As Musk is learning, content moderation is a messy job

December 5 2022, by DAVID KLEPPER and MATT O'BRIEN

Elon Musk, Tesla CEO, attends the opening of the Tesla factory Berlin Brandenburg in Gruenheide, Germany, March 22, 2022. Musk said during a presentation Wednesday, Dec. 1, 2022, that his Neuralink company is seeking permission to test its brain implant in people soon. Musk's Neuralink is one of many groups working on linking brains to computers, efforts aimed at helping treat brain disorders, overcoming brain injuries and other applications. Credit: Patrick Pleul/Pool via AP, File

Now that he's back on Twitter, neo-Nazi Andrew Anglin wants somebody to explain the rules.

Anglin, the founder of an infamous neo-Nazi website, was reinstated Thursday, one of many previously banned users to benefit from an amnesty granted by Twitter's new owner Elon Musk. The next day, Musk banished Ye, the rapper formerly known as Kanye West, after he posted a swastika with a Star of David in it.

"That's cool," Anglin tweeted Friday. "I mean, whatever the rules are, people will follow them. We just need to know what the rules are."

Ask Musk. Since the world's richest man paid $44 billion for Twitter, the platform has struggled to define its rules for misinformation and hate speech, issued conflicting and contradictory announcements, and failed to full address what researchers say is a troubling rise in hate speech.

As the " chief twit " may be learning, running a global platform with nearly 240 million active daily users requires more than good algorithms and often demands imperfect solutions to messy situations—tough choices that must ultimately be made by a human and are sure to displease someone.

A self-described free speech absolutist, Musk has said he wants to make Twitter a global digital town square. But he also said he wouldn't make major decisions about content or about restoring banned accounts before setting up a " content moderation council " with diverse viewpoints.

He soon changed his mind after polling users on Twitter, and offered reinstatement to a long list of formerly banned users including ex-President Donald Trump, Ye, the satire site The Babylon Bee, the comedian Kathy Griffin and Anglin, the neo-Nazi.

And while Musk's own tweets suggested he would allow all legal content on the platform, Ye's banishment shows that's not entirely the case. The swastika image posted by the rapper falls in the "lawful but awful" category that often bedevils content moderators, according to Eric Goldman, a technology law expert and professor at Santa Clara University law school.

While Europe has imposed rules requiring social media platforms to create policies on misinformation and hate speech, Goldman noted that in the U.S. at least, loose regulations allow Musk to run Twitter as he sees fit, despite his inconsistent approach.

"What Musk is doing with Twitter is completely permissible under U.S. law," Goldman said.

Pressure from the EU may force Musk to lay out his policies to ensure he is complying with the new law, which takes effect next year. Last month, a senior EU official warned Musk that Twitter would have to improve its efforts to combat hate speech and misinformation; failure to comply could lead to huge fines.

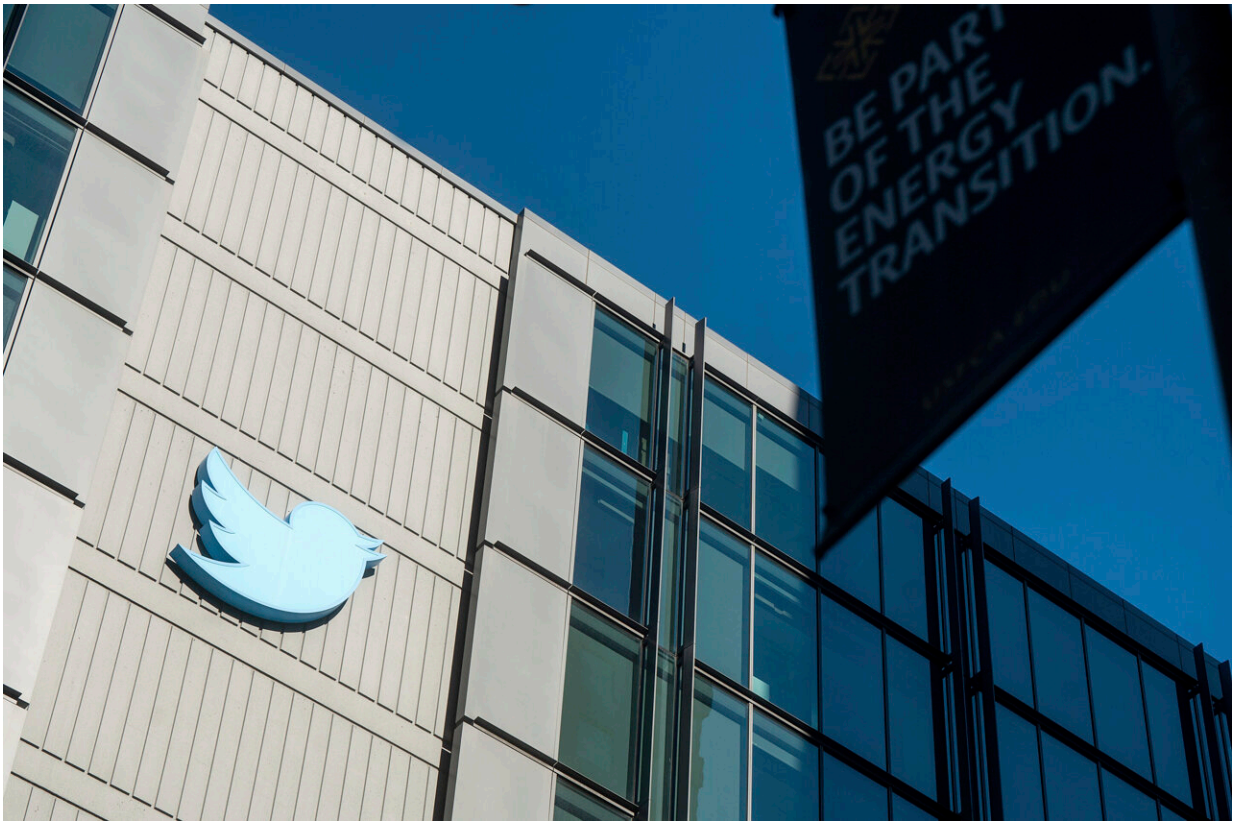
In another confusing move, Twitter announced in late November that it would end its policy prohibiting COVID-19 misinformation. Days later, it posted an update claiming that "None of our policies have changed."

On Friday, Musk revealed what he said was the inside story of Twitter's decision in 2020 to limit the spread of a New York Post story about Hunter Biden's laptop.

Twitter initially blocked links to the story on its platform, citing concerns that it contained material obtained through computer hacking. That decision was reversed after it was criticized by then-Twitter CEO Jack Dorsey. Facebook also took actions to limit the story's spread.

The information revealed by Musk included Twitter's decision to delete a handful of tweets after receiving a request from Joe Biden's campaign. The tweets included nude photos of Hunter Biden that had been shared without his consent—a violation of Twitter's rules against revenge porn.

Instead of revealing nefarious conduct or collusion with Democrats, Musk's revelation highlighted the kind of difficult content moderation decisions that he will now face.



A Twitter logo hangs outside the company's San Francisco offices on Nov. 1, 2022. A top European Union official warned Elon Musk on Wednesday Nov. 30, 2022 that Twitter needs to beef up measures to protect users from hate speech, misinformation and other harmful content to avoid violating new rules that threaten tech giants with big fines or even a ban in the 27-nation bloc. Credit: AP Photo/Noah Berger, File

"Impossible, messy and squishy decisions" are unavoidable, according to Yoel Roth, Twitter's former head of trust and safety who resigned a few weeks into Musk's ownership.

While far from perfect, the old Twitter strove to be transparent with users and steady in enforcing its rules, Roth said. That changed under Musk, he told a Knight Foundation forum this week.

"When push came to shove, when you buy a $44 billion thing, you get to have the final say in how that $44 billion thing is governed," Roth said.

While much of the attention has been on Twitter's moves in the U.S., the cutbacks of content-moderation workers is affecting other parts of the world too, according to activists with the #StopToxicTwitter campaign.

"We're not talking about people not having resilience to hear things that hurt feelings," said Thenmozhi Soundararajan, executive director of Equality Labs, which works to combat caste-based discrimination in South Asia. "We are talking about the prevention of dangerous genocidal hate speech that can lead to mass atrocities."

Soundararajan's organization sits on Twitter's Trust and Safety Council, which hasn't met since Musk took over. She said "millions of Indians are terrified about who is going to get reinstated," and the company has stopped responding to the group's concerns.

"So what happens if there's another call for violence? Like, do I have to tag Elon Musk and hope that he's going to address the pogrom?" Soundararajan said.

Instances of hate speech and racial epithets soared on Twitter after

Musk's purchase as some users sought to test the new owner's limits. The number of tweets containing hateful terms continues to rise, according to a report published Friday by the Center for Countering Digital Hate, a group that tracks online hate and extremism.

Musk has said Twitter has reduced the spread of tweets containing hate speech, making them harder to find unless a user searches for them. But that failed to satisfy the center's CEO, Imran Ahmed, who called the rise in hate speech a "clear failure to meet his own self-proclaimed standards."

Immediately after Musk's takeover and the firing of much of Twitter's staff, researchers who previously had flagged harmful hate speech or misinformation to the platform reported that their pleas were going unanswered.

Jesse Littlewood, vice president for campaigns at Common Cause, said his group reached out to Twitter last week about a tweet from U.S. Rep. Marjorie Taylor Greene that alleged election fraud in Arizona. Musk had reinstated Greene's personal account after she was kicked off Twitter for spreading COVID-19 misinformation.

This time, Twitter was quick to respond, telling Common Cause that the tweet didn't violate any rules and would stay up—even though Twitter requires the labeling or removal of content that spreads false or misleading claims about election results.

Twitter gave Littlewood no explanation for why it wasn't following its own rules.

"I find that pretty confounding," Littlewood said.

Twitter did not respond to messages seeking comment for this story.

Musk has defended the platform's sometimes herky-jerky moves since he took over, and said mistakes will happen as it evolves. "We will do lots of dumb things," he tweeted.

To Musk's many online fans, the disarray is a feature, not a bug, of the site under its new ownership, and a reflection of the free speech mecca they hope Twitter will be.

"I love Elon Twitter so far," tweeted a user who goes by the name Some Dude. "The chaos is glorious!"