Facebook's response to the Oversight Board's first decisions

Breast cancer symptoms and nudity

OVERSIGHT BOARD'S RECOMMENDATION | **FACEBOOK'S RESPONSE**

Revise the Instagram Community Guidelines around adult nudity. Clarify that the Instagram Community Guidelines are interpreted in line with the Facebook Community Standards, and where there are inconsistencies the latter take precedence.

Committed to action

OUR COMMITMENT

In response to the board's recommendations, we updated the Instagram Community Guidelines on nudity to read: "...photos in the context of breastfeeding, birth-giving and after-birth moments, health-related situations (for example, post-mastectomy, breast cancer awareness or gender confirmation surgery) or an act of protest are allowed."

We'll also clarify the overall relationship between Facebook's Community Standards and Instagram's Community Guidelines, including in the Transparency Center we'll be launching in the coming months (see hydroxychloroquine, azithromycin and COVID-19 recommendation 2 for more detail).

CONSIDERATIONS

Our policies are applied uniformly across Facebook and Instagram, with a few exceptions — for example, people may have multiple accounts for different purposes on Instagram, while people on Facebook can only have one account using their authentic identity. We will update Instagram's Community Guidelines to provide additional transparency about the policies we enforce on the platform. Our teams will need some time to do this holistically (for example, ensuring the changes are reflected in the notifications we send to people and in our Help Center), but we'll provide updates on our progress.

NEXT STEPS

We'll build more comprehensive Instagram Community Guidelines that provide additional detail on the policies we enforce on Instagram today and provide people with more information on the relationship between Facebook's Community Standards and Instagram's Community Guidelines.

When communicating to users about how they violated policies, be clear about the relationship between the Instagram Community Guidelines and Facebook Community Standards.

Committed to action

OUR COMMITMENT

We'll continue to explore how best to provide transparency to people about enforcement actions, within the limits of what is technologically feasible. We'll start with ensuring consistent communication across Facebook and Instagram to build on our commitment above to clarify the overall relationship between Facebook's Community Standards and Instagram's Community Guidelines.

CONSIDERATIONS

Over the past years we've invested in improving the way we communicate with people when we remove content, and we have teams dedicated to continuing to research and refine these user experiences. As part of this work, we've updated our notifications to inform people under which of Instagram's Community Guidelines a post was taken down (for example, was it taken down for Hate Speech or Adult Nudity & Sexual Activity), but we agree with the board that we'd like to provide more detail. As part of our response to the recommendation in the case about Armenians in Azerbaijan, we are working through multiple considerations to explore how we can provide additional transparency.

In addition to confirming the need to provide more specificity about our decisions, the board's decision also highlighted the need for consistency in how we communicate across Facebook and Instagram. In this case, we did not tell the user that we allow female nipples in health contexts, but the same notification on Facebook would have included this detail. As we clarify the overall relationship between Facebook's Community Standards and Instagram's Community Guidelines, we commit to ensuring our notification systems keep up with those changes.

NEXT STEPS

We will continue to work toward consistency between Facebook and Instagram and provide updates within the next few months.

Improve the automated detection of images with text-overlay to ensure that posts raising awareness of breast cancer symptoms are not wrongly flagged for review.

Committed to action

OUR COMMITMENT

We agree we can do more to ensure our machine learning models don't remove the kinds of nudity we allow (e.g., female nipples in the context of breast cancer awareness). We commit to refining these systems by continuing to invest in improving our computer vision signals, sampling more training data for our machine learning, and leveraging manual review when we're not as confident about the accuracy of our automation.

CONSIDERATIONS

Facebook uses both: 1) automated detection systems to flag potentially violating content and "enqueue" it for a content reviewer, and 2) automated enforcement systems to review content and decide if it violates our policies. We want to avoid wrongfully flagging posts both for review and removal, but our priority will be to ensure our models don't remove this kind of content (content wrongfully flagged for review is still assessed against our policies before any action is taken).

In this case, our automated systems got it wrong by removing this post, but not because they didn't recognize the words "breast cancer." Our machine learning works by predicting whether a piece of content violates our policies or not, including text overlays. We have observed patterns of abuse where people mention "breast cancer" or "cervix cancer" to try to confuse and/or evade our systems, meaning we cannot train our system to, say, ignore everything that says "breast cancer."

So, our models make predictions about posts like breast cancer awareness after "learning" from a large set of examples that content reviewers have confirmed either do or do not violate our policies. This case was difficult for our systems because the number of breast cancer-related posts on Instagram is very small compared to the overall number of violating nudity-related posts. This means the machine learning system has fewer examples to learn from and may be less accurate.

NEXT STEPS

We will continue to invest in making our machine learning models better at detecting the kinds of nudity we do allow. We will continue to improve computer vision signals, sampling more training data for our machine learning, and increase our use of manual review when we're less sure about the accuracy of our automation.

Ensure users can appeal decisions taken by automated systems to human review when their content is found to have violated Facebook's Community Standard on Adult Nudity and Sexual Activity.

Assessing feasibility

OUR COMMITMENT

Our teams are always working to refine the appropriate balance between manual and automated review. We will continue this assessment for appeals, evaluating whether using manual review would improve accuracy in certain areas, and if so how best to accomplish it.

CONSIDERATIONS

Typically, the majority of appeals are reviewed by content reviewers. Anyone can appeal any decision we make to remove nudity, and that appeal will be reviewed by a content reviewer except in cases where we have capacity constraints related to COVID-19.

That said, automation can also be an important tool in re-reviewing content decisions since we typically launch automated removals only when they are at least as accurate as content reviewers.

NEXT STEPS

We'll continue to monitor our enforcement and appeals systems to ensure that there's an appropriate level of manual review and will make adjustments where needed.

Inform users when automation is used to take enforcement action against their content, including accessible descriptions of what this means.

Assessing feasibility

OUR COMMITMENT

Our teams will test the impact of telling people whether their content was actioned by automation or manual review.

CONSIDERATIONS

Over the past several years we've invested in improving the experience that we provide people when we remove content. We have teams who think about how to best explain our actions and conduct research to help inform how we can do this in a way that's accessible and supportive to people. We also need to ensure that this experience is consistent across billions of people all over the world, with differing levels of comprehension. From prior research and experimentation, we've identified that people have different perceptions and expectations about both manual and automated reviews. While we agree with the board that automated technologies are limited in their ability to understand some context and nuance, we want to ensure that any additional transparency we provide is helping all people more accurately understand our systems, and not instead creating confusion as a result of pre-existing perceptions. For example, we typically launch automated removal technology when it is at least as accurate as content reviewers. We also don't want to overrepresent the ability of content reviewers to always get it right.

Additionally, many decisions made are a combination of both manual and automated input. For example, a content reviewer may take action on a piece of content based on our Community Standards, and we may then use automation to detect and enforce on identical copies. We would need to research to identify the best way of explaining these and other permutations to people.

NEXT STEPS

We will continue experimentation to understand how we can more clearly explain our systems to people, including specifically testing the impact of telling people more about how an enforcement action decision was made.

Expand transparency reporting to disclose data on number of automated removal decisions, and the proportion of those decisions subsequently reversed following human review.

Assessing feasibility

OUR COMMITMENT

We need more time to evaluate the right approach to share more about our automated enforcement. Our Community Standards Enforcement Report currently includes our "proactive rate" (the amount of violating content we find before people report it), but we agree that we can add more information to show the accuracy of our automated review systems.

CONSIDERATIONS

The board uses the term "automation" broadly, however many decisions are made with a combination of both manual and automated input. For example, a content reviewer may take action on a piece of content based on our Community Standards, and we may then use automation to detect and enforce on identical copies. We need to align on the best way to study and report this information.

NEXT STEPS

We'll continue working on this recommendation and the most appropriate and meaningful metrics reported in our Community Standards Enforcement Report that take into account the complexities of scale, technology, and manual review.

Hydroxychloroquine, azithromycin and COVID-19

OVERSIGHT BOARD'S RECOMMENDATION | **FACEBOOK'S RESPONSE**

Clarify the Community Standards with respect to health misinformation, particularly with regard to COVID-19. Facebook should set out a clear and accessible Community Standard on health misinformation, consolidating and clarifying existing rules in one place.

Committed to action

OUR COMMITMENT

In response to the board's recommendation, we have consolidated information about health misinformation in a Help Center article, which we now link to in the Community Standards. This article includes details about all of our Community Standards related to COVID-19 and vaccines, including how we treat misinformation that is likely to contribute to imminent physical harm. We also added a "Commonly Asked Questions" section to address more nuanced situations (e.g. how humor and satire relate to these policies, how we handle personal experiences or anecdotes).

We have also clarified our health misinformation policy as part of a larger COVID-19 update earlier this month. As part of that update, we added more specificity to our rules, including giving examples of the type of false claims that we will remove.

CONSIDERATIONS

Our policies and principles for enforcement of health misinformation are continuously updated to reflect the feedback we get from our global conversations with health experts.

NEXT STEPS

We'll continue to update the Help Center as necessary as our policies evolve with the pandemic.

Facebook should 1) publish its range of enforcement options within the Community Standards, ranking these options from most to least intrusive based on how they infringe freedom of expression, 2) explain what factors, including evidence-based criteria, the platform will use in selecting the least intrusive option when enforcing its Community Standards to protect public health, and 3) make clear within the Community Standards what enforcement option applies to each rule.

Committed to action

OUR COMMITMENT

In the coming months, we will launch the Transparency Center. The website will be a destination for people to get more information about our Community Standards and how we enforce them on our platform, including when and why we remove violating content, and when we choose to provide additional context and labeling.

CONSIDERATIONS

As our content moderation practices have grown in sophistication and complexity, our efforts to provide people with comprehensive but clear information about our systems have to catch up. The Transparency Center is a step in this effort, building on our Community Standards to help people understand our integrity efforts overall. The Transparency Center will add more detail about what isn't allowed, as well as how we use interventions like downranking and labels for content that we think may benefit from more context.

NEXT STEP

Launch the Transparency Center in the coming months.

To ensure enforcement measures on health misinformation represent the least intrusive means of protecting public health, Facebook should clarify the particular harms it is seeking to prevent and provide transparency about how it will assess the potential harm of particular content.

Committed to action

OUR COMMITMENT

In response to the board's guidance, we updated our Help Center to provide greater detail on the specific harms that our COVID-19 and vaccine policies are intended to address. The Help Center explains that we will "remove misinformation when public health authorities conclude that the information is false and likely to contribute to imminent violence or physical harm." As noted in the Help Center, some of these examples of imminent physical harm include "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."

CONSIDERATION

For COVID-19, we assessed harm by working closely with public health authorities, who are better equipped to answer the complex question of causality between online speech and offline harm. We also consulted with experts from around the world with backgrounds in public health, vaccinology, sociology, freedom of expression, and human rights on updates we made to our policies on vaccine misinformation. These experts came from academia, civil society, public health organizations, and elsewhere. We rely on these experts to help us understand whether claims are false and likely to contribute to the risk of increased exposure and transmission or to adverse effects on the public health system. We then remove content that includes these claims.

NEXT STEPS

We won't take any additional actions since based on the board's recommendation we've already updated our Help Center.

To ensure enforcement measures on health misinformation represent the least intrusive means of protecting public health, Facebook should conduct an assessment of its existing range of tools to deal with health misinformation and consider the potential for development of further tools that are less intrusive than content removals.

Committed to action

OUR COMMITMENT

We will continue to develop a range of tools to connect people to authoritative information as they encounter health content on our platforms, starting with information about COVID-19 vaccines.

CONSIDERATIONS

We continually assess and develop a range of tools, in consultation with public health experts, to address potential health misinformation in the least intrusive way depending on the risk of imminent physical harm. Our current range of enforcement tools include:

- Working with independent third-party fact-checking partners to debunk claims that are found to be false, but do not violate our Community Standards. Once third-party fact-checkers rate something as false, we reduce its distribution and inform people about factual information from authoritative sources.
- Sending notifications to people who shared false content to let them know it's since been rated false. We add a notice and an overlay to the post and show a fact-checker's articles when someone tries to share the content.
- Connecting people to authoritative information based on their behavior. For example, if someone searches for "COVID-19" or "vaccines," we will redirect them to our COVID-19 Info Center on Facebook. And, we may show educational modules to people who we know have interacted with misinformation we removed for violating our Community Standards.

These tools are part of our larger effort to respond proportionally to content, as the board recommends, while keeping people safe on the platform.

NEXT STEPS

Our immediate focus for this recommendation is to work on tools to connect people with authoritative information about COVID-19 vaccines.

In cases where users post information about COVID-19 treatments that contradicts the specific advice of health authorities and where a potential for physical harm is identified but is not imminent, Facebook should adopt a range of less intrusive measures.

No further action

OUR COMMITMENT

We agree with the board that less intrusive measures should be used where a potential for physical harm is identified but is not imminent. That said, we disagree with the board that the content implicated in this case does not rise to the level of imminent harm. We will continually evaluate and calibrate our response to content about COVID-19 treatments based on information from public health authorities.

CONSIDERATIONS

Our global expert stakeholder consultations have made it clear that, that in the context of a health emergency, the harm from certain types of health misinformation does lead to imminent physical harm. That is why we remove this content from the platform. We use a wide variety of proportionate measures to support the distribution of authoritative health misinformation. We also partner with independent third-party fact-checkers and label other kinds of health misinformation.

We know from our work with the World Health Organization (WHO) and other public health authorities that if people think there is a cure for COVID-19 they are less likely to follow safe health practices, like social distancing or mask-wearing. Exponential viral replication rates mean one person's behavior can transmit the virus to thousands of others within a few days.

We also note that one reason the board decided to allow this content was that the person who posted the content was based in France, and in France, it is not possible to obtain hydroxychloroquine without a prescription. However, readers of French content may be anywhere in the world, and cross-border flows for medication are well established. The fact that a particular pharmaceutical item is only available via prescription in France should not be a determinative element in decision-making.

NEXT STEPS

We'll take no further action on this recommendation since we believe we already do employ the least intrusive enforcement measures given the likelihood of imminent harm. We restored the content based on the binding power of the board's decision. We will continue to rely on extensive consultation with leading public health authorities to tell us what is likely to contribute to imminent physical harm. During a global pandemic, this approach will not change.

Publish a transparency report on how the Community Standards have been enforced during the COVID-19 global health crisis.

Committed to action

OUR COMMITMENT

We will continue to look for ways to communicate the efficacy of our efforts to combat COVID-19 misinformation.

CONSIDERATIONS:

We regularly publish information on the efforts we are taking to combat COVID-19 misinformation. For example, we have previously shared detailed data points on our response to COVID-19 misinformation, including the number of pieces of content on Facebook and Instagram we removed for violating our COVID-19 misinformation policies, the number of warning labels applied to content about COVID-19 that was rated by independent third-party fact-checkers, the number of visits to the COVID-19 Information Hub, and the number of people who clicked through these notifications to go directly to the authoritative health sources. We have also shared information with the EU Commission's COVID-19 monitoring programme reports.

NEXT STEPS

We began consistently sharing COVID-19 metrics in the Spring of 2020, and we will continue to do so for the duration of the pandemic. Given the temporary and unique circumstances of COVID-19, we are not planning to add it into the Community Standards Enforcement Report as an additional policy area.

Conduct a human rights impact assessment with relevant stakeholders as part of its process of rule modification.

Committed to action

OUR COMMITMENT

We will ask the board to clarify if its recommendation relates to all rule modifications or those related to COVID-19 misinformation. We will explore approaches to strengthen the incorporation of human rights principles into our policy development process.

CONSIDERATIONS

Facebook has a dedicated Human Rights Policy Team that consults on policy development and rule changes. Given the frequency with which we update our policies conducting a full human rights impact assessment for every rule change is not feasible.

The Human Rights Policy Team, informed by authoritative guidance and an independent literature review, advised on access to authoritative health information as part of the right to health and on permissible restrictions to freedom of expression related to public health. It also participated in structuring an extensive global rights holder consultation. These elements were directly incorporated into Facebook's overall strategy for combating misinformation that contributes to the risk of imminent physical harm.

NEXT STEPS

We will ask the board to clarify if its recommendation relates to all rule modifications or those related to COVID-19 misinformation. Based on this, we will assess whether there are opportunities to strengthen the inclusion of human rights principles in our policy development process, including the possibility of additional formal human rights impact assessments.

Armenians in Azerbaijan

OVERSIGHT BOARD'S RECOMMENDATION | **FACEBOOK'S RESPONSE**

Go beyond the Community Standard that Facebook is enforcing, and add more specifics about what part of the policy they violated.

Assessing feasibility

OUR COMMITMENT

We will continue to explore how best to provide transparency to people about enforcement actions, within the limits of what is technologically feasible.

CONSIDERATIONS

Over the past several years, we've invested in improving the experiences for people when we remove their content, and we have teams dedicated to continuing to improve these. As part of this work, we updated our notifications to inform people under which Community Standard a post was taken down (for example, Hate Speech, Adult Nudity & Sexual Activity, etc.), but we agree with the board that we'd like to provide more.

When a content reviewer reviews a post and determines it violates a policy, they often provide some additional data to our systems about the type of violation, but not always to the granularity of each line in the policy. Additionally, when we build technology to take automated action, it is often at the level of a policy area (e.g., Hate Speech) as it is not technologically feasible to create separate AI systems for each individual line in the policy. We understand the benefit in additional detail and will continue to explore how best to provide additional transparency.

NEXT STEPS

Our teams will continue to explore potential ways to address this challenge. We will provide updates with any future developments.

Nazi quote

OVERSIGHT BOARD'S RECOMMENDATION | **FACEBOOK'S RESPONSE**

Ensure that users are always notified of the Community Standards Facebook is enforcing.

Committed to action

OUR COMMITMENT

We've fixed the mistake that led to the user not being notified about the Community Standard used for our enforcement action.

CONSIDERATIONS

People should be able to understand our decisions when we take action on their content. This is why we've worked to ensure a consistent level of detail is provided when content is removed from our platforms, specifically by referencing at least the Community Standard or Community Guideline in question.

NEXT STEPS

After the board surfaced this issue, we fixed the mistake.

Explain and provide examples of the application of key terms used in the Dangerous Individuals and Organizations policy. These should align with the definitions used in Facebook's Internal Implementation Standards.

Committed to action

OUR COMMITMENT

We commit to adding language to the Dangerous Individuals and Organizations Community Standard clearly explaining our intent requirements for this policy. We also commit to increasing transparency around definitions of "praise," "support," and "representation."

CONSIDERATIONS

Facebook agrees with the board that we can be clearer about how we define concepts like "praise," "support" and "representation," and we're committed to increasing transparency here. Ahead of sharing more details about these terms, we need to ensure that this information doesn't inadvertently allow bad actors to circumvent our enforcement mechanisms. Over the next few months, our teams will determine the best way to explain these terms and how they are used in our policy.

NEXT STEPS

We will add language to our Dangerous Individuals and Organizations Community Standard within a few weeks explaining that we may remove content if the intent is not made clear. We will also add definitions of "praise," "support" and "representation" within a few months.

Provide a public list of the organizations and individuals designated "dangerous" under the Dangerous Individuals and Organizations Community Standard.

Assessing feasibility

OUR COMMITMENT

We commit to increasing transparency around our Dangerous Individuals and Organizations Policy. In the short term, we will update the Community Standard and link to all of our Newsroom content related to Dangerous Individuals and Organizations so that people can access it with one click.

CONSIDERATIONS

Ahead of sharing more details about these terms, we need to ensure that this information will not allow bad actors to circumvent our enforcement mechanisms.

Our teams need more time to fully evaluate whether sharing examples of designations will help people better understand our policy, or if we should publish a wider list. Before publishing, we also have to be confident it will not jeopardize the safety of our employees.

NEXT STEPS

We will update the link in the Community Standards within a few weeks. We will continue to work toward more clarity on our Dangerous Individuals and Organizations policies while protecting the safety of our employees and platform.