergencies*** policy and some claims have been removed or amended.

100. The UN Special Rapporteur on freedom of expression has stated that all individuals should have "genuine access to the tools of communication necessary to learn about the public health crisis" (A/HRC/44/49, para. 63(b)). Adding a change log to the Help Center would be consistent with the principle of legality, apprising users clearly when specific claims are removed. Greater transparency on how the list of claims evolves would benefit users as scientific consensus and understanding of the public health impact of the COVID-19 pandemic continues to evolve.

101. The addition of a change log in the Help Center would also help users with different views to challenge the assessment made by the public health authority on falsity or likelihood of directly contributing to the risk of imminent physical harm, in connection with recommendations one and two. This approach would address Meta's human rights responsibility with regard to public health, while enabling dissenting voices to contest claims with which they disagree.

102. **Recommendation 12:** *Meta should provide quarterly enforcement data on Misinformation in the Quarterly Enforcement Report, broken down by type of misinformation (i.e., physical harm or violence, harmful health misinformation, voter or census interference, or manipulated media) and country and language. This data should include information on the number of appeals and the number of pieces of content restored. The Board will consider this recommendation implemented when Meta starts including enforcement data on the Misinformation policy in the company's enforcement reports.*

103. The Community Standards Enforcement Report (CSER) that Meta releases every quarter shows how many pieces of content were actioned under the various Community Standards. However, this report does not contain any enforcement data on the company's Misinformation policy. The Board understands that this is partly because the Misinformation Community Standard was formally established in March 2022. Meta informed the Board that the company does not have data on COVID-19 misinformation prevalence on its platforms. According to Meta, this is the case due to evolving definitions of what constitutes COVID-19 misinformation as well as the difficulty in establishing a meaningful comparison between pre- and post-policy prevalence.

104. Meta, however, has been able to measure prevalence for brief periods of time for smaller subsets. According to the company, between March 1, 2022, and March 21, 2022, COVID-19-related content comprised 1-2% of views of Facebook posts in the United States. Among these views, Meta estimates that about 0.1% concerned content that violates the Misinformation and Harm policies.

105. The Board received many comments from stakeholders throughout the world highlighting that the lack of publicly available information on the number of pieces of content actioned under the Misinformation policy, among other relevant data points, undermines the ability of researchers and stakeholders to evaluate the effectiveness of Meta's existing responses to COVID-19 misinformation. Meta must provide data to evaluate whether the enforcement of the policy results in too many false positives and needs to be amended to lessen the risk of overenforcement. In this regard, the UN Special Rapporteur on freedom of expression has highlighted the "lack of transparency and access to data that hampers an objective assessment of the effectiveness of the measures" adopted to counter disinformation online. This also prevents stakeholders

from knowing whether policies have been applied consistently throughout the world. (A/HRC/47/25, para. 65).

106. The Board has previously recommended that Meta disaggregate its Community Standard Enforcement Report Data by country and language ("Punjab concern over the RSS in India," [2021-003-FB-UA] case decision, recommendation one). In response, Meta has committed to changing its metrics and has set the goal of launching them by the end of 2023. Disaggregating enforcement data by country or language is vital for understanding the scope of the problem in different parts of the world and the relative effectiveness of the company's enforcement measures. The Board and, importantly, Meta's stakeholders are prevented from fully and meaningfully understanding the effectiveness of the company's current global policy and enforcement approach to address COVID-19 misinformation absent any relevant data that would allow researchers and civil society to evaluate the company's efforts.

107. **Recommendation 13:** *Meta should create a section in its "Community Standards Enforcement Report" to report on state actor requests to review content for the policy on* ***Misinformation about health during public health emergencies*** *violations. The report should include the details on the number of review and removal requests by country and government agency, and the number of rejections and approvals by Meta. The Board will consider this implemented when Meta publishes a separate section in its "Community Standards Enforcement Report" information on requests from state actors that led to removal for this type of policy violation.*

108. In the "UK drill music" case [2022-007-IG-MR], the Board recommended that Meta "publish data on state actor content review and removal requests for Community Standard violations." At the height of the COVID-19 pandemic, concerns were raised about Meta reviewing COVID-19-related content at the behest of governments. This trend could be exacerbated in countries where governments make such requests to crack down on peaceful protestors or human rights defenders that criticize government policies or to silence public debate. During the pandemic, the UN Special Rapporteur on the rights to freedom of peaceful assembly and of association has raised concerns about governments around the world using the pandemic as a pretext to impose a state of emergency or otherwise circumvent due process requirements and institutional checks and balances inherent in a democratic society. This affected fundamental human rights, such as the right to peaceful protest (A/HRC/50/42, para. 18; A/77/171, para. 40, 67). A detailed report on state actor requests to review content under the policy on ***Misinformation about health during public health emergencies*** would provide due process to users, in line with the principle of legality, especially those in at-risk countries with weak civic spaces.

109. The UN Special rapporteur on the rights to freedom of peaceful assembly and of association recommended that technology companies must ensure that their products

are not used by governments to surveil or control "rights-supporting social movement activists" (A/77/171, para. 71). The Board commends Meta's commitment to empower human rights defenders against online harassment, surveillance, and censorship demands from governments as outlined in the company's Corporate Human Rights Policy. Transparency on government requests to review and/or remove content under Meta's Misinformation Community Standard would demonstrate this commitment.

110. **Recommendation 14:** *Meta should ensure existing research tools, such as CrowdTangle and Facebook Open Research and Transparency (FORT) continue to be made available to researchers. The Board will consider this recommendation implemented when Meta publicly states its commitment to sharing data through these tools to researchers.*

111. **Recommendation 15:** *Meta should institute a pathway for external researchers to gain access to non-public data to independently study the effects of policy interventions related to the removal and reduced distribution of COVID-19 misinformation, while ensuring these pathways protect the right to privacy of Meta's users and the human rights of people on and off the platform. This data should include metrics not previously made available, including the rate of recidivism around COVID-19 misinformation interventions. The Board will consider this recommendation implemented when Meta makes these datasets available to external researchers and confirms this with the Board.*

112. The UN Special Rapporteur on freedom of expression noted the difficulty in addressing disinformation, partly because of a lack of sufficiently publicly available information that would enable users, researchers and activists to understand and articulate the nature of the problem (A/HRC/47/25, para. 3, 67, 81). To this end, the UN Special Rapporteur on freedom of expression recommends making data available for "research, policymaking, monitoring and evaluation" (A/HRC/47/25, para. 104).

113. CrowdTangle is a tool that external researchers can use to track "influential public accounts and groups across Facebook and Instagram" and analyze relevant trends, including on misinformation. The tool's database includes all verified/public users, profiles, and accounts, such as those of politicians, journalists, media and publishers, celebrities, sports teams and other public figures. It also includes public groups and pages above a certain size threshold depending on the country. It shares data on the date and type of content posted, which page, account or group shared the content, the number and type of interactions with the content, and which other public pages or accounts shared it. It does not track the reach of the content, any data or content posted by private accounts, paid or boosted content, or the demographic information of the users that interact with the content. CrowdTangle has data on over seven million Facebook pages, groups and verified profiles and over two million public Instagram accounts.

114. In 2022, news reports claimed that Meta plans to end CrowdTangle. Although Meta has not publicly confirmed this, the Board emphasizes that research tools should be strengthened rather than discontinued by the company. This would enable external researchers to understand the impact of Meta's products, including on COVID-19 misinformation.

115. The Board notes Meta's efforts in setting-up the Facebook Open Research and Transparency (FORT) tool, which provides various privacy-protected data sets to academics and researchers. According to the company, the FORT's Researcher Platform grants social scientists access to sensitive information in a controlled environment concerning "large-scale behavioral data in order to study and explain social phenomena." However, reports have surfaced of the tool's shortcomings for academic research, such as Meta's "restrictive terms of use" or researchers being provided with insufficient information to undertake meaningful analysis. That said, the Board recognizes that in the context of other social media companies, Meta has taken significant steps to share data with external researchers and the Board encourages Meta to do even more.

116. Throughout the stakeholder engagement activities conducted for this policy advisory opinion, researchers have repeatedly reinforced the need for these tools to track trends related to COVID-19 misinformation. The lack of access to relevant data has also created challenges for the Board when assessing the merits of the policy advisory opinion request. To the Board's understanding, some of this data is not available to the company itself, whereas some is available but cannot be shared with external stakeholders, including the Board. Meta should provide researchers with access to relevant data to track COVID-19 misinformation prevalence on the platform and the effectiveness of specific measures to address it. Such information is also essential to the conduct of the human rights impact assessment discussed above.

117. **Recommendation 16:** *Meta should publish the findings of its research on neutral and fact-checking labels that it shared with the Board during the COVID-19 policy advisory opinion process. The Board will consider this recommendation implemented when Meta publishes this research publicly in its Transparency Center.*

118. The Board appreciates the information that Meta shared with the Board on the effectiveness of NITs and demotion of content, such as the results of the experiments the company ran to evaluate the continued effectiveness of NITs. The Board is of the view that the findings of these experiments should be shared more widely with external researchers seeking to understand the impact of the company's responses to COVID-19 misinformation.

119. **Recommendation 17:** *Meta should ensure equitable data access to researchers around the world. While researchers in Europe will have an avenue to apply for data*

access through the Digital Services Act (DSA), Meta should ensure it does not over-index on researchers from Global North research universities. Research on prevalence of COVID-19 misinformation and the impact of Meta's policies will shape general understanding of, and future responses to, harmful health misinformation in this and future emergencies. If that research is disproportionately focused on the Global North, the response will be too. The Board will consider this recommendation implemented when Meta publicly shares its plan to provide researchers around the world with data access similar to that provided to EU countries under the DSA.

120. A majority of the research on disinformation flows disproportionately reflects trends and patterns that occur in the U.S. and Western Europe. This could lead to content policy interventions being framed according to the specific problems of these geographies. The UN Special Rapporteur on freedom of expression noted calls from civil society for further research on the impact of disinformation on "vulnerable and minority communities," citing the identity-based disinformation campaigns that fueled ethnic conflict in Ethiopia and Myanmar (A/HRC/47/25, para. 26). As Meta expands access to external researchers in compliance with the Digital Services Act as well as maintaining its own FORT, the company should ensure representation of academics and researchers from around the world.

121. **Recommendation 18:** *Meta should evaluate the impact of the cross-check Early Response Secondary Review (ERSR) system on the effectiveness of its enforcement of the Misinformation policy and ensure that Recommendations 16 and 17 in the Board's policy advisory opinion on Meta's cross-check program apply to entities that post content violating the* **Misinformation about health during a public health emergency** policy. *The Board will consider this recommendation implemented when Meta shares its findings with the Board and publicizes it.*

122. According to Meta, the cross-check program was put in place to minimize the highest-risk false-positive moderation errors. The first part of the cross-check program is the Early Response Secondary Review (ERSR) system that guarantees additional human review of potentially violating content posted by specific entitled entities. Meta maintains lists of entitled entities based on who the company has decided is entitled to receive the benefits of ERSR provides. Entities can take the form of Facebook pages, Facebook profiles, and Instagram accounts and can represent individual persons and groups or organizations. Many of the users included in these lists are celebrities, major companies, government leaders, and politicians.

123. In its policy advisory opinion on Meta's cross-check program, the Board recommended that Meta establish "clear and public criteria for entity-based mistake-prevention eligibility" that would differentiate between "users whose expression merits additional protection from a human rights perspective" and those "included for business reasons."

124. Politicians are exempt from the company's third-party fact-checking program. This means that false information shared by politicians, which is not otherwise removed under the Harmful Health Misinformation policy, cannot be reviewed and labeled by third-party fact-checkers. Meta also explained that it has not evaluated the impact of the ERSR system on the effectiveness of the COVID-19 Misinformation policy as the ERSR was set up to prevent mistakes in enforcement rather than to evaluate the effectiveness of a specific Community Standard. As such, the company's internal teams do not track or analyze data on its impact on the COVID-19 Misinformation policy.

125. Based on internal research as well as stakeholder engagement for this COVID-19 policy advisory opinion, the Board found that many spreaders of misinformation are prominent speakers such as celebrities, politicians, state actors, and religious figures who may be entities entitled to the benefits of the ERSR program. For example, the submission from Media Matters for America (PC-10758) called attention to the impact of Meta's cross-check system in undermining efforts to address misinformation. As celebrities, politicians, journalists and other prominent users were "afforded slower or more lenient enforcement" for content violations, misinformation was allowed to remain on the platform. The UN Special Rapporteur on freedom of expression similarly raised concerns about "unreliable information" disseminated by "individuals with significant platforms," noting that this "can cause grave harm, whether maliciously intended or not." State actors have also disseminated "often reckless claims" about the origins of the virus, the availability of drugs to counter symptoms, and the state of COVID-19 in their country, among others. The UN Special Rapporteur recommended holding public officials accountable for their statements and actions. (A/HRC/44/49, para. 41, 45, 63(c); see also A/HRC/47/25, para. 18 identifying celebrities as purveyors of misinformation more generally.)

126. Misinformation posted by entitled entities under the ERSR program that is likely to directly contribute to the risk of imminent physical harm to public health and safety will be removed consistent with the Harmful Health Misinformation policy. However, misinformation that would ordinarily merit a fact-check or label, but because it is posted by entitled entities such as politicians is not only exempt from third-party fact-checking but also benefits from delayed enforcement due to additional human review of potentially violating content provided by the ERSR system. This means that COVID-19 misinformation that is not one of the defined 80 claims posted by entitled entities can stay on the platform without a fact-check label and may not be reviewed at all.

***Procedural note:**

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

For this policy advisory opinion, independent research was commissioned on behalf of the Board. The Board was assisted by an independent research institute headquartered at the University of Gothenburg which draws on a team of over fifty 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. The Board was also assisted by Memetica, an organization that engages in open-source research on social media trends, which also provided analysis.