FACEBOOK, INC.
Community Standards Enforcement Report, Q3 2020
November 19, 2020
11:30 a.m. ET

Operator: Hello and welcome to today's Community Standards Enforcement Report. There will be prepared remarks and a Q&A to follow. To ask a question after the prepared remarks conclude, please press star 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, ahead of this call. We are on the record and this call is embargoed until 10:00 a.m.

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 go ahead and kick it over to Guy.

Guy Rosen: Thank you and good morning, everyone. Thanks for joining us today.

Today, we're publishing our latest Community Standards Enforcement Report and providing metrics on how we enforced our policies from July from September 2020. And this report, which is now quarterly, includes metrics across 12 policies on Facebook and 10 policies on Instagram.

I'd like to talk about our first release of hate speech prevalence and how our enforcement numbers are doing as COVID-19 continues to disrupt our workforce. But first, I would like to briefly touch on our efforts around the recent U.S. elections and provide an update on our global efforts to combat misinformation about COVID-19.

Here's an update on the latest election data points. All of these efforts were part of our goals of, first, protecting the integrity of the election by fighting foreign interference, misinformation and voter suppression; and secondly, helping more Americans register and vote.

From March 1 through Election Day, we removed more than 265,000 pieces of content from Facebook and Instagram in the U.S. for violating our voter interference policies. In that same period, we displayed warnings on more than 180 million pieces of content viewed on Facebook by people in the U.S that were debunked by third party fact checkers. We also rejected ad submissions before they could be run about 3.3 million times for targeting the U.S. with ads about social issues, elections and politics without having completed the required authorization process.

Over the past few years, we've also built systems and the teams to tackle foreign interference and domestic influence operations. And we removed over a hundred networks of coordinated inauthentic behavior from our platforms.

As part of the voter registration efforts we undertook, we estimate that we helped about 4.5 million people register to vote this year across Facebook, Instagram and Messenger and helped over a 100,000 people sign up as poll workers. This is based on conversion rates that we calculated from a few states that we partnered with.

140 million people have visited the voting information center on Facebook and Instagram since it launched over the summer. Over 33 million people visited on Election Day.

Now with COVID surging, accurate information is more important than ever. And between March and October of 2020, we removed more than 12 million pieces of content on Facebook and Instagram for containing misinformation that may lead to imminent physical harm such as content relating to fake preventative measures or exaggerated cures. During the same time, we also displayed warnings on about 167 million pieces of content on Facebook based on COVID-19 related debunking articles written by our fact checking partners.

Now let's turn to our Community Standards Enforcement Report, which I should remind everyone covers the period from July until September. Let's talk about our metrics for hate speech and in particular the prevalence metric that we're sharing today for the first time.

Our metrics around enforcement such as how much content we act on and how proactively we find it are indications of the progress we've made on catching harmful content. When we first began reporting our metrics for hate speech, this is for the fourth quarter of 2017; our proactive rate for hate speech was 23.6 percent on Facebook. This means that of the hate speech, we removed 23.6 percent of it was found by our systems before a user reported it to us. The remaining majority at the time was removed after a user reported it.

To date, this rate is about 95 percent on both Facebook and on Instagram and this means of all the content that we removed for violating our hate speech policies, 95 percent of it was first detected by our systems.

Now while these metrics indicate our progress on catching harmful content, the real question is what do we not catch, what do we miss? And that's were prevalence comes in. And it's why we consider it to be the most important measure.

You can think of prevalence like an air quality test to determine the concentration of pollutants. So just as an environmental regulator might periodically sample air quality to calculate what percent of the air we breathe is let's say nitrogen dioxide, we periodically sample content that's viewed on Facebook to calculate what percent violates our policies. And we focus on how much content is seen, not how much of your content is out there that violates our rules. That's important because a small amount of content can go viral and get a lot of distribution in a very short span of time, whereas other content could be on the internet for a very long time and not been seen by anyone.

Now, in the case of hate speech, given the language and the cultural context, in this process, we sent samples of content to reviewers across different languages and regions. Based on this methodology, the prevalence of hate speech between July and September, 2020, was 0.1 percent to 0.11 percent. In other words, out of every 10,000 views on Facebook on average, 10 to 11 of them were content that we consider hate speech under our policies.

We evaluate the effectiveness of our enforcement by trying to keep this prevalence of hate speech on the platform to a minimum as well as reducing mistakes we make when we remove content.

My colleagues will shortly speak to how AI has enabled better enforcement on hate speech as well as about policy changes on what we do and don't define as hate speech.

Now finally, while the COVID-19 pandemic continues to disrupt our content review workforce, we are seeing some enforcement metrics return to pre-pandemic levels. We're also now able to report on metrics that were not included in our Q2 report for the same reasons. We're glad these numbers are getting better, but I do want to remain cautious as our human review capabilities are still less than what they were pre-pandemic and we're continuing to rely heavily on AI. We still prioritize the most sensitive content for human review, which includes areas like suicide and self-injury and child nudity.

Now, I'd like to turn it over to Monika, who has more details on recent policy updates that back these enforcement numbers.

Monika Bickert: Thanks, Guy, and hi everyone. Guy talked about our hate speech prevalence, and these numbers are really important not only because we're the first company to publish those numbers, but also because they will help people track our progress overtime and serve as a metric for accountability. But it's equally important to understand that people, the policies and the product work that backs those numbers.

I lead the team that writes our policies and that team is made up of about 200 people based in 11 offices around the world and they include experts in everything from child safety to cyber security to hate organizations and have a diverse range of backgrounds, experiences and political views. That team, my team, also works regularly with hundreds of organizations and experts outside the company around the world who help us understand diverse perspectives when we craft our policies. And of course our policies don't stand still. We will continue to refine the lines as speech and society evolve.

For example, over the past few months, we've made a couple of key updates. Last month, we updated our hate speech policy to prohibit any content that denies or distorts the Holocaust. And that decision is one that's supported by a well documented rights in anti-Semitism colloquially and the alarming level of ignorance about the Holocaust, especially among young people.

In fact, according to a recent survey of adults in the United States aged 18 to 39, almost a quarter said that they believe the Holocaust was a myth, that it had been exaggerated or they weren't sure.

We've also expanded our policy against dangerous individuals and organizations to also address militarized social movements and violence inducing conspiracy networks, like QAnon. Separately, we worked to protect elections from foreign interference, misinformation and voter suppression and we're continuing our efforts to combat misinformation on vaccines and specifically on COVID-19.

Now separate from the policy updates, today we're also updating our community standards website to include some existing policies that require additional information for us to properly enforce. So these policies aren't new. A number have been announced previously publicly and are well known. But today, we're sharing more details to be even more transparent about our enforcement practices.

So just to be clear what these are, these are policies that require specialized teams to gather more information before we can apply them. So, basically, we don't have our broader team of content reviewers make these decisions at scale.

For example, we require additional information when we're enforcing our policy of removing content that reveals the identity of an informant or a witness in a non-legal proceeding. Our scale review teams just would not have the context to assess whether a person in a witness – is a witness in such a proceeding.

We also require additional information to remove a deceased family member's account because we need to know that the request is really coming from a family member, somebody with the appropriate authority.

In the case of deceased family member accounts, that information can be submitted through our help center. But in other cases, we also look to public news sources or trusted partners with local expertise who were on the ground to gather the information we need to enforce these context dependent policies.

We will continue to update the community standards website monthly as new context dependent policies are developed just as we already update the site with refinements to our at scale policies.

Publishing our policies and our quarterly enforcement reports is part of how we hold ourselves accountable, especially as we look at the prevalence of harmful content and our work to reduce that prevalence overtime. So we really learn from these numbers and we think this report and the types of categories it tracks can be a model for other companies to adopt and for policymakers to consider when updating the rules of the internet.

As Mark said recently, as we talk about putting in place regulation or reforming Section 230 in the United States, we should be considering how to hold companies accountable for acting on harmful content before it gets seen by a lot of people.

The numbers in today's report can help inform that conversation because they show our performance and our progress in these areas. And we think that good content regulation could create a standard like that across the entire industry.

With that, I'll turn it over to Schroep.

Mike Schroepfer: Thanks, Monika. AI has been a powerful tool to support our content enforcement work at scale. The update from this quarter's report reinforces that our investments and advancements are crucial to providing the safest communication platforms possible. But we realize this is a complex, nuanced and rapidly evolving issue.

Hate speech is changing – is challenging for any tech company to detect in all of its forms across hundreds of languages, regions and countries. And over the last few months, we've made progress creating more efficient and effective AI to address these challenges.

In November of 2019, we announced new state of the art research, XLM-R, and a year later, it's in production helping us detect harmful content in multiple languages across Facebook and Instagram. This is a prime example of how Facebook's strength in AI innovation comes from the ability to quickly bring the latest research into large scale production. More importantly, we ensure that this technology us open sourced so the entire industry can benefit from it.

During this time, we also invested in what we believe was a promising way to reduce the prevalence of hate speech across languages through post level self-supervised models and just six months later whole post integrity embeddings give us the ability to train not just on text or images in isolation, but the ability to pull all of them together to better understand the post as a whole.

Over the last few months, we've continued to enhance this technology capability, improving its performance by training it on more violations and data and detecting a wider range of harmful content on Facebook and Instagram.

In late Q3, Reinforced Integrity Optimizer, RIO, went into production on Instagram. RIO was available for Facebook in early Q4 2020, which learns and evolves using online real-world data from our production systems to optimize the AI models that detect complex hate speech, utilizing end to end optimization for significantly better performance and faster iteration.

Also in Q3, Linformer went into production on Instagram, available on Facebook in early Q4. Linformer's a technology for analyzing complicated texts without requiring a lot of computing resources. Along with other AI advances, we're hopeful that Linformer can help us make steady progress in catching hate speech and content that incites violence.

We now use RIO and Linformer in production to analyze billions of pieces of Facebook content and Instagram content in different regions all around the world. We will continue to invest in forward looking AI tech and work with others to take an open and transparent approach to discussing and handling harmful content on our platforms. We share our breakthroughs with the AI research community so we can build on each other's work and accelerate progress for everyone.

This includes Linformer, XLM-R, the Deepfake Detection Challenge, and soon, our Hateful Memes Challenge. We strongly believe that this is the best way for the industry to step – stay in step with the evolving challenge of hate speech and misinfo. And as evidenced by the tech discussed today, it's clear that the best solutions come from open collaboration with experts across the global community.

While AI can't solve this problem on its own, it has given us an incredible advantage to take on this issue at scale, adapting and innovating under unimaginable circumstances and conditions and giving us a fighting chance to deliver on our mission to give people the power to build community and bring the world closer together.

I will now turn the call over to the operator for questions.

Operator: And we will now open the line for questions. To ask a question, press star, followed by the number one. We will pause for a moment to compile the questions.

Your first question comes from the line of Issie Lapowsky from Protocol. Please go ahead.

Issie Lapowsky: Hi, guys. Thank you so much for taking the time to do this. I have two questions, one is for Monika. Monika, you told us on the last call in August that you guys were bringing a small number of moderators back to the office. Can you tell us now what percentage of moderators are back working in an office and do you attribute – it looks like you guys made some major gains in sensitive issue areas like (inaudible) (spam) and suicide related content. In Q3, do you attribute those gains to the decision to bring those moderators back online?

And then if I could, Guy, at the top of the call, you gave us a lot of stats about warning labels, I'm wondering what evidence you guys have that warning labels work and by what measure they work? How that's being studied? Thank you, guys.

Guy Rosen: Hey, thanks for the question and this is Guy, I'll try to take both pieces of that. So the majority of our review work force is still working from home. Stepping back, we're not able to route some of the most sensitive and graphic content to outsource reviewers at home. This is really sensitive content, this is not something you want people reviewing from home with their family around and so we've – in this unprecedented situation, we have – we're taking all the steps to balance the safety of our users, our workforce, the surrounding communities.

So we've done a few things. Our full time employees have taken a larger role in content review, in particular helping to review content in the areas that are more sensitive. As possible in some areas, we're enabling some workers to return to the offices while working within very strict safety standards. We've always had some employees that – in critical jobs who needed to come into some of our global facilities, this is things like data centers and so forth and we have very strict safety standards that we are employing at those locations.

Now in parallel, we've been employing AI and our human review workforce to ensure that we are making progress on the most sensitive content that you call out and we've made progress of a combination both of increased availability in the workforce and increased progress on our AI systems to continue to ensure in a few ways, not just to – both to address the contents that's getting to reviewers, ensuring we're prioritizing and they're spending their time on the most important things but also ensuring that we make best use of each decision by a reviewer.

This is things like technology that removes content that's identical or nearly identical to something a reviewer has made a decision on so that each one of those decisions really can have the maximum impact.

On your second question on warning screens. So on the metric I shared relates to mislabeled – warnings we put on misinformation and we know on – for those labels that 95 percent of people don't actually click to even uncover that warning screen. So we believe those are effective in ensuring that people aren't exposed to that misinformation.

Operator: Your next question comes from the line of Queenie Wong from CNET. Please go ahead.

Queenie Wong: Hi. I was – I had a question about updates to the community standards. Facebook has received some criticism for not pulling down these, Bannon's account after he called for the beheading of Dr. Fauci, and I was wondering if there's going to be more context unless I missed it in the community standards that explains at what point does an account gets disabled versus the content just gets taken down. Is that something Facebook will provide to the public? Like how many strikes your account has to have before it gets taken down?

And building on the question about the labels, have you studied the labels put under the posts about the election versus the ones by fact-checkers? And how exactly are you measuring the effectiveness of those?

Monika Bickert: Hi, maybe I'll start and answer the first question. So, the comments – just to be clear the comments that Steve Bannon made do violate our policies and we removed them promptly from the site. And we just didn't remove the first instance, we removed all versions of it and we blocked additional uploads of it.

Steve Bannon doesn't actually have a Facebook profile, but to be clear, we removed the video from his branded page. And we're clear in the community standards when there are categories of content where we will actually remove somebody from having an account. For instance, we list dangerous organizations and individuals and define them, who we don't allow to have an account. And then we're also clear that there are escalating consequences for people based on what they post.

Now, that varies by the nature and severity of violations. For instance, somebody could post an image of child sexual abuse and we would remove that account right away. And other situations if somebody posted a violation, say it was a bullying violation, they might just get a warning about how not to violate our policies.

I do want to emphasize though that in addition to removing that video and preventing its upload, we also removed several clusters of activity for using inauthentic behavior tactics to artificially boost how many people saw that context. And that affected a lot of pages that were used to – or a series of pages that were used to artificially amplify Steve Bannon's branded page.

So, it's not just about removing contents and stopping it from coming back again, it's also about understanding how inauthentic pages might be used to try to quickly amplify that and responding to that as well.

Guy Rosen: On your – on the second part of your question, the labels. So, the political speech is the most scrutinized speech on our platform. And as of the election, we developed these information labels, which we applied to a number of different posts about the election and from candidates so that people had easy access to the reliable sources about the election. For example, content from the Bipartisan Policy Center and links to our Voting Information Center, which is really sort of a one-stop-shop and it helped over about 4.5 million people register to vote and the goal is really to ensure that we are providing authoritative information.

We always continue to study all of the products and the enforcement levers that we build. And when we share stats on a call such as this one, they're heavily vetted; they go through rigorous analysis and ensure that we are sharing accurate and comprehensive stats. But would also caution anyone from drawing any early conclusion based on things they might have seen.

Operator: Your next question comes from the line of Julie Jammot from AFP. Please go ahead.

Julie Jammot: Hi, thank you for taking my question. Yesterday, 200 content moderators signed a letter calling for better working conditions, but they also criticized the recent development in AI. They were saying that important speech got removed were risky contents, like self-harm data. Is that true? And what is your answer to this problem?

Guy Rosen: Hey, this is Guy. So look, people are an important part of the equation for content enforcement, as we spoke about earlier. People say this is a – these are incredibly important workers who do an incredibly important part of this job. And our investments in AI are helping us detect and remove this content to keep people safe.

The reason, as I mentioned earlier, we are bringing some workers back into offices is exactly in order to ensure that we can have that balance of both people and AI working on these areas.

Now, AI, it's getting better. It's helping us make difficult calls. As I mentioned earlier, it helps us also ensure that any decision made by a reviewer can be applied to as many pieces of content as possible so that those reviewers don't need to review that content again. And we can really make sure that their work is inefficient as – is as efficient as possible.

As I mentioned, we've always had some full-time employees and employees across our workforce who are unable to perform their jobs from home and needed to come into global facilities. This is things like datacenters, security teams, hardware engineers, and so the facilities meet or exceed the guidance on a safe workspace.

We have a variety of health and safety measures put in place for anyone who is returning to office. This is significantly reduced capacity so that there's physical distancing, occupancy limits, mandatory temperature checks, mandatory facemasks, deep cleaning on a daily basis, high-touch surfaces cleaned throughout the day, lots of personal supplies available at the office such as hand sanitizers and wipes and so forth, air filter changes, frequent changes to the air pressure.

All of these are protocols that are in place for Facebook facilities and those in which our – the reviewers work at in order to ensure that we are providing a workspace for them to do this incredibly important work to keep our community safe as well

Operator: Your next question comes from the line of Kurt Wagner from Bloomberg. Please go ahead.

Kurt Wagner: Hey, thank you. Good morning. Two questions, one, I noticed that the number of posts that were restored last quarter seemed to be higher than usual and I'm wondering if that's the result of, kind of, expanding your net, if you will, about what you ultimately decided to take down. And I'm wondering if you could just talk about why those numbers were higher.

And then secondly, Guy, you mentioned those numbers around posts that were labeled for misinformation it sounds like a really, really high number. I think it was almost 350 million posts. Can you give us just a sense of how prevalent misinfo is on Facebook? Based on the other numbers you reported, it seems like that would be the highest category by far. And I'm just wondering if you can give us some perspective around how big a problem that is.

Guy Rosen: Hey, this is Guy. Thanks for the question. So there's a few things. Let me try the first piece.

You asked about our restores, so there's a few aspects here on restores. Generally, some of those numbers are growing in tandem with the growth in content action. So as we take more action, we remove more content, there's more opportunities also for those to be in error. And then, we also provide people the ability to indicate that they think we made a mistake and for us to – for us to restore that.

The other thing to keep in mind is we did – and we talked about this in previous calls as part of our response to COVID and the reduced availability of our reviewer workforce, we did change how the appeal process works. We didn't provide people with direct appeal but we provided people an ability to indicate that they think we made a mistake. And what our teams do is then look at those in the aggregate and understand where we've made mistakes and where a system has an issue and has taken down a series of posts in error and restore those. And so you see some of those indeed – you see that in the report as things that are indicated as restored without appeal, the darker blue on the graphic, if you have that.

We also, for example, on bullying and harassment, we had an error in our system, we made an error there which we detected pretty quickly and restored pieces of content within an hour. But that is also indicated here on this chart.

So I think it's actually a really important area because we can talk about taking down content and all of these big numbers. But we also need to make sure we're constantly balancing that with giving people the ability indicate when we've made a mistake and ensuring that we are held accountable for also correcting those mistakes because this is always a balance of areas that we need to constantly look at and ensure that we're not over or under enforcing in either area.

Your second question around misinformation, we don't have any more stats to share here beyond the enforcement stats that I shared. What happens, just to remind, is our fact checkers write these debunking articles based on – based on certain – based on the right of ratings, based on what they find to be misinformation and they debunk. We then take those and we apply the same AI that we do in many other areas to ensure that whoever finds those warning screens to copies of that information as well. And when that misinformation is – when that is applied then we put a warning screen on top of the content, we notify people who shared the content, we also put an (interstitial) for people who try to share that content so that we're limiting the distribution.

Our goal with misinformation, we don't remove it from the platform, but we want to limit its distribution and ensure people have the accurate context and that it doesn't go viral. And so that changes some of the dynamics here, but otherwise I don't have any other metrics to share at this point.

Operator: Your next question comes from the line of (Diego Lovato) from (Exam) Magazine. Please go ahead.

Thiago Lovado: Hey, guys, thank you for taking this question. Last year, (Schroep), for doing the (Inaudible), you announced improvements of the usage of AI for removing posts before people could see them. And now the system's getting even better and I wanted to know if you intend to reduce the amount of investment you make in human reviewers and moderators? And even if you made any major change (at account that it's) present to human moderators with this technological advance? I know this question has been already answered, but I wanted little more detail on this. Thank you.

Mike Schroepfer: Yes, it's a great question. I mean I will Guy and Monika talk if they're interested about the sort of the long term. I don't see any short term reduction or near term or probably even long term in the human involvement in this.

What happens when our AI systems get better is a few things. One is we catch things faster. So instead of waiting for someone to report it and then us to have to review it, we catch it in advance. And so that's an advantage to the people who use our products. You know the second, as Guy said, we can amplify the work of human reviewers. So in a particularly hard categories like misinformation or some of the most awful forms of content, you do make sure we get it right and have human oversight.

But once they make that decision we catch that in all examples, we can catch it in all variations that people may try to re-upload that. We've had a lot of progress on that even in just the last few months.

And then the third thing is instead of having people do sort of more (roach) review of things that looks fairly obvious to them, we can spend their time and energy working on the harder more nuanced, more subtle things that are harder for even our most advanced AIs.

So to some degree what happens is we get faster, more accurate, more powerful and then we can use our – the amazing staff we have to work on the more nuanced problems that really require human review.

Operator: Your next question comes from the line of Shannon Vaughn from NPR. Please go ahead.

Shannon Vaughn: Thanks for taking the question. In terms of the announcement about the hate speech prevalence metric, can you just talk a little more about why the system that you've decided to share now, sort of what was the impetus between – about making this a public number, and yes, just a little more about the thinking there. Thanks.

Guy Rosen: Hey, this is Guy. Thanks. Thanks for that question. If you look at our report, we have always since we first launched this structure (gateway) number of metrics, content action, proactive rate prevalence, and from the beginning, we have mentioned how prevalence we believe is really the correct north star metric for work like this because it represents not just what we caught but what we missed and what the actual experiences are of people on the platform.

We have been working over the past several years to, if you will, fill in the report across more violation types and across more metrics and as well as across both Facebook and Instagram. And so, as part of this ongoing process, hate speech prevalence, always something we wanted to get to. There was always a sell for that on the grid that we share every time and it just took longer to do that work given the global contextual nature of how we do language review as part of publishing prevalence metric. And so we're glad that we're able to do that and to fill in this metric, which has always been in the works.

Operator: Your next question comes from the line of Jeff Horwitz from WSJ. Please go ahead.

Jeff Horwitz: Hi. So, first question is what's the timeline for getting independent appraisal of prevalences on the platform? Second, would be following up on Issie's and Queenie's question, will you guys commit to making public your research on the impact of the semi-misinformation labels placed on misleading posts regarding election.

And then the final one, sorry for the third, is on Facebook the number of actioned SSI posts that were taken down as a result of user reports have fell by around 90 percent in second quarter and 80 percent in the third, I believe, compared to pre-COVID.

And the company says that every SSI report that a user makes is getting reviewed by a human. What explanation would there be for why the number of reports being valid has fallen so severely?

Guy Rosen: Hey, Jeff. Thanks for those questions. I'll try to remember each of the three.

Jeff Horwitz: Sorry for three.

Guy Rosen: Manage to get – right, no worries. OK. So, your first question was around timeline for an independent appraisal. So, as we mentioned, I think over the last couple calls, we are planning and working towards an audit of these metrics. Like it's very important to have that independent appraisal.

We issued an RFP, a Request for Proposal in August for an auditing firm to work with, so we're currently in process of selecting that auditing firm and we expect to be able to conduct an audit over the course of 2021. This is – these are complex systems and this will be a process that takes time but it's something we're very much committed to.

On your second question, I will reinforce that the goal of labels around elections is to provide more information. We are of course, as everything, we're continuing to study to impact of them and of all other things we do on the platform, but I don't have anything further to share on that point.

On your third question on this SSI, do you mind maybe repeating that question just so I make sure I have it down?

Operator: Give me moment while I locate the participant's line.

Guy Rosen: Sorry about that.

Operator: Jeff Horowitz, your line is open.

Jeff Horowitz: Sorry for the third question myself. The third one was the number of actions SSI posts that were taking down as a result of user reports fell by about 90 percent it looks like, if you back out those numbers from before COVID. I was trying to figure out why the number of valid user reports of SSI content would have fluctuated and fallen so severely over the course of COVID?

Guy Rosen: Got it, OK. So the math I assume you were doing is calculating the action based on the user reports with the inverse of proactive rates, as it were. Now it's a bit tricky to do that necessarily because a piece of content can be both proactively detected and user reported. So, if a – if our improved proactive detection gets to a piece of content that would later have been reported by user otherwise, then it would appear to lower this reaction – reactive report, reactive action calculation. So it needs to be a little bit – it's not necessarily the correct way to look at this.

The other thing to keep in mind if you look at the longer time series is in Q4 of 2019, so prior to pandemic and prior to the year, there was a onetime spike in content action on suicide and self-injury. That's due to rerunning our latest technology on all the old content that was posted previously.

Now, this type is mostly, as you would expect, classifies as proactive as something we're running sort of on our side, but what may happen from the counting perspective and you do – you would see that in your math is an old piece of content that was perhaps reported as much as maybe years ago but wasn't taken down at the time would be classified as a reactive action under the way that our accounting works because what we asked is what came first. Was it first proactively detected or was it first user reported.

The underlying question though that you're asking across this really is how are we doing given these limitations on human review capacity, particularly for this very – can I say, an important area. So just to reiterate a few things, we are prioritizing the most sensitive areas and we – for example, the ones where there's an imminent risk to someone's safety, and this is part of how we overall think about prioritization based on severity and based on morality.

And we accelerated our AI efforts on two paths that are really important, particularly for this area. One is addressing the content that our AI perhaps isn't confident enough to delete automatically and it's awaiting human review.

Now, in some of those cases, even before something gets to a reviewer, we may temporarily limit the visibility of the content, we may automatically send resources to the person who is in distress, all of these even before it gets to a reviewer.

The second thing – the second area of AI is something myself and Schroep also talked about on this call is, ensuring that we're making the best use, we're amplifying each decision by our reviewers, that's more tech to remove content of identical or near identical that we've taken action on.

A couple of other things that I think are worth keeping in mind impact these numbers on suicide and self-injury. One is, viral challenges, which do happen from time to time in this very distressing area and ensuring we're getting ahead of them.

We've had cases where these really dangerous viral challenges started on another platform even and we have teams that get ahead of this, that understand what's happening around the world and we (banned) where we understand the content and ensure that if we can even prevent it from being uploaded so that we prevent these memes, these viral issues from spreading on our platform in the first place.

And final note just around this content or this violation type, how we handle it in the EU, where GDPR does limit our ability to use proactive detection technology. We, after many discussions, our lead data protection regulator has agreed that we can use proactive detection, but it is in a limited way and for moderation only. So, we will make content less visible, we can remove posts automatically, but only if the tech is sufficiently confident that it is very likely to be something that violates our policies.

But to clarify, in the European Union, we can't send content to reviewers, which is often necessary if the tech believes there might be an issue but isn't confident enough to review. We can connect people to safety organizations automatically and we can alert first responders based on proactive detection.

So, all of these are really how we're approaching this area, which I agree, is incredibly important area and we have to continue to use both humans and technology to ensure that we're doing the best job here.

Operator: Your next question comes from the line of Mike Isaac from New York Times. Please go ahead.

Mike Isaac: Hey, everyone. Thanks for taking the time. I want apologize in advance, because I only have one question. But I – sorry, I just (inaudible). I just want to know if you guys can give me any idea on, you did some emergency – well, let’s say extended measures around the election and what types of content you're willing to open to taking down. And I just wanted to know like how long that stuff is going to be in effect or if that has already stopped sort of being the case, if there's sort of less – there was some movement around like more (flag) – more friction to different posts or down-rankings from sort of misinformation related to stuff.

So, I just wanted to know where you all are on that, and if it's essentially still in effect or you decided to step it back?

Guy Rosen: Mike, thanks for the question. So look, ahead of the election, this has been years of preparation and weeks – we've created new product, new partnerships, policies to prepare, including some temporary measures which we have talked to address challenges of potential uncertainty, particularly after Election Day. And there's never been a plan to make these permanent. They will be rolled back just as they were rolled out, which is very carefully.

This is very similar to what we did around other elections around the world, including in the U.S. in the 2018 mid-terms. So we’re continuing to study these, but these are temporary measures that will be rolled back.

Operator: Our last question comes from the line of Hagay Hacohen from Jerusalem Post. Please go ahead.

Hagay Hacohen: Hello. Thank you so much. My question is – could I get a simple explanation about RIO? How does that work?

Mike Schroepfer: Yes. The idea here is the way most machine learning models work today is we build a data set to train and then you build a data set to test. Think of it as a test control group in other context. You run all of these experiments offline, you calculate results, usually precision recall curves, which is sort of how accurate you are in these things and then at some point you're satisfied that the new model you hope is better than the existing model you have in production and then you take this thing and you put it into production.

All of this is basically overseen manually by research engineers, research scientists, and ML experts of the company and then you sort of rinse and repeat this as we go. So as we talk about all these new models improving our capability, this is how it works.

The idea of RIO is instead of doing all this manual stuff offline, you basically build a reinforcement learning system, which is a similar technology that you've seen play games for example, this is sort of the origins of the technology (common), beating Go champions and chess and others is how do I build a system that can optimize towards an objective. In a game playing world, the objective is the score, how do I get the best score of this game.

For RIO, for example, what we do for hate speech is we say prevalence is the thing that we want to reduce, that's our sort of guide, that's our gold star metric. So we set up this system to reduce prevalence of hate speech in real time on the site. So the metric it's using is our sampled prevalence of this thing on the site and then it is sort of back optimizing a whole bunch of aspects of the model, the parameters, and other things like that to sort of real-time tune the system in order to reduce the metric we care about.

So this, RIO is a very first implementation of this. But the idea of moving from sort of a hand crafted, offline system to an online system optimizing in real time, end to end towards the objective we care about is a pretty big deal. And I think that technology is going to be interesting for us over the next few years and is one of many things that we have sort of either early in production or in the research pipeline that will help continue to cause improvements in all of these things.

So as much as we've made progress, A, we're not doing in terms of where we all want to be in terms of taking this content down; but probably even more importantly, B, we're no where close to out of ideas on how to continuously improve these automated systems. It won't be next quarter, but as we roll (over) the next few years, we're going to continue to make improvements through state changes like this as we've gone through multiple state changes over the last five years in our capability of our systems.

Sabrina Siddiqui: All right. Thank you, guys. We appreciate everyone joining the call and appreciate everyone taking the time. I know we weren't able to get to everyone, so just send your questions to Press@Facebook and we'll make sure we get to them. Thank you all.

Operator: This concludes the Facebook Press Call. Thank you for joining. You may now disconnect your line.

END