Twitter's highest-profile users get VIP treatment when trolls strike

December 8 2021, by Kurt Wagner

Credit: Unsplash/CC0 Public Domain

Twitter Inc.'s highest-profile users—those with lots of followers or

particular prominence—often receive a heightened level of protection from the social network's content moderators under a secretive program that seeks to limit their exposure to trolls and bullies.

Code-named Project Guardian, the internal program includes a list of thousands of accounts most likely to be attacked or harassed on the platform, including politicians, journalists, musicians and professional athletes. When someone flags abusive posts or messages related to those users, the reports are prioritized by Twitter's content moderation systems, meaning the company reviews them faster than other reports in the queue.

Twitter says its rules are the same for all users, but Project Guardian ensures that potential issues related to prominent accounts—those that could erupt into viral nightmares for the users and for the company—are dealt with ahead of complaints from people who aren't part of the program.

This VIP group, which most members don't even know they're a part of, is intended to remove abusive content that could have the most reach and is most liable to spread on the social-media site. It also helps protect the Twitter experience of those prominent users, making them more likely to keep tweeting—and perhaps less apt to complain about abuse or harassment issues publicly.

"Project Guardian is just the internal name for one of many automated tools we deploy to identify potentially abusive content," Katrina Lane, vice president for Twitter's service organization, which runs the program, said in a statement. "The techniques it uses are the same ones that protect all people on the service."

The list of users protected by Project Guardian changes regularly, according to Yoel Roth, Twitter's head of site integrity, and doesn't only

include famous users. The program is also used to increase protection for people who unintentionally find the limelight because of a controversial tweet, or because they've suddenly been targeted by a Twitter mob.

That means some Twitter users are added to the list temporarily while they have the world's attention; others are on the list at almost all times. "The reason this concept existed is because of the 'person of the day' phenomenon," Roth says. "And on that basis, there are some people who are the 'person of the day' most days, and so Project Guardian would be one way to protect them."

The program's existence raises an obvious question: If Twitter can more quickly and efficiently protect some of its most visible users—or those who have suddenly become famous—why couldn't it do the same for all accounts that find themselves on the receiving end of bullying or abuse?

The short answer is scale. With more than 200 million daily users, Twitter has too many abuse reports to handle all of them simultaneously. That means that reports are prioritized using several different data points, including how many followers a user has, how many impressions a tweet is getting, or how likely it is that the tweet in question is abusive. An account's inclusion in Project Guardian is just one of those signals, though people familiar with the program believe it's a powerful one.

Roth said the distinction can't apply to everybody, or it would mean that there's no point in having a list.

"If the list becomes too big, it stops being valuable as a signal," he added. "We really want to focus on the people who are getting an exceptional or unprecedented amount of prominence in a specific moment … this is really focused on a small slice of accounts."

Project Guardian has been used to protect users from a range of

different professions. YouTube star and makeup artist James Charles was added to the program earlier this year after being harassed online. Egyptian internet activist Wael Ghonim has also been part of Project Guardian, as has former U.S. Food and Drug Administration Commissioner Scott Gottlieb, who tweets often about COVID-19 vaccines. The program has also included journalists—even news interns—who write about topics that can result in harassment, like 8chan or the Jan. 6 riot at the U.S. Capitol.

Twitter has used Project Guardian to protect its own employees, including Roth. After the company first fact-checked then-President Donald Trump's tweets in May 2020, Roth was singled out by Trump and his supporters as the employee behind the decision, leading to attacks and death threats. Roth, who wasn't actually the employee who made the call, says he was temporarily added to the Project Guardian list at the time. "All of a sudden I became a lot more famous than I was the day before," Roth explained. He said he was removed from the program after the harassment started to slow down.

Accounts are added to the list in several ways, including by recommendation from Twitter employees who witness a user getting attacked and request added protection. In some cases, a famous Twitter user's manager or agent will approach the company and ask for extra protection for their client. Social media managers at news organizations have also requested extra protection for their colleagues who write high-profile or controversial stories. Users who are in the program don't necessarily know they are receiving any extra attention.

"We look at it as, who are the people who we know have been the targets of abuse or who are predicted to be likely targets of abuse?" Roth said.

Twitter said it is getting better at detecting abuse and harassment automatically, meaning it doesn't need to wait for a user to report a

problem before it can send it to a human moderator. The company says its technology now flags 65% of the abusive content it removes or asks people to delete before a user ever reports it.

Lane said Twitter uses both technology and human review "to proactively monitor Tweets and Trends, especially when someone is put in the spotlight unexpectedly or there is a significant uptick in abuse or harassment."

It's not clear whether there was any one event or incident that sparked Project Guardian, though it has existed for at least a couple of years, people familiar with the program said.The list doesn't just protect prominent users; it also helps protect Twitter's reputation.

In years past, Twitter's image has suffered when high-profile users publicly criticize the service—or abandon it entirely—because of a failure to combat abuse and harassment. It's been particularly common with famous women. Model Chrissy Teigen, singer Lizzo, actor Leslie Jones and *New York Times* journalist Maggie Haberman have all publicly stepped back from the service after being swamped with negative tweets and messages. (They've all since returned.)

More recently, celebrities calling out Twitter for constant harassment seems to be happening less often, though, and some people familiar with the company believe Project Guardian is one reason.

Twitter's program is another instance of the different treatment that social media apps provide to certain pre-eminent users and accounts. A Wall Street Journal investigative report from September found that Meta Platforms Inc., which owns Facebook and Instagram, was giving some prominent users special exemptions from some of its rules, leaving up content from these people that would have been flagged or removed from others.Twitter officials are adamant that Project Guardian is

different, and that all users on its platform are held to the same rules. Reports related to users who are part of Project Guardian are judged the same way as all other content reports—the process usually just happens faster.

While Twitter's rules may apply to everyone, punishments for breaking those rules aren't always equal. World leaders, for example, have more leeway when breaking Twitter's rules than most of its users. Twitter and Meta have also spent years cultivating relationships with high-profile users, creating teams to help those people use their products and to provide hands-on assistance when needed. In 2016, Twitter stopped showing ads to a small group of prominent users with the goal of improving their experience.