# FACEBOOK, INC.
# COMMUNITY STANDARDS ENFORCEMENT REPORT, Q4 2021
# February 11, 2021
# 11:30 a.m. ET

Operator: Hello and welcome to today's Community Standards Enforcement Report Press Call.

There will be prepared remarks and a Q&A to follow. To ask a question after the prepared remarks conclude, please press "star" followed by the number "one."

Now, I'd like to turn the call over to Sabrina Siddiqui, who will kick this off.

Sabrina Siddiqui: Good morning, everyone. Thank you for joining us. You should have received embargoed materials including our data snapshots and a copy of our report ahead of this call. We are on the record and this call is embargoed until 10:00 a.m. Pacific Time.

Today, you will hear opening remarks from Vice President of Integrity, Guy Rosen; Vice President of Content Policy, Monika Bickert; and CTO, Mike Schroepfer. We will then open up the call for questions.

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

Guy Rosen: Thank you, Sabrina. And good morning, thank you everyone for joining us today.

Today, we're publishing our Community Standards Enforcement Report for the fourth quarter of 2020. And we view this report as a quarterly custom, a way to give updates on our enforcement results, policy changes, technology updates. This report includes metrics on how we enforced our policies from October through December across 12 policies on Facebook and 10 on Instagram.

Throughout this coming year, we plan to share additional metrics on Instagram and to add new policy categories on Facebook as well. Our goal here is to lead the industry in transparency about this kind of work. And we'll continue to share more as part of this ongoing effort.

We also know and believe that no company should grade its own homework. So we've committed to undertaking an independent third-party audit of our content moderation systems. This year, we'll begin working with an external auditor to validate the numbers that we publish here.

Last quarter, we shared the prevalence of hate speech on Facebook for the first time. Prevalence, as a reminder, is the percent of times that violating content is seen on our platform and it is the main metric we use to judge how we're doing on enforcement. And it matters because it captures not what we took down but what we missed.

In Q4 of 2020, hate speech prevalence on Facebook dropped from between 0.10 to 0.11 percent to between 0.07 to 0.08 percent, or between 7 and 8 views of hate speech for every 10,000 views of content. The prevalence of violent, graphic content also dropped from 0.07 percent to 0.05 percent. And the prevalence of adult nudity content dropped from between 0.05 and 0.06 percent to between 0.03 and 0.04 percent.

These improvements in prevalence are mainly due to changes we've been making to reduce problematic content in News Feed. The way it works is this, each post a person sees is ranked by processes that take into account combination of integrity signals, among other things, such as how likely a piece of content is to violate our policies, as well as certain signals that we receive from users.

Things like actions they can take, like hiding or reporting a post. And improving how we've been using these signals has helped us tailor newsfeeds to each individuals preference, and it also reduces the number of times we display posts that later may be determined to violate our policies.

We've also been seeing improvements in our AI's performance in a number of areas, with our proactive rate on bullying and harassment content increasing on Facebook from 26 percent in Q3 to 49 percent in Q4; on Instagram, from 55 percent in Q3 to 80 percent in Q4. This means that of the bullying and harassment content that we removed, 49 percent on Facebook and 80 percent on Instagram was found by our systems before the user reported it to us.

We had a drop this quarter in our content action numbers in the child nudity and sexual dislocation of children policy area. There were two reasons for this. First, towards the end of last year, we've made a number of changes to our media matching systems. We later discovered a technical issue in the implementation of those changes that started in mid-November. When that error was discovered, we fixed it and are in the process of going back to retroactively remove all of that content that was missed.

Secondly, Q3 was actually higher because there was a spike at the time in viral content that was shared and which we removed. When content, such as that, is not shared anymore, the enforcement numbers will decrease.

Now finally before I wrap, I would like to share our enforcement numbers for COVID. In Q4 of 2020, we removed over 1 million pieces of content on Facebook and Instagram globally for containing misinformation on COVID-19 that may lead to imminent physical harm.

This is things such as content relating to fake preventative measures or exaggerated cures. This Monday, we announced an update to our COVID-19 related policies, which Monika will speak to shortly.

And with that, I'll hand it over to her. Over to you, Monika.

Monika Bickert: Thanks, Guy, and hi everybody. Thanks for joining. Guy talked about our decrease in prevalence across harmful content and I really want to underscore, these members are important because in showing how much content we miss, they help people hold us accountable.

As I think you know, we remain the only company to publish these numbers, and I'd like to share a little bit about the people, the policies and the (proudest) work that has led to those numbers. If you've listened to one of these calls in the past, then you might know that I lead the team that writes our policies.

This team is made up of more than 200 people based in 11 offices around the world. And we have people from a diverse range of experiences, backgrounds and countries. That team continually works to refine our policies as a world and speech trend evolves, and we work regularly with hundreds of organizations and experts outside the company to make sure that we're doing that in the best way that we can.

It will come as no surprise to you that in the month of October through December of 2020, we were incredibly busy as a content policy team as the global pandemic surged in the U.S. and Europe and as elections took place in the U.S., Myanmar and Brazil.

During those months, we made several key policy updates. We updated our hate speech policy to prohibit any content that denies or distorts the Holocaust. In countries with elections, we made localized policy refinements to protect those elections from misinformation and voter interference. We expanded our dangerous individuals and organizations policy to address militarized social movements and violence inducing conspiracy networks like QAnon.

And we continued our efforts to combat misinformation on vaccines and the COVID-19 virus, ensuring that our policies track expert health advice and medical breakthroughs. For example, in December, as countries like the U.K. and the U.S. moved to approve COVID-19 vaccines, we updated our polices

to ensure that key organizations like the World Health Organization could run ads that promote ways to safely get the vaccine.

And while not reflected in the data in this report this week, we also announced that we are expending our efforts to remove false claims on Facebook and Instagram about COVID-19 or COVID-19 vaccines or vaccines in general during the pandemic.

Specifically following consultations with leading health organizations including the WHO, we've now expanded the list of false claims that we will remove during this pandemic to include additional debunked claims about the coronavirus and about vaccines in general.

And Facebook pages and groups and Instagram accounts that repeatedly share these debunked claims will be removed from our platforms all together. We are also requiring some administrators for groups with admins or members who have violated our COVID-19 policies to approve all posts within their group.

Now a little bit about the Oversight Board, since we post our last community standards enforcement report, the Oversight Board has become operational, it's selected its first cases back in December 2020, and it published its first decisions two weeks ago.

The board's first rulings were a significant moment for content regulation and accountability because they marked the first time that an independent body has reviewed a private social media company's content decision at that company's request. And along with their decisions, which we've already implemented, the board gave us recommendations tied to content policy and content moderation practices.

My team is currently reviewing those recommendations and working out a plan for how we will consider and address them. Now, some of their recommendations will be relatively simple to implement. For example, adding more detail to our community standards on COVID misinformation, which we did earlier this week alongside our vaccine misinformation announcement.

But other recommendations will take considerable time and effort to scope. Such as their recommendation to let people know if humans or artificial intelligence reviewed their content. Overall, we believe that the board included some important suggestions and we will take those to heart. Their recommendations will have a lasting impact on how we structure our polices and we look forward to continuing to receive the board's decisions in the months and years to come.

And now I'd like to talk a little bit about regulation, content regulation. The board is an important step in building accountability and oversight into what we do, but we know that the board isn't going to solve every problem and it certainly doesn't replace the need for responsible regulation.

As we talk about putting in place regulation or reforming Section 230 in the U.S., we should be considering how to hold companies accountable to take action on harmful content. And as I've said before when we've released these reports, we think that the numbers we're providing today can help inform that conversation because they show our performance and our progress in these areas.

We think that good regulation could create a standard like that across the entire industry. Before I turn it over to Schroep let me also say a quick word about the January 6, attack on the U.S. Capitol. As we said immediately following the attack, we were appalled by the violence and put a series of emergency measures in place. Even before the attack, we had removed networks related to some of the extremist groups that were allegedly involved in the violence.

We've always had a strong relationship with law enforcement, we were monitoring the assault in real-time and made appropriate referrals to law enforcement to assist their efforts to bring those responsible to account. This includes helping them identify people who posted photos of themselves from the scene, even after the attack was over.

We're continuing to share more information with law enforcement in response to valid legal requests as they continue their investigation. And with that, I'll hand it over to you, Schroep.

Mike Schroepfer: Thanks, Monika. AI is a key part of how our content enforcement works at scale and the numbers released today show how that enforcement benefits from continuous technological advancements. Throughout 2020 our team has made steady progress in improving the capabilities of the AI systems used to enforce our policies.

That progress came in many forms, and I'd like to highlight a couple of them today, just to give you a sense of how these things are improving. The impact of these improvements is clear, the amount of hate speech spotted proactively by our automated systems before anyone reported it rose to 97 percent, up from 94 percent in previous quarter and 80 percent in July of 2019.

But more importantly, that number was just 24 percent as recently as late 2017. Quarterly gains are impressive but making them time and time again over a number of years is where we see the major impact of continuous technological advancement. One example is the way AI is getting better at

detecting violating content in the comments of a post. This has been a big challenge historically because judging whether a comment violates our policy often depends on the context in which it's replying to.

So, a content – a comment saying "This is great news" can mean entirely different things if it's left beneath a post announcing the birth of a child or the death of a loved one. This kind of contextual understanding might seem obvious to us, to people, but it can be really hard for computers.

Throughout 2020, our team's involved in how AI analyzes comments considering both the comments themselves and their surrounding context. To do this, we needed our AI to develop a deeper understanding of language as well as the ability to combine the analysis of images, text, and other information to make a judgment. This is cutting edge technology and it keeps getting better and having more impact. Improvements to our comment classifiers help boost the amount of bullying and harassment content action on Facebook by over 50 percent over the previous quarter.

Additionally, at the beginning of 2020, just 16 percent of this content was being practically detected by AI. By the end of the year, that number has increased almost 49 percent. We also saw progress in the way our systems operate in multiple languages thanks largely to improvements in Spanish and Arabic, the number of pieces of hate speech content that were taken down increased from 26.9 million up from 22.1 million in the previous quarter.

There wasn't one single breakthrough here, there was a whole package of AI technologies getting better. We have a new architecture called Linformer, which we put into production last October that allows AI models to train on larger and more complex pieces of text. And there's RIO, a system we put into production in November that allows content moderation tools that constantly learn from new content being posted to Facebook each day.

You can read more about these advancements in our Facebook AI Blog. But I'm happy that we didn't just build these technologies, we published the research behind them and released the code because we want the research and engineers – researchers and engineers around the world to be able to work with what we've made and use it to make the internet safer.

There's still much more work to be done here, we're especially focused on getting AI even better at understanding the context of speech across different languages, cultures, and geographies. The same words could be considered harmless or hateful depending on where they're spoken and who's speaking them. And training machines to capture this nuance is a really hard technical problem.

But the message I hope you'll take away from this report is we continue to be committed to transparency and we're making progress and we won't stop. Our research scientists have made fundamental breakthroughs in 2020 that are moving from their laboratories to the core systems faster than ever. This year we'll be rolling out entirely new technologies that will join forces with those that drove so much progress last year.

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 "star" followed by "1" on your telephone keypad.

Your first question will come from Steven Levy from Wired. Thanks you line is open.

Steven Levy: Thanks. You guys can hear me? Hello?

Mike Schroepfer: Yes.

Steven Levy: OK, great. You know thanks a lot for the call. I want to follow-up a little maybe Monika with the response to the Oversight Board. If you were to pull together a thread of the first set of decisions, I think it was that Facebook doesn't take as much time to consider the challenges to – the content decisions. Basically the message they're sending is that you guys just need to spend more time taking a look at it and you know before you make the decisions there.

Did you extract that same result from looking at the decisions as a bunch? And are you thinking of changing the way, getting more time to the decisions to get channels?

Monika Bickert: They definitely raised – they definitely raised questions about our overall processes and the notice that we provide around, and I mentioned this earlier, but the specificity in our policies but then also they raised things like should we be telling people when their content was reviewed by person or by automation. So they did – they did raise some really interesting questions.

And, like I said, we'll be putting out responses in 30 days where we'll address those. As I think you know, the content decisions themselves are binding, we've implemented those. We're now looking more fully at what they said around general policy guidance and general practices.

But I think from my – from my standpoint, this was what we intended the board to do. You know we wanted them to be truly independent so that they could approach these issues from a perspective where they could not only say this piece of content must be reinstated or must stay down. But they could

also point out to us areas where we should be investing in terms of making the overall process better.

Operator: Your next question comes from Octavio Castillo from El Universal. Please go ahead, your line is open.

Octavio Castillo: Good morning to everyone. My question is also for Monika because it is something about content moderation. It is something that is happening here in Mexico. As you may well know, our government is going to discuss in about two weeks a proposal of law that is contemplating that all the providers of social media platforms, Facebook included, are going to be somehow responsible for fake news, hate speech and so on. But in a – in a more legal way. And they are also questioning the space that these social media platforms have for freedom of speech. And they are arguing that these platforms are not allowing freedom of speech. While doing this homework or doing this work of moderating the content, what is your opinion about it?

Monika Bickert: You know, this is – this is a debate that we're seeing in so many places in the world right now. People recognize the need for a refreshing of the laws and regulations that govern online speech. I mean many of these regulations are so, so old. They need to be updated and we very much agree with that. But we are seeing this debate.

How do you impose restrictions on social media companies and make them be more accountable for removing harmful content but also ensure that they don't just start removing anything remotely close to the line and impinge on what should really be considered free expression.

Our thoughts, and we put out a paper on this about a year ago through our Newsroom, if you search for – actually if you search for my name and search our Newsroom, you'll find it but our thought is that there's a couple different ways to go about regulation. We think regulation would be a very good thing. We think that we need to start collectively by having systems in place for transparency about what the social media companies are actually doing and how well it's working.

Basically the kinds of stuff that we're putting out in the report today, we think that should really be a foundation that will inform the conversation on how you can have regulation that achieves the restriction without also creating bad incentives to either go after content that's actually less harmful but easier to regulate.

For instance, we've seen regulations around time to take down of content that is reported by users when we've seen from our own experience that often using our systems is the best way to find content that is the most harmful or is

going viral the quickest. And so sometimes regulation could create those sorts of bad incentives or it can create an incentive to take down too much speech.

So we think the starting point is getting the right systems for transparency in place. And we are working worldwide, including in Mexico, we're working worldwide to make sure that we are involved in the conversation around regulations and can help bring it to a thoughtful place.

Operator: Your next question comes from Shannon Bond from National Public Radio. Please go ahead, your line is open.

Shannon Bond: Thanks for taking the question. Another question for Monika. When you talked about the changes this week to COVID vaccine policies and vaccine misinformation in general, you talked about removing those – that kind of misinformation during the pandemic. When it comes to vaccine misinformation in general, is that a change that will extend – I mean hopefully at some point, we'll be beyond this pandemic, is that a permanent change to Facebook's policies or are you – are you looking at just in the context of the pandemic?

Monika Bickert: Well, we're starting by looking at it through the lens of the pandemic. And let me explain a little bit why. Our approach to misinformation generally has been that we want to give people more information so they can see the fuller picture. That's why we work with fact checking organizations to label content false, to reduce distribution and to provide links to authoritative information.

However, starting around in 2018, we said if we think there is a risk that content that is false could lead to or contribute to an increased risk of imminent physical harm, we'll remove it.

Now we first – when we started using that policy, it was mostly in the context of things like there's been a riot in a location and a rumor that is false could actually lead to further riots or further violence. But we first used this policy in a health context when there was a measles outbreak in Samoa. And we said here, misinformation about vaccines could actually contribute to imminent risk through transmission from people not getting this vaccine while there was this outbreak.

And then we used that same policy at the start of this pandemic for the same reason, that there is an imminent risk here, people are actively transmitting this virus, and it's a – it's a global threat, and we specifically looked to the fact that this is then recognized as a global pandemic by the WHO. So that's how we got here.

Now recently, as we've been talking to World Health Organization and others, we've been looking at the full list of misinformation that could actually

contribute to that risk of imminent harm, and that does include some of these more general vaccine false claims, such as vaccines cause the disease that they are supposed to prevent, or vaccines are part of a government plot, or vaccines are just composed of toxins, or whatever the case may be.

So we do think that it is important to remove those. After the pandemic is over, we will continue to talk to health authorities and make sure that we are striking the right approach going forward consistent with our misinformation policies and principles.

Operator: Your next question comes from Issie Lapowsky from Protocol. Please go ahead, your line is open.

Issie Lapowsky: Hi, thank you so much. So Guy, you said in the beginning and said in your blog post that your goal is to make Facebook as transparent as possible and to lead the industry. So at the risk of asking for a little bit more than what is in this report, I'm wondering if Facebook is considering sharing information on recommendations, say the amount of recommended content, pages, groups that contain a large percentage of the policy violations.

And then second, this is a quick one, I might be missing it, but I don't see where you guys have ever reported stuff on violence and incitement for violence. I see violent and graphic content, but I understand that to be a separate category. So if that is indeed a category you don't report on, can you share why? Thank you.

Guy Rosen: Hey, thanks for the question. This is – yes, this is a multi-year journey, and this report even has expanded over more and more areas just over the course of the now a few years in which we've been publishing it.

So we absolutely want to keep expanding it to ensure that we're sharing, first of all, the policies and the approach and just sharing with the public how we treat these types of issues, but then also expand how we report to – metrics around these different types of entities on the platform.

On recommendations, we published our recommendation policies, we thought it was important to make sure that that is out there and people can see that and scrutinize it. We don't have any numbers yet to share in terms of the extent of that enforcement, and indeed even in the community standards enforcement report, it covered the actions we take on individual posts across a number of policy areas.

One thing that is certainly on our mind and a part of the road map is to expand that to also cover what we call conflict entities, in like pages or groups or accounts that are also taken down under those – under those policies.

So we don't have any immediate timeline for that, but it's absolutely on the list of things that we want to get to and expand how we are transparent about that.

The same goes for policy areas. We started with a few policy areas, mostly on Facebook, we've grown to a larger number of policy areas on Facebook and Instagram. We want to keep expanding those. The violence and incitement policy area is one that's certainly on our road map, and I do hope we'll be able to share more about that soon enough.

Operator: Your next question comes from Julie Jammot from AFP. Please go ahead, your line is open.

Julie Jammot: Hi. My question is regarding your blog post you had yesterday from Instagram talking about the direct messages. I was wondering, how do you police hate speech and other problems. Is it mostly information? Is it because it gets flagged or reported? Thank you very much.

Monika Bickert: Sure. Well, first, let me be clear that the policies do apply across Messenger. And we do use – we do rely on user reports as well as use some other ways to try to detect content that violates our policies.

But I want to talk a little bit more specifically about the – what's been happening in the U.K. with the racism against football players there. This is something that – and we've engaged and listened to this community and this is abuse that is truly horrific. And it's something that we want to make sure is not happening on our services. And so, we're doing a couple of things here.

In addition to putting in place additional measures to remove accounts that are engaging in this sort of hate and harassment, we're also developing new controls that will allow people to reduce the chance that they could ever receive abusive messages in the first place. So that's definitely two prongs of the approach. But then there's a third prong, which is working to help the community use social media to actually push back on this racism in the first place.

I mean, what – the abuse that we're seeing is horrible but we also know that most people who are football fans are not engaged in this type of abuse. So, part of the goal here is helping people leverage social media, helping that majority of people who are really well intentioned to pushback on discrimination when they see it. And we've been working on this for a while.

Last year, we teamed up with Kick It Out in the U.K. to launch and anti-discrimination initiative that's called Take a Stand. And we're continuing to look at other things we can do to both remove the abuse, give people control over their messages but also help people push back on this from happening in the first place.

Operator: Your next question comes from David Ingram with NBC News. Please go ahead, your line is open.

David Ingram: Hey, everybody. I have a question about groups. And I know that came up a little bit earlier. But it looks like most of the report – and actually I could use some clarification on whether the report is just covering problematic content in news feed or if it also includes groups. And more generally, what steps is Facebook taking to improve unfortunate groups given there have been stories recently about the company being slow in taking down content from groups in particular.

Guy Rosen: Hey, this is Guy again. Thanks for those questions. It's definitely important. So first to clarify on what is covered in the report, it reports the content action numbers and the metrics are with respect to posts that are taken down. But those posts may be in groups, they may be just regular posts people post on their profile and so forth because all of those ultimately are part of the experience and part of the – part of what people are seeing on their devices, which is also what we use then to understand the prevalence of something actually showing up when we serve content to people that ends up actually being something that violates our policies and that may have surfaced through a group they're a part of, a page they're following or a friend that may have posted or re-shared something.

Now, groups are definitely an incredibly important area for us and perhaps the most important thing to understand for us as we work through this product is we need to differentiate between good groups with some bad actors in them to groups where the admins are actually a part of the abuse that may be happening and are supporting that behavior.

And one of the mechanisms we use to do that is admins can moderate the content in their groups and, in fact, even can enable a setting where any post in the group is actually surfaced for their approval. And that gives them an opportunity to clean up their group essentially, and to police their group and ensure that they – that group is a place where there isn't that kind of abusive content.

Now, we use that also exactly to differentiate between those cases because if an admin consistently approves posts which later are taken down for violating our policies, that is a signal to us that the admins are – have actually bad intent in sort of how they're running that group. And that will lead to the group being taken down.

So we use a number of signals to help facilitate this. In some groups – as we also mentioned, this week as part of our efforts around protecting against misinformation around COVID-19 vaccines, we require admins to activate

that and, in fact, ensure that in groups around these sensitive topics or during sensitive times, admins must approve all of these posts. That enables us more quickly to clean up those groups. The admins can moderate the content in those groups. And when they don't, those groups may come down.

We also – overall under this approach and under broadly our policies against what kind of groups are and aren't allowed to exist on our platform over the last year, we took down more than 1 million groups for violating our policies.

We've also done a number of other things around controlling which groups are eligible to be recommended – suggested to people. As we've mentioned in the past several months, and as Mark said on the earnings call a couple weeks ago, we're also not recommending any groups that are civic or political in nature to people. We did the same last year for groups that are about health topics.

And the goal here is really – these are areas of high responsibility, we want to make sure people can control their experience. And as we work on this area, we're going to continue to make sure that we provide admins the right set of tools and products to keep their groups clean, as well as take a really hard stance on enforcing where groups may be problematic and may be fostering abuse in them.

Operator: Your next question comes from Elizabeth Dwoskin from The Washington Post. Please go ahead, your line is open.

Elizabeth Dwoskin: Hi, Monika, Guy, Schroep. Thanks for doing this. This one's a question, I think, probably for Monika, but anyone can answer it. So shortly after the Capitol riots, Sheryl did a media interview where she intimated that the riots were largely not planned or organized on Facebook. But we know that it was. We know there was extensive promotion of "Stop the Steal." We know there were bus trips that were organized on Facebook. And of course, no surprise to anyone, it was extensively promoted by Trump as well.

So I kind of have a compound question that is about how you view all of this in retrospect in hindsight. The first question is whether you're considering any more changes to the newsworthiness exception and kind of to the longstanding approach of giving wide latitude to political figures and leaders.

Now, I know you have curved it (inaudible) in the last year and said that there have been (inaudible) said that there's categories such as incitement that the newsworthiness exception does not (inaudible) any further or changing it any further based on what happened?

I'm also wondering why did you, a week after the election you banned one large Stop the Steal group and a hashtag but then you let it back up. What was the rationale for letting it back up at the time?

And then thirdly, sort of just do you take any responsibility for the events that were organized on the platform? Yes.

Monika Bickert: OK, that's – let me – there's a couple things to address there. So, let me – let me walk through those. So, first specifically on the response to the riots at the Capitol, this is an area that we had a team that worked on before the riots, we were taking steps to make sure that we were reducing the chance that violent actors could organize and this was through a number of measures.

We were taking down militarized social movements, that's a policy we've put in place last summer, and as of today, we've removed more than 890 such movements and some of those movements had multiple pages. So, in total, we've removed more than 3,000 pages, more than 19,000 groups, more than 100 events. So, we've removed a lot of content by groups that would try to organize with arms.

We also removed any calls for people to bring weapons to locations. And we also were removing QAnon groups and pages and accounts. So, there is a lot that we were doing in the run-up to make sure that our services would not be abused. And then there's also what we were doing with law enforcement and that's both in the run-up to the violence but also throughout the violence and afterwards.

And I mentioned this a little bit in my remarks earlier, but we were actively looking for content being posted by people involved in the violence and we were making appropriate referrals to law enforcement, starting whenever we became aware of that and certainly continuing through the violence and afterwards. And we will of course respond to valid law enforcement requests. So there's a lot that we did specifically with regard to the Capitol attacks to make sure that we were doing our part.

Now, more broadly on Stop the Steal and QAnon. With Stop the Steal, we first removed – we removed the original Stop the Steal group back in November and then began looking for any time that that term was used by groups or pages or accounts that were also encouraging violence and we would remove those as well.

And then when we looked at the way that that term was being used, in January, we started to see that there was this association with praising the Capitol attacks or praising violence in general. We said, you know, this – we're just not going to allow this term at all. And so, we put that policy in

place in January where we're not removing content that contains the phrase Stop the Steal.

We have not changed that policy, so I can – I can follow-up with you. I'm not sure if there's content that you're seeing that violates and we just haven't removed. If so, that's something that we should be removing, but this does violate our policy.

And then on QAnon, we've continued to remove QAnon content, we've removed more than 3,000 pages, more than 10,000 groups, more than 500 events that have been associated with QAnon.

And then finally on the newsworthiness policy, you know this is a policy that I think people often misunderstand and so I just want to kind of make sure that everybody is clear on what this policy is and what it is not.

The newsworthiness policy is not about protecting a certain group of speakers, it's basically a policy that says that is content posted by anybody violates our policies but we think that the public interest in seeing that content outweighs any risk that that content can contribute to harm, then we may allow the content on the site.

And mostly when we're making newsworthiness calls and leaving content up on the site that otherwise violates our policies, it's usually stuff like somebody sharing a nude image from a situation where let's say it's coverage of a war or something like that where we say gosh, it's really in the public interest for people to see this, even though technically this is a nudity violation. And sometimes we see graphic content the same way. It might be a violation of our policies but it's very important for raising awareness, we would leave it on the site. So that's usually how we use that policy.

We, this summer, clarified that we would not be making newsworthiness exceptions if something is a call to violence or incitement. But that is building on or I guess just sort of articulating the test that we've long had in place, which is this is about a balancing of the public interest and seeing content against the risk that content could actually lead to real world harm. And those principles remain very much our guiding principles in deciding where our policy line should be.

Operator: Your next question comes from Queenie Wong from CNET. Please go ahead, your line is open.

Queenie Wong: Hi, thank you so much for taking the time to answer my question. I was just wondering if the majority of content that Facebook moderates, is it mostly still text and photo based? During the pandemic, it seems like there are other types of media that are rising in popularity, including video and audio. So are

you seeing that type of shift in content moderation as well and does that pose unique challenges compared to moderating video and audio versus text and photos?

Guy Rosen: Hey, Queenie, thanks for the question. We definitely look at a broad range of different formats, and actually even going back to the past one, when we started out with this work. Images were always a big part of it. Graphic content where it's violent or nudity, always an important part of the work that we've done and a part of the system that we built to detect content that's violating.

If anything, our progress on hate speech and bullying harassment, for example, in the past several years has actually required us to make more investments in understanding text and make sure that our systems are able to understand the nuances of language and context in which it's used.

But overall, I don't have a specific metric or stats to share here, but we definitely see, continually, a shift in how different formats come into play in the – and that drive how we think about the systems and technologies that we build and how we moderate content across – ultimately, we have to handle all of the different kinds of things that are posted on our platforms, whether they're textual, they're graphic, they're a video or others.

I don't know perhaps Schroep wants to talk a little bit about the technology behind some of those things.

Mike Schroepfer: Yes. Yes, I mean, obviously we're seeing – we've seen a long-term multi-year shift to richer media, just broadly on how people use these products. Ten, 12 years ago was predominantly text and now it's predominantly imagery, plus text and video is on the rise. So certainly people are using those media more.

I'd say there's two or three areas that are more difficult technically to work on that we've been investing on for years. One is, for example, mixed media. So a meme, when you have an image with text overlaid. And the text overlaid on top of the image can completely change the meaning of what that thing is.

So and these are – I'm sure you've seen these, these are very difficult for computers to understand. So this is actually one of the things we launched recently was the Hateful Memes Challenge is to organize a worldwide challenge to build open source technologies to better detect this. So we are sort of investing in advance of these things on the platform.

The other that has been a very long term investment for us is video understanding. From a technical perspective, video is obviously a lot harder to classify than a single photo. Single photo is a one thing. When you have a

video, you may have a 10-minute long video where five minutes and 22 seconds into the video, there's a five second clip that is the piece of content that's violating that makes the whole thing a problem.

And so that is a place where we've been investing for multiple years on how to apply our technologies to video technology as well. And so we're investing in all of these different media and as people shift how they use the products, we will build the right technology to make sure we keep people safe.

Operator: Your next question comes from (Raphael Balenieri) from (Les Echos). Please go ahead, your line is open.

Raphael Balenieri: Yes, hello. My question is a bit similar to the CNET question just now. There was a Bloomberg story today saying you're working on chat – audio related chat products. And so what does this imply for content moderation to which extent your algorithm today works well or not with audio.

We've seen the (Clubhouse) story this past few weeks. Could you work with the admins, admins of those groups as well? What do you think about this?

Mike Schroepfer: Yes, I'll jump in here. And as I said, I think that we're investing in technology across all of the different sorts of ways in which people share. So the trend really has been for us, building technologies that understand multiple, what we call modalities, at once.

So we don't just understand audio, we understand audio, video; we understand the comments around those things, who shared it. And sort of build a broader picture of what's happening there. And this has actually been one of the big things that has shifted the improvement in the performance of our tools. Because if you go back five years ago, we might have had something that was looking just at the text and something just looking at the images, something just looking at some of the audio.

And today, what we're building is these multi model classifiers, which if there's a video, it will look at the pixels and the audio and the text around it. If it's audio only, it would look at the audio – it would be trying to understand the audio and the comments around it.

So I think there's a lot we're doing here that can apply to these different formats and we obviously look at how the products are changing and invest ahead of those changes to make sure we have the technological tools we need to, again, keep people safe and using whatever means they want to collaborate.

Sabrina Siddiqui: And I think with that we actually have time for just one more question.

Operator: Thank you. Your next question comes from Adi Robertson from Verge. Please go ahead, your line is open.

Adi Robertson: Yes. So going back to the question of Facebook and regulation, what does Facebook think of the SAFE TECH Act that was introduced by Senators Warner and Klobuchar last week?

Monika Bickert: Yes, you know, I don't have specific commentary other than to say that you know we remain committed to having this dialogue with everybody right now in the United States who is working on finding the way forward with regulation. We've obviously seen a number of proposals in this area and we've seen different – we've seen different focuses from different people on the Hill in terms of what they want to pursue. And we want to make sure that we are part of all those conversations.

Sabrina Siddiqui: Thank you all for joining today's press call. If you have any additional questions, just reach out to me or you can reach out to our press email. Thanks. Have a great day.

Operator: This concludes the Community Standards Enforcement Report Press Call. Thank you for joining. You may now disconnect your line.

END