Days after the March 9 bombing of a maternity and children's hospital in the Ukrainian city of Mariupol, comments claiming the attack never happened began flooding the queues of workers moderating Facebook and Instagram content on behalf of the apps' owner, Meta Platforms.

The bombardment killed at least three people, including a child, Ukraine's President Volodymyr Zelenskiy said publicly. Images of bloodied, heavily pregnant women fleeing through the rubble, their hands cradling their bellies, sparked immediate outrage worldwide.

Among the most-recognized women was Mariana Vishegirskaya, a Ukrainian fashion and beauty influencer. Photos of her navigating down a hospital stairwell in polka-dot pajamas circulated widely after the attack, captured by an Associated Press photographer.

Online expressions of support for the mother-to-be quickly turned to attacks on her Instagram account, according to two contractors directly moderating content from the conflict on Facebook and Instagram. They spoke to Reuters on condition of anonymity, citing non-disclosure agreements that barred them from discussing their work publicly.

The case involving the beauty influencer is just one example of how Meta's content policies and enforcement mechanisms have enabled pro-Russian propaganda during the Ukraine invasion, the moderators told Reuters.

Russian officialdom seized on the images, setting them side-by-side against her glossy Instagram photos in an effort to persuade viewers that the attack had been faked. On state television and social media, and in the chamber of the U.N. Security Council, Moscow alleged – falsely – that Vishegirskaya had donned make-up and multiple outfits in an elaborately staged hoax orchestrated by Ukrainian forces.

Swarms of comments accusing the influencer of duplicity and being an actress appeared underneath old Instagram posts of her posed with tubes of makeup, the moderators said.

At the height of the onslaught, comments containing false allegations about the woman accounted for most of the material in one moderator's content queue, which normally would have contained a mix of posts suspected of violating Meta's myriad policies, the person recalled.

"The posts were vile," and appeared to be orchestrated, the moderator told Reuters. But many were within the company's rules, the person said, because they did not directly mention the attack. "I couldn't do anything about them," the moderator said.

Reuters was unable to contact Vishegirskaya.

Meta declined to comment on its handling of the activity involving Vishegirskaya, but said in a statement to Reuters that multiple teams are addressing the issue.

"We have separate, expert teams and outside partners that review misinformation and inauthentic behavior and we have been applying our policies to counter that activity forcefully throughout the war," the statement said.

Meta policy chief Nick Clegg separately told reporters on Wednesday that the company was considering new steps to address misinformation and hoaxes from Russian government pages, without elaborating.

Russia's Ministry of Digital Development, Communications and Mass Media and the Kremlin did not respond to requests for comment.

Representatives of Ukraine did not respond to a request for comment.

Image source: Independent

Spirit of the policy

Based at a moderation hub of several hundred people reviewing content from Eastern Europe, the two contractors are foot soldiers in Meta's battle to police content from the conflict. They are among tens of thousands of low-paid workers at outsourcing firms around the world that Meta contracts to enforce its rules.

The tech giant has sought to position itself as a responsible steward of online speech during the invasion, which Russia calls a "special operation" to disarm and "denazify" its neighbor.

Just a few days into the war, Meta imposed restrictions on Russian state media and took down a small network of coordinated fake accounts that it said were trying to undermine trust in the Ukrainian government.

It later said it had pulled down another Russia-based network that was falsely reporting people for violations like hate speech or bullying, while beating back attempts by previously disabled networks to return to the platform.

Meanwhile, the company attempted to carve out space for users in the region to express their anger over Russia's invasion and to issue calls to arms in ways Meta normally would not permit.

In Ukraine and 11 other countries across Eastern Europe and the Caucasus, it created a series of temporary "spirit of the policy" exemptions to its rules barring hate speech, violent threats and more; the changes were intended to honor the general principles of those policies rather than their literal wording, according to Meta instructions to moderators seen by Reuters.

For example, it permitted "dehumanizing speech against Russian soldiers" and calls for death to Russian President Vladimir Putin and his ally Belarusian President Alexander Lukashenko, unless those calls were considered credible or contained additional targets, according to the instructions viewed by Reuters.

The changes became a flashpoint for Meta as it navigated pressures both inside the company and from Moscow, which opened a criminal case into the firm after a March 10 Reuters report made the carve-outs public. Russia also banned Facebook and Instagram inside its borders, with a court accusing Meta of "extremist activity."

Meta walked back elements of the exceptions after the Reuters report. It first limited them to Ukraine alone and then canceled one altogether, according to documents reviewed by Reuters, Meta's public statements, and interviews with two Meta staffers, the two moderators in Europe and a third moderator who handles English-language content in another region who had seen the advisories.

The documents offer a rare lens into how Meta interprets its policies, called community standards. The company says its system is neutral and rule-based.

Critics say it is often reactive, driven as much by business considerations and news cycles as by principle. It's a complaint that has dogged Meta in other global conflicts including Myanmar, Syria and Ethiopia. Social media researchers say the approach allows the company to escape accountability for how its policies affect the 3.6 billion users of its services.

The shifting guidance over Ukraine has generated confusion and frustration for moderators, who say they have 90 seconds on average to decide whether a given post violates policy, as first reported by the New York Times. Reuters independently confirmed such frustrations with three moderators.

After Reuters reported the exemptions on March 10, Meta policy chief Nick Clegg said in a statement the next day that Meta would allow such speech only in Ukraine.

Two days later, Clegg told employees the company was reversing altogether the exemption that had allowed users to call for the deaths of Putin and Lukashenko, according to a March 13 internal company post seen by Reuters.

At the end of March, the company extended the remaining Ukraine-only exemptions through April 30, the documents show. Reuters is the first to report this extension, which allows Ukrainians to continue engaging in certain types of violent and dehumanizing speech that normally would be off-limits.

Inside the company, writing on an internal social platform, some Meta employees expressed frustration that Facebook was allowing Ukrainians to make statements that would have been deemed out of bounds for users posting about previous conflicts in the Middle East and other parts of the world, according to copies of the messages viewed by Reuters.

"Seems this policy is saying hate speech and violence is ok if it is targeting the 'right' people," one employee wrote, one of 900 comments on a post about the changes.

Meanwhile, Meta gave moderators no guidance to enhance their ability to disable posts promoting false narratives about Russia's invasion, like denials that civilian deaths have occurred, the people told Reuters.

The company declined to comment on its guidance to moderators.

Denying violent tragedies

In theory, Meta did have a rule that should have enabled moderators to address the mobs of commenters directing baseless vitriol at Vishegirskaya, the pregnant beauty influencer. She survived the Mariupol hospital bombing and delivered her baby, the Associated Press reported.

Meta's harassment policy prohibits users from "posting content about a violent tragedy, or victims of violent tragedies that include claims that a violent tragedy did not occur," according to the Community Standards published on its website. It cited that rule when it removed posts by the Russian Embassy in London that had pushed false claims about the Mariupol bombing following the March 9 attack.

But because the rule is narrowly defined, two of the moderators said, it could be used only sparingly to battle the online hate campaign against the beauty influencer that followed.

Posts that explicitly alleged that the bombing was staged were eligible for removal, but comments such as "you're such a good actress" were considered too vague and had to stay

up, even when the subtext was clear, they said.

Guidance from Meta enabling commenters to consider context and enforce the spirit of that policy could have helped, they added.

Meta declined to comment on whether the rule applied to the comments on Vishegirskaya's account.

At the same time, even explicit posts proved elusive to Meta's enforcement systems.

A week after the bombing, versions of the Russian Embassy posts were still circulating on at least eight official Russian accounts on Facebook, including its embassies in Denmark, Mexico and Japan, according to an Israeli watchdog organization, FakeReporter.

One showed a red "fake" label laid over the Associated Press photos of Mariupol, with text claiming the attack on Vishegirskaya was a hoax, and pointing readers to "more than 500 comments from real users" on her Instagram account condemning her for participating in the alleged ruse.

Meta removed those posts on March 16, hours after Reuters asked the company about them, a spokesperson confirmed. Meta declined to comment on why the posts had evaded its own detection systems.

The following day, on March 17, Meta designated Vishegirskaya an "involuntary public person," which meant moderators could finally start deleting the comments under the company's bullying and harassment policy, they told Reuters.

But the change, they said, came too late. The flow of posts related to the woman had already slowed to a trickle.

Reporting by Katie Paul in Palo Alto and Munsif Vengattil in New Delhi; Additional reporting by Paresh Dave in Oakland, and Dasha Afanasieva and Mark Trevelyan in London; Editing by Kenneth Li and Marla Dickerson

The views and opinions expressed in this article are those of the author and do not necessarily reflect the views of The Kootneeti Team