Quarterly Integrity Update Press Call
May 17, 2022
11:30 a.m. ET

Operator: Hello and welcome to Meta's Quarterly Integrity Update Call. There will be prepared remarks and a Q&A to follow. To ask a question after the prepared remarks conclude, please press the “1” followed by the “4” on your telephone.

As a reminder, this conference is being recorded Tuesday, May 17th, 2022.

Now I'd like to turn the call over to Carolyn Glanville, who will kick this off. Please go ahead.

Carolyn Glanville: Thank you so much, and thank you all for joining us. You should have received embargoed material ahead of this call with our community standards enforcement report, widely viewed content report, the Oversight Board quarterly report, and the transparency report for the second half of 2021, as well as the results of the EY assessment of our Q4 2021 community standards enforcement report.

To kick off our call today, you'll hear from Vice President of Integrity, Guy Rosen; Vice President of Content Policy, Monika Bickert; and Director of Product Management, Anna Stepanov.

We will then open the call for questions. This call is on the record and it's embargoed until 1:00 PM Eastern, 10:00 AM Pacific.

With that, I'll kick it over to Guy.

Guy Rosen: Thank you, Carolyn. Hey everyone, I am Guy Rosen and I lead the product and engineering team that work on safety and security.

I'd like to start first today with the community standards enforcement report. We use this report and this call as a quarterly touchstone to update you all about our progress in this work.

Now for a number of years, we've been reporting on the metrics here which are the same ones that we use internally. We believe prevalence is the best way to hold us accountable for this work, as it measures the views of content that violate our policies. Violating posts might be seen by people because we missed it altogether, or we did catch it but not quickly enough.

In this report, across most of our policy areas, we have seen prevalence levels relatively stable in most areas. This means that the vast majority of the content that people see on our platforms is not in violation of our policies. So for example, hate speech prevalence on Facebook, which we first started reporting in November of 2020, was then between 0.10 to 0.11 percent; in this most recent report for Q1 of 2022, it was 0.02 percent, so just a little lower than last quarter.

We also continued to see a slight decrease in the prevalence of (bullying harassment) on Facebook, down from between 0.11 to 0.12 percent in Q4 to 0.09 percent in Q1 of this year.

In one area, adult nudity, prevalence slightly increased on Facebook, from 0.03 percent in Q4 to 0.04 percent in Q1. This is due to an increase in spam actors who shared large volumes of videos that contain nudity. We've since taken additional measures to combat these kinds of behaviors.

Next I'd like to talk about the independent assessment of the report. And when you think about any report like this and others, there's two things to consider. First, are we asking the right questions and second, are our answers correct?

We started this process in 2018 we began tackling the first, are these the right questions. And in 2018, that same year, a group of experts in statistics, law, economics and governance published an independent assessment of the methodology that we use here. Are these the right ways to measure these kind of issues.

Two years ago we also committed to address the second part, are the answer correct. By having these metrics reviewed for accuracy because no company should be grading their own homework. In the past two years we engaged with EY for this assessment. We provided EY with an in-depth understanding of our processes, our systems, the controls we have. We also provided them with the data and the evidence they requested in order to conduct the assessment.

And we welcome that they have concluded that we presented these metric accurately and we have the right internal controls in place to ensure accuracy. As we keep growing this report we will also keep working on way to make sure it is independently verified.

An independent third-party assessment like this demonstrates the commitment we have to these reports and this approach. While we are the first of our peer companies to undertake this kind of assessment we believe these should be standard and more companies should pursue similar verification.

Now let me turn it over to Monika.

Monika Bickert: Thanks, Guy. And hi everyone, thanks for joining today. I am Monika Bickert, VP of Content Policy at Meta. I lead the team that writes our policies on what content is and is not allowed on our platforms.

It's an international team, they work every day in partnership with local experts and organizations to identify how we can best manage our policies and keep up with new shifts in global dynamics.

I want to cover a few things today starting with the steps that we're taking to consolidate the information that we share publicly to make it easier to access and understand. As part of this effort we're creating one central page for the public regulatory reports that we share around the world including reports we've published in India, Germany, Austria, Turkey and E.U.

To be clear, these reports are already publicly available on our website but we'll be creating a dedicated page in our transparency center to make it easier for people to access all of these reports and to access them by region.

We're also releasing bi-annual transparency report which includes the number of requests we have received from governments for user data, content restrictions based on local law, service disruptions, and intellectual property takedowns. Now this report covers July 1, 2021 through December 31, 2021. So any requests pertaining to the war in Ukraine will be reflected in our next report.

Next, I wanted to highlight a few elements of the quarterly update we're releasing today on the oversight board. As you may know, in addition to the (binding) decisions they've issued, all of which we've complied with, they've also issued over 100 recommendations for our polices and processes.

We responded to all of those recommendations publicly and this report includes an update on our response to 55 of them including details on recent changes in actions.

Just to highlight a few, we've initiated two policy reviews which will likely result at meetings in the policy forum to consider specific policy changes. We've undertaken several new research projects to better understand how we can incorporate user voice into our appeals and review processes. And we've translated the community standards into additional languages, include Assamese and Farsi.

People around the world have broad ranging views on how to limit online speech, including how to balance freedom of expression and safety and who should draw those lines. The oversight board provides and independent sources of guidance on those important issues, and we look forward to updating you in the future on our progress and the additional recommendations that we expect to receive from the oversight board.

With that, I'll turn it over to Anna to talk to the Widely Viewed Content Report.

Anna Stepanov: Thanks, Monika. Hi, everyone. I'm Anna Stepanov, and I lead the Facebook Integrity Team. So today we're publishing the Widely Viewed Content Report for the first quarter of 2022. This report highlights the most viewed organic content in feed, including domains, links, pages, and posts. It includes content recommended by Facebook and excludes advertising content.

So I want to talk a little about how we are operationalizing these reports. One of the reasons for understanding this data and releasing these reports is to help improve our product. For example, we've seen promising results from our test to reduce engagement bait, including expanding our signals and introducing spacing roles to help prevent multiple posts that are identified as engagement-based from showing up one after the other in feed.

We anticipate that these changes will lead to a reduction in flow quality content, but expect it'll take several reporting cycles for these changes to make a noticeable impact in the quarterly data.

In the last quarter, we saw the quality of our posts engagements improve but did still see some lower quality links get widespread distribution. We will continue to test alternative solutions to reduce engagement bait and other problematic content in feeds.

So since releasing the inaugural WVCR, we have engaged with academics and experts to identify the parts based on valuables, which metrics needed more context and how we can best support their understanding of content distribution on Facebook.

Based on these discussions, we're improving our link and domain data methodologies. Previously, we counted a link view any time a post or video containing a link was viewed, even if the link was not front and center. Moving forward, links will need to run their preview in order to be counted as a view as that more accurately represents what people are seeing.

As part of the transition, the Q1 2022 report includes top viewed links using both our old and new methodologies. Starting next quarter, the WVCR will use only (the new ones).

So as we've seen in previous reports, some lower quality posts had widespread distribution last quarter. Although it's important to note that the top 20 links in this report represent only 0.3 percent of all feed content views in the U.S. during the quarter.

The fourth linked URL was a YouTube video of a panel discussion held by a U.S. Senator that was rated false by one of our fact check partners. When that happened, we took a number of steps to limit the reach of this link, including adding a warning screen that covered content with a link with more information about the claim, showing a notification warning to someone when they tried to share the link, and reducing the distribution of the link in feed.

Without these features, this feed would likely have received even more reach and people who viewed it would not have seen additional information and context from the false fact check.

Lastly, in this report there were pieces of content that have since been removed from Facebook for violating our policy of inauthentic behavior. The removed links were all from the same domain, and links from that domain are no longer allowed on Facebook.

When we shared the Q4 2021 WVCR in February, we were rightly pressed to share information on a link that had been removed from the report. We understand the expectation that we would be more transparent about our more – most widely viewed content, even that content which has been removed from Facebook.

We've taken the feedback we've received seriously and we'll attempt to disclose as much information as possible moving forward. We've updated the report and the companion guide to explain our updated removal disclosure framework.

However, we want to be clear that at times preventing additional harm to our community will outweigh disclosing specific details on removed content. Put simply we don't want to direct traffic to come that violates our community standards.

And thanks everyone for joining the call and with that I'll turn it over to the operator for questions.

Operator: Thank you. We will now open the line for questions. To ask a question please press the "1" followed by the "4" on your telephones. Our first question comes from the line of Brian Fung with CNN. Please proceed with your question.

Brian Fung: Hi, folks. Two if I may. You noted in the report that the oversight board (related to the) one content referral related to Russia's invasion of Ukraine. Can you say a little bit more about the nature of that request and the question underlying the referral?

And then second, I wonder if you can talk a little bit about how you're thinking about Texas' new content moderation law HB 20 and there's a lot of implications both your policies and (a portion of those) policies. Thanks.

Monika Bicker: Thanks, Brian, for the question. And I'll start with Ukraine and just say that we assessed that going through with our policy advisory opinion referral to the oversight board on that subject prevents an ongoing safety and security concern. And we need to take that into account. I really can't comment beyond that on the Ukraine referral.

On the question of the Texas law, we are of course watching that case and we understand that the Supreme Court is considering an appeal right now on the state of the injunction and we are watching that to see what happens going forward. Thanks.

Operator: Thank you. Our next question comes from Alex Heath with The Verge. Please proceed with your question.

Alex Heath: Hey, thanks for taking the question. Mines a little bit future looking. I'm curious what the shift in strategy that Mark talked about on the last earnings call to what he called the discovery engine for the feed and video primarily taking up more and more time in the BLUE App, especially (inaudible).

Does that change how you guys approach integrity? I assume it presents some new challenges. I'd be curious to hear from Guy or anyone about how you have to kind of adapt to that environment that's video and more kind of (probilistic) in the way I.A. works versus maybe (deterministic)? Thanks.

Guy Rosen: Hi, Alex. Thanks for the question. So generally speaking, the work that's happening to build these new product experiences, like any new product the company builds, we do it with integrity teams really embedded in the work from the start.

We have teams, not just a sort of one central team but we also have teams such as Anna's team, which is embedded in the Facebook app, which partners with folks that are building these new experiences so that we do think about these things as we – as we go about this work.

Of course, any content, including content such as in real surfaces whether it's on Facebook or Instagram, we have our community standards. We have recommendation policies for content, which is of this nature on connected content that people are being offered up.

We are definitely working in partnership across all integrity teams to ensure that we're bringing our best foot forward applying all of the technologies that we've built and the systems that we've created over the years to ensure that we are monitoring, that we are filtering out content appropriately so that we uphold our policies and build the best product experience for folks that are using these new experiences.

Operator: Thank you. Our next question comes from the line of Glenn Chapman with AFP. Please proceed with your question.

Glenn Chapman: Thank you. Good morning, everyone. Since this is a broad policy discussion, I very much appreciate your thoughts. Obviously there's been a big focus on the degree to which content should be moderated with the current effort by someone to buy Twitter, and I'm wondering given that this discussion has focused on your approach to content moderation and policies and transparencies what your thoughts are on Musk's recommended approach of just minimally moderating content just to a legal standard and doing away with the expense and the controversy that comes with mitigating content?

Monika Bickert: Thanks. The question of how to best provide a service that lets people express themselves and connect with one another while at the same time doing what we can to protect safety and dignity and privacy, those are the issues that we've grappled with from the beginning of this service, and I've been – I've been managing the content policies now for almost 10 years.

And from day one it's always been how do we balance this. And we do want to maximize expression. We want to create as much room as possible for people to connect with one another even if – even if they're saying some things that are – that are difficult for others to hear or even if they're engaging on controversial topics we want to provide a space for that kind of discussion.

At the same time we also want to do our part to keep that online discussion safe and to also protect people's dignity and privacy. And so, our approach to that is to have an understanding of what those norms are, what the trends are, what the speech trends are, what the safety risks are, what the norms are for speech around the world. We do that by engaging with external partners, and this is literally hundreds of experts, NGOs, community groups, and so forth.

And then we also build expertise on the team. So we have a human rights team. We have a civil rights team. And these are people who have spent their careers in these areas and have an understanding of these issues and how we should think about, for instance, human rights law as it pertains to speech and how we should value the important rights of freedom of expression with some of the safety risks.

So the issue that, Glenn, here are not new. They're the ones that define the work that we do every day and have done every day for the past years. It's a space that people have many different strongly-held opinions about the right way to draw these lines, it's also one of the reasons that we have put in place the oversight board to get more independence guidance on how we should think about the balance of (safety) in expression and what our role should be in drawing those lines.

Operator: Thank you. Our next question comes from the line of Elizabeth Culliford with Reuters. Please proceed with your question. Please proceed with your question.

Elizabeth Culliford: Hi, guys. Thanks for doing this. I wanted to ask a bit about instances of the Buffalo shooter videos that have circulated on Facebook and understand sort of what was some of the issues of permanently blocking (inaudible) video, for instance, I know one link took about 10 hours to remove. So I was wondering if there were more you could say about some of the issues whether it's technical, (apetherial) just what went on there?

Guy Rosen: Hi, Elizabeth, thank you for the question. First of all, on the Buffalo shooting, I want to say is a horrific incident and our thoughts go out to the victims, to their families and to the Buffalo community.

So looking at the response, on Saturday right after the event, as soon as we became aware we quickly designated this event as a violating terrorist attacks. This triggers internal process that we have into action to identify and remove accounts and content. Any copies of the video and the manifesto. And we do this as part of a process with industry through GIFCT, the global internet foundation to counter terrorism, as part of crisis response protocols that (inaudible) has developed.

Any copies or links to the video or manifesto or any content that praises or supports or represents the event or the shooter violates our policies and will be removed. And we've had teams working around the clock on this.

One of the challenges we see through events like this is people create new contents, new versions, new external links to try and evade our policies and evade our enforcement. As an incident we are – we're going to continue to learn, to refine our processes, refine our systems to ensure that we can detect, we can take down violating content and links more quickly in the future.

Operator: Thank you. Our next question comes from the line of (Sylvie Kineman) of BBC. Please proceed with your question.

(Sylvie Kineman): Hi and thanks very much. I've got two questions. (Inaudible) whether you can tell us anymore about what you've been seeing regarding content from Russia and Ukraine? I know that you said it's going to come out in the next quarter but surely it's been going on prior to that.

And my second question is about whether you are working on a new way of automating moderation? And I ask because I've had a flurry of conversations with people who say they've been – faced 24-hour bans or comments (inaudible) deleted. And to say (inaudible) really sort of classify in any way as being – as breaching terms (I could murder a) gin and tonic, that sort of thing.

Are you changing the way you're automating and are you experiencing (teething) problems with that? Thank you.

Monika Bickert: Thanks. Maybe Guy, I'll start on the Ukraine stuff …

Guy Rosen: OK, super.

Monika Bickert: … and then I'll – OK. On Ukraine, there's two things to think about here and one is content that we are removing for violating our community standards and then the other is content that we are restricting based on government reports.

And with the content that we're seeing that violates our policies, the content moderation data of course that's in the report that we just released today, those are global numbers and so those are going to affected minimally by the Russian invasion of Ukraine, but I will say that we've seen an increase in violating content in Ukraine and Russia for categories like graphic violence, hate speech, and (violence and incitement), and that's consistent with what we often see when there are major events in the world, we'll see more speech and even more violations.

And then separately on the issue of government requests pertaining to the war in Ukraine, those will be reflected in our next report, but I'll note that we did publish a few case studies real time in our transparency center, which you can access, that pertain to Russian government requests to restrict content on Facebook and Instagram.

And that was specifically so that researchers and journalists could understand the situation as it's happening and scrutinize those requests. Thanks.

Monika Bickert: Thank you.

Guy Rosen: I'll add just on Ukraine before going to the second part of your question. On Ukraine, one thing we can see is an increase in some restores on this report of violating – of, sorry, of graphic and violent content.

That's (us making sure) that we are leaving up the right kind of content that may not actually violate our policies, that is calling awareness to the events unfolding on the ground, certain content is marked as disturbing, so it will have a sort of warning screen on top of it, but actually is allowed to remain up on the site, and as we work throughout the incident and the ongoing war, we are making sure that we are leaving the right kind of content up.

Now to the second part of your question, so for years we've invested in building technology and enforcement (combination) of both human and technology to improve how we detect violating content, but we know that (with the progress) we also make mistakes and it has been equally important along the way, and we're increasingly continuing to focus on how we refine the policies and enforcement, because we hear feedback just like what you mentioned.

So there's a few things that we're mentioning today and you can read in our newsroom post, including AI systems that identifies and prevents potential cases of over-enforcement, learning from content that's been appealed and subsequently restored, you think about things like there's words that may be offensive slurs in one country and common words in another, even in the English language, the British word for a cigarette does not violate our policies but (it's actually) a slur, particularly in the U.S., and so ensuring that context is used more appropriately is important and is part of the systems that we are testing and deploying.

We are also – we are also evaluating the effectiveness and testing ways we can better inform our community about our policies, and give people additional warnings and more information before we trigger penalties on their account. It is still too early to share more, but we're testing different approaches to this because we recognize there is feedback on the current system.

And finally I think it's worth just pointing out also in this context that we are refining how we approach proactive detection in certain spaces and groups or in comments between friends, spaces like that, to ensure that we're taking the right context and nuance, or in the case of groups, ensuring admins have the right tools to better nurture the community that they oversee as part of thinking through just these overall refinements to our policy and our enforcement in operating and managing content at this scale.

Operator: Thank you. Our next question comes from the line of Queenie Wong with CNET. Please proceed with your question.

Queenie Wong: Hi, I had two questions. One kind of building off of you comments, Guy, about the Buffalo shooting. I was wondering if there were any specific new steps Meta was looking at taking to address the fact that some of these videos were reshared and were up on the site more than nine hours.

And then when I was reading the community standards, enforcement report had mentioned there was some sort of bug that – in the media matching technology that impacted, like for example, organized hate and terrorism content.

There was like an uptick in the number of pieces being restored. What exactly was this bug and can we share anymore details about that because that seemed to impact a large amount of content.

Guy Rosen: Hi there. Thanks for the question. So on Buffalo I don’t have any more details to share. As I said, for this incident much as for any incident as we go about this work we’re going to continue to learn and to refine and to make sure that we’re improving our systems so that we’re more ready for the next time.

We’re only a couple days after the incident over the weekend so I don’t have any more to share at this point.

On your second question, essentially the media matching systems is one of the things we are very careful about and I think this may be the issue you’re referring to here is insuring that it doesn’t over enforce. So one thing that may happen in systems like this is that if they – a mistake, a false positive essentially is fed into (medium and patching) it will fan out and take down a large amount of content that doesn’t actually violate.

And so we have to be very diligent about the so called seeds that go into these systems before that fan out occurs. What we had in this case is introduction of some new technology, which introduced some false positives into the system. We subsequently went, restored those posts that were taken down and made sure that they were up because they didn't actually violate our policies.

Operator: Thank you. Your last question comes from the line of Mike Swift with MLex. Please proceed with your question.

Mike Swift: Yes, hi. I have two quick questions. One is Europe recently passed the Digital Services Act and I'm wondering if you can talk at all about how this will affect your reporting going forward and possible enforcement of content moderation policies.

And secondly, can you talk a little bit about why your A.I. filters seem to be improving? Is it just a matter of having more data to train the algorithm on or is the introduction of new technology or some combination of those two things. Thanks.

Monika Bicker: Guy, maybe I should start and then – and then turn it to you. Yes, on digital services we will – we will work with regulators to make sure we understand and are complying with the act. It's a little too early, I think, to speak to the details of exactly how that will look in practice but I'll note that for several years now we have a – an internal pain that's focused on content regulation and there's really two parts to that.

One is making sure that we are providing a – providing any value that we can in the conversation that's going into those regulations. So in other words, we talk to regulators, help them understand some of the challenges that we face. We've called for regulation in some areas. We've tried to put some – put some details into what we think can be helpful in regulation.

And then the second part of that internal team is understanding how regulation, how any one piece of regulation – and there are many around the world – will affect the services that we provide and how we can make sure that we're compliant with those.

And so, these are – these are teams that have been working on understanding this regulation, other regulation, and are now working on making sure that we're able to comply with it.

Guy Rosen: On the second part of your question on A.I., so throughout the years we have absolutely developed new technologies to help support the work we do here. This is not just about having additional data and training, although that is, indeed, and important part of any artificial intelligence and its systems, but we have really been developing and moving forward the state of the art.

If you go back eight, nine years ago it was really just texts and key words matching. In the last years there has been a lot of development in computer vision and then onto multimodal understanding of images and texts and videos together. And we've spoken, including on this call, and we can share some of the previous posts on this as well, some of the technologies we have developed including XLM-R (our former fushot learner) that helps to bootstrap classification in newer areas.

Pieces of this, for example, or XLM methods uses a single shared encoder to train (of our matters) multilingual data, and that helps us train better across different languages, which is a challenge in areas like this particularly for areas like hate speech, which are text and language-based.

The other development in particular in the past couple of years have been consolidating a lot of our A.I. systems, for example, across areas like hate speech or bullying and harassment or violence and incitement, which are quite sort of adjacent to each other if you will from a sort of textural perspective.

We actually have – we've made progress by (trading) sort of unified models for this where the systems sort of cross train across the areas, and that has enabled us to improve the quality, the accuracy, and the effectiveness of the systems that do this work, so we're absolutely continuing to develop more A.I. technology to make sure that work (can be state of the art) and as well as being able to take and respond to things faster.

Operator: Thank you. I will now turn the call back to Carolyn Glanville for some closing remarks.

Carolyn Glanville: Thank you all so much for joining. Just as a reminder, the embargo lifts at 10 a.m. Pacific today. If you have any follow up questions, please feel free to reach out to press @ so that we can get them answered for you. Thank you for joining.

Operator: Thank you. This concludes Meta's Quarterly Integrity Update Call. Thank you for joining. You may now disconnect your lines.

END