

FOR FREE PEOPLE

EXTRA!

Listen to 'The Witch Trials of J.K. Rowling'

HOME | WITCH TRIALS | HONESTLY | SUPPORT US | CAREERS | COMMUNITY | ARCHIVE | ABOUT



Twitter Headquarters is seen in San Francisco on November 18, 2022. (Tayfun Coskun via Getty Images)

Twitter's Secret Blacklists

Teams of employees were tasked with suppressing the visibility of accounts or subjects deemed undesirable or dangerous—all in secret, without informing users.

By The Free Press
December 15, 2022

Subscribe



This story is by [Bari Weiss](#), [Abigail Shrier](#), [Michael Shellenberger](#) and [Nellie Bowles](#).

Twitter—the social-media platform founded in 2006 with more than 330 million users that has fueled political uprisings, launched presidential campaigns, and bequeathed to us the 280-character rant—has claimed to be a global “public square,” [its mission](#) “to give everyone the power to create and share ideas and information instantly without barriers.”

But barriers—many of them—were erected.

After the 2016 election of Donald Trump, Twitter started to play a more active role in policing, or “moderating,” in tech parlance, that public square. It operated a secret blacklist, with teams of employees tasked with suppressing the visibility of accounts or subjects deemed undesirable or dangerous.

All this was happening at the same time that major institutions across America—in the media, education, tech and elsewhere—were becoming less tolerant of views that, a few years ago, were considered well within the 40-yard lines of American politics but are now deemed far outside the parameters of acceptable discourse.

The people in charge of these institutions enforced the new parameters by expanding the definitions of words like “violence,” “harm” and “safety.” Things once considered part of everyday life in America—like disagreeing about whether a global pandemic started in a market or a lab in Wuhan—were increasingly off limits.

Twitter’s waning appetite for ideas or points of view outside the mainstream, in other words, was part of a broader trend sweeping the country.

Whether the platform was a catalyst for that trend or a mirror of it is a question better left to historians.

What is perfectly clear today is that the platform is an indispensable tool for journalists and politicians and that it has deeply affected which stories get covered and how. It has the power to determine the heroes and villains of contemporary news cycles, and to decide which areas of inquiry are legitimate and which are strictly off-limits—even wrong.

So the fact that Twitter shielded users from perspectives it deemed extremist—and that it did so while pretending it wasn't—is a fact relevant to all Americans, whether they have ever logged onto Twitter or not.

Tweets or accounts or hashtags that offended the powers that be were not publicly shamed, but quietly throttled, meaning users frequently did not know they were being deprived of arguments or data that did not support the prevailing wisdom or the politically favored narrative.



For years, Twitter denied that it did this.

“We do not shadow ban.” So declared Vijaya Gadde, then Twitter’s Head of Legal Policy and Trust, and Kayvon Beykpour, at the time the company’s Head of Product, in July 2018 on the Twitter blog. “And we certainly don’t shadow ban based on political viewpoints or ideology.”

People are asking us if we shadow ban. We do not. But let's start with, "what is shadow banning?"

The best definition we found is this: deliberately making someone's content undiscoverable to everyone except the person who posted it, unbeknownst to the original poster.

We do not shadow ban. You are always able to see the tweets from accounts you follow (although you may have to do more work to find them, like go directly to their profile). And we certainly don't shadow ban based on political viewpoints or ideology.

Their statement—"Setting the record straight on shadow banning"—was Twitter's attempt to clear up what they had dismissed as a conspiracy theory: The idea that the site discriminates against users and viewpoints it dislikes by limiting the reach and visibility of their posts.

Not six months before, Olinda Hassan, then Twitter's policy manager of Trust and Safety, was caught on film by Project Veritas saying: "We're trying to get the shitty people to not show up. It's a product thing we're working on."

Which is perhaps why Gadde and Beykpour felt the need to quash that particular conspiracy theory and let users know: You're imagining things. Stop being so paranoid.

"Even paranoids have enemies," Golda Meir is reported to have said to Henry Kissinger. When it came to Twitter, the public wasn't being paranoid, and it definitely wasn't imagining things. Not by a long shot.

So how could executives claim they weren't shadow banning?

There are two meanings of shadow banning. One is techie argot, and it refers to something specific: a practice in which a social media platform makes a particular user's posts invisible to everyone else on the site but the

user, without informing the user of the action. In this very technical sense, Twitter apologists claim, the company did not shadow ban.

But, of course, this is not what the public meant when it accused social-media companies of suppressing their views. What “shadow banning” has long meant to lay critics of Twitter was not that no one on the platform could see their posts, but that far fewer people could. They meant something fishy was happening, screens were being thrown up, the air let out of their tires, so that posts by conservative or non-woke thinkers never seemed to reach as many eyeballs as those that espoused left-approved ideas. They weren’t techies themselves and didn’t know how to describe the mechanism of their suppression, but they sensed its presence.

Not surprisingly, when we searched the Twitter Files for evidence of “shadow banning”—or “shadowbanning”—nothing came up. That’s mostly a matter of word choice. Twitter executives prefer “visibility filtering,” or “VF,” which is all over the files. (Multiple high-level sources at Twitter confirmed the meaning of these terms.)

“Think about visibility filtering as being a way for us to suppress what people see to different levels. It’s a very powerful tool,” one senior Twitter employee told us.

Actually, it’s a set of tools that include locking users out of searches and preventing some users’ tweets from trending—which is how countless other users discover what’s popular or being talked about on Twitter.

“We control visibility quite a bit. And we control the amplification of your content quite a bit. And normal people do not know how much we do,” one Twitter employee told us. Two additional Twitter employees confirmed this.

“They say they didn’t put their thumb on the scale,” Musk, who became CEO of Twitter in October, told The Free Press. “But they were pressing the thumb hard in favor of the left. If left, you could get away with death

threats, and nothing would happen. If right, you could get suspended for retweeting a picture of a Trump rally.”

Nor was the anger directed at Twitter confined to the right.

Congressman Ro Khanna, a California Democrat whose district encompasses much of Silicon Valley, told The Free Press: “The problem that’s happening here is that people are conflating hate speech with viewpoint discrimination.”

He added: “The essence of this story is that Twitter is telling some people that, based on the viewpoints that they have, that they aren’t allowed to ask a question or share their point of view in the same way as everyone else in the room. And that’s just anti-democratic.”

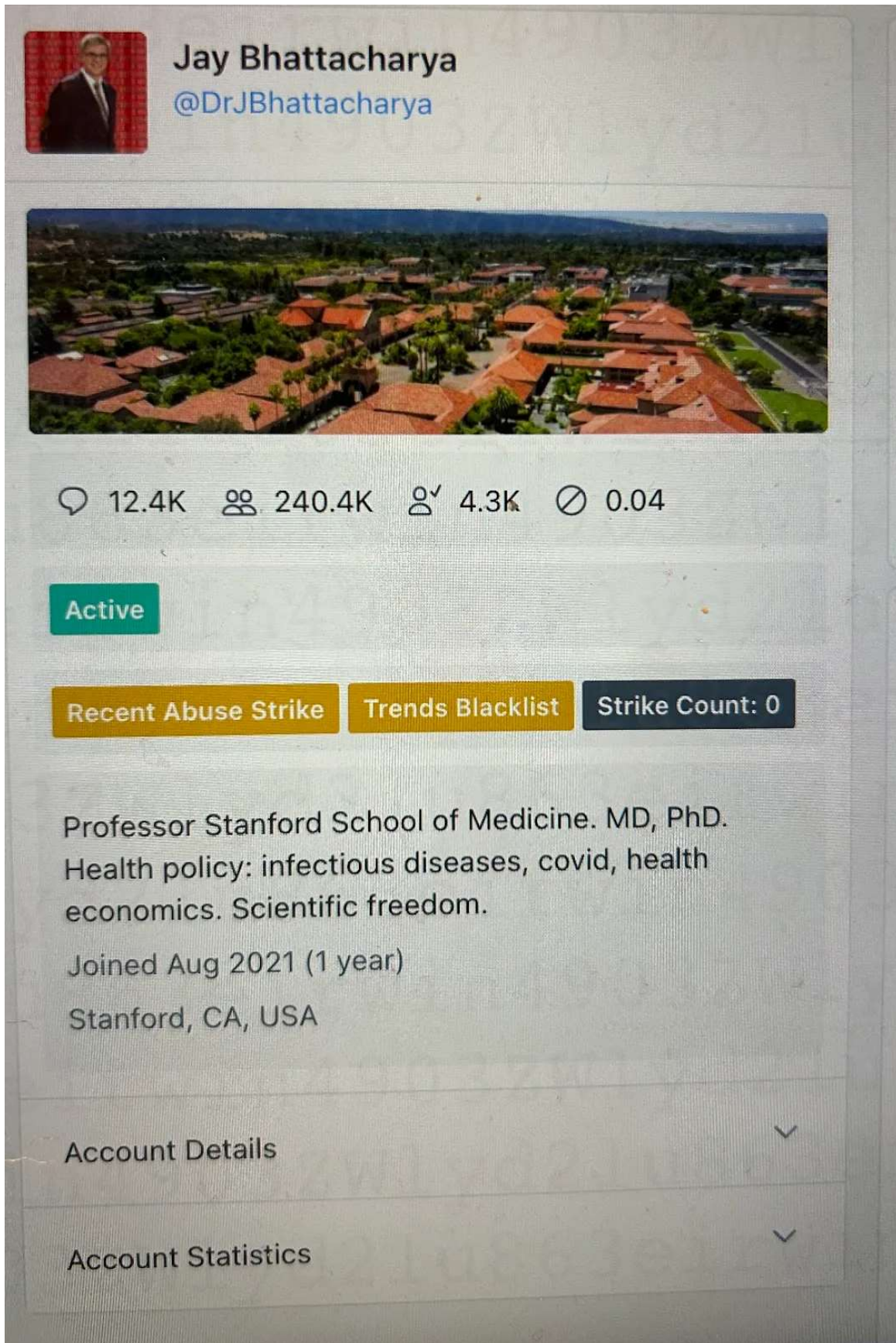


So who—and what—did Twitter mask from view?

One user was Stanford’s Dr. Jay Bhattacharya, an M.D., an economist, and a professor of health policy at Stanford.

In 2020, he and Dr. Martin Kulldorff, then a professor of medicine at Harvard, and Dr. Sunetra Gupta, a professor of epidemiology at Oxford, [warned](#) in an open letter of the dangerous impact of lockdowns, especially on children, the working class, and the poor. They argued for “focused protection” for the most medically vulnerable, and a return to normal life for the rest of society.

This made the three scientists [targets](#) of Washington’s public health bureaucracy. Twitter put Bhattacharya’s account on the Trends Blacklist, which meant that, no matter how many likes or views one of his tweets racked up, it could never “trend”; its visibility to users on the platform would be sharply curtailed.



The image is a screenshot of a Twitter profile page for Jay Bhattacharya. At the top left is a profile picture of a man in a suit. To the right of the picture, the name "Jay Bhattacharya" is displayed in bold black text, with the handle "@DrJBhattacharya" below it in blue. Below the name and handle is a large rectangular profile picture showing an aerial view of a campus with many buildings with red-tiled roofs. Underneath the profile picture, there are four icons representing different metrics: a speech bubble for "12.4K", a group of people for "240.4K", a person with a checkmark for "4.3K", and a crossed-out circle for "0.04". Below these metrics is a green button with the word "Active" in white. Underneath the "Active" button are three yellow buttons: "Recent Abuse Strike", "Trends Blacklist", and "Strike Count: 0". Below the buttons is a bio section with the text: "Professor Stanford School of Medicine. MD, PhD. Health policy: infectious diseases, covid, health economics. Scientific freedom." followed by "Joined Aug 2021 (1 year)" and "Stanford, CA, USA". At the bottom of the profile card, there are two expandable sections: "Account Details" and "Account Statistics", each with a downward-pointing chevron icon to its right.

Two prominent right-wing talk show hosts also came to the attention of Twitter's censors. One was talk-radio host and Turning Point USA executive

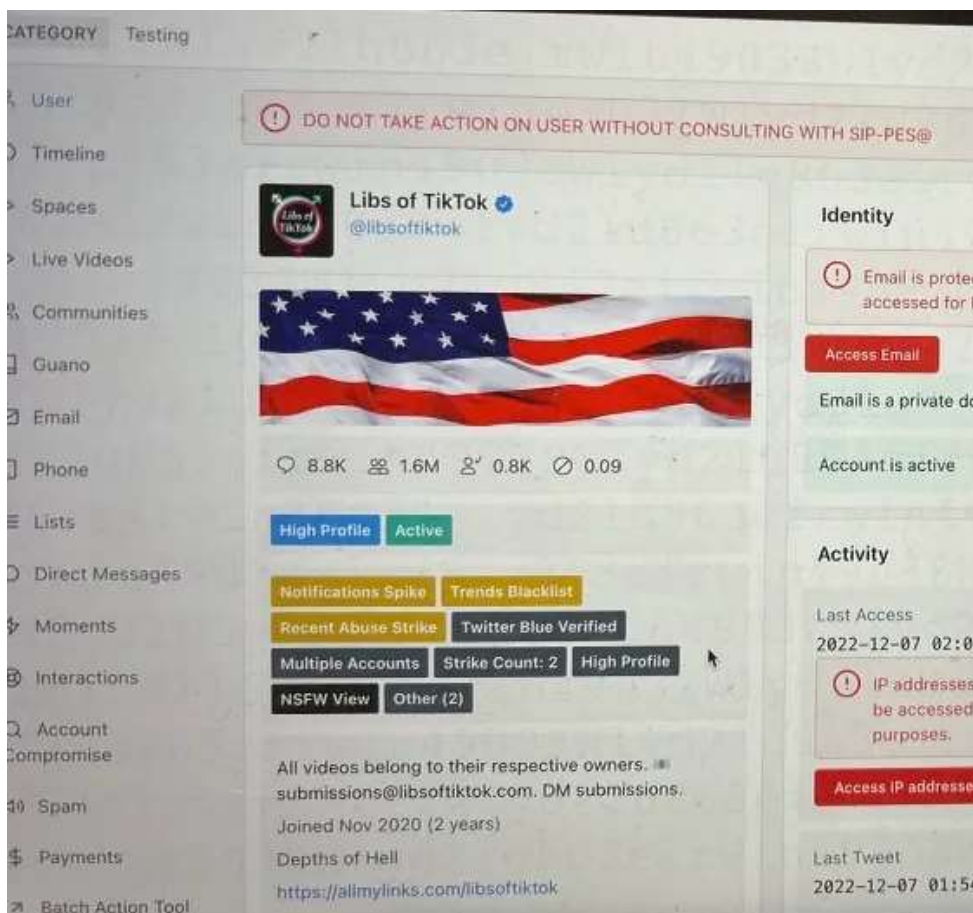
director Charlie Kirk, whose account was set to Do Not Amplify. Another was right-wing talk-show host Dan Bongino, whose account, at one point, was slapped with a Search Blacklist.

The group at Twitter tasked with overseeing all this was the Strategic Response Team-Global Escalation Team, or SRT-GET. It handled up to 200 cases a day.

But there was another, more secretive team hovering in the background, senior to SRT-GET. That was the Site Integrity Policy-Policy Escalation Support team, or SIP-PES. It comprised the most powerful people at the company—including then-CEO Jack Dorsey and, later, his successor, Parag Agrawal; Vijaya Gadde; and Yoel Roth, Twitter's Global Head of Trust & Safety.

This was where the biggest, most politically sensitive decisions got made. “Think highly sensitive, high follower account, controversial,” one senior Twitter employee told us.

One account that rose to this level of scrutiny was @libsoftiktok, which was placed on the Trends Blacklist and designated as Do Not Take Action on User Without Consulting With SIP-PES.



The account was begun in November 2020 by Chaya Raichik, a former Brooklyn real-estate agent. It now boasts over 1.6 million followers.

Raichik often reposts videos other people have made public on social media. For example, this year, Raichik posted videos of physicians at Boston Children’s Hospital discussing the gender-transition procedures they offer to kids and teens. One physician described performing [hysterectomies](#) on female patients transitioning to male. [Another said](#), “We have parents tell us that their kids, they knew from the minute they were born practically” that their child was transgender.

Twitter suspended the Libs of TikTok account repeatedly for up to a week at a time. The reason Twitter provided Raichik for the suspensions was that she had violated Twitter’s Hateful Conduct Policy.

But in an internal SIP-PES memo from October 2022, the committee acknowledged that this wasn’t actually true. “LTT has not directly engaged

in behavior violative of the Hateful Conduct policy,” it said.

Libs of TikTok was not guilty of any “explicitly violative tweets,” the memo stated, but “the user has continued targeting individuals/allies/supporters of the LGBTQIA+ community for alleged misconduct.” The memo noted that, as a result of her “continued pattern of indirectly violating” Twitter’s Hateful Conduct Policy, Raichik’s posts had tended “to incite harassment against individuals and institutions that support LGBTQ communities.”

Site Policy Recommendation

Site Policy recommends placing @LibsOfTikTok ([LTT] 1.3M followers, not verified) in a 7-day timeout at the account level [meaning, not for a specific Tweet] based on the account’s continued pattern of indirectly violating Twitter’s Hateful Conduct Policy by tweeting content that either leads to or intends to incite harassment against individuals and institutions that support LGBTQ communities. At this time, Site Policy has not found explicitly violative Tweets, which would result in a permanent suspension of the account.

This type of enforcement action [repeated 7-day timeouts at the account-level] will not lead to permanent suspension, however: should LTT engage in any other direct Tweet-level violations of any of Site Policy’s policies, we will move forward with permanent suspension.

Assessment

Since its most recent timeout, while LTT has not directly engaged in behavior violative of the Hateful Conduct policy, the user has continued targeting individuals/allies/supporters of the LGBTQIA+ community for alleged misconduct. The targeting of at least one of these institutions

Compare this to what happened when Raichik herself was doxxed on November 21, 2022. A photo of her home with her address was posted in a tweet that has garnered more than 10,000 likes.

But when Raichik informed Twitter that her address had been disseminated, she said that Twitter Support responded with this message: “We reviewed the reported content, and didn't find it to be in violation of the Twitter rules.”

No action was taken. The doxxing tweet is still up.

The bottom line was: If someone in SIP-PES didn’t like you, they could find a way to shut you down.

That was the subtext of this direct message from Yoel Roth to colleagues in early 2021, written in a nearly indecipherable Twitter-ese:

“A lot of times, SI [Twitter’s Site Integrity team] has used technicality spam enforcements as a way to solve a problem created by Safety [team at Twitter] under-enforcing their policies.”

In a follow-up message with a colleague, Roth said he was looking for ways to marginalize accounts that had fallen into disfavor without banning them outright. Banning a user, especially a prominent one with many followers, generated bad publicity. Possible workarounds included “disabling engagements” and “deamplification/visibility filtering.”

Roth claimed the sub rosa censorship amounted to a public service. “The hypothesis underlying much of what we’ve implemented,” he said, “is that if exposure to, e.g., misinformation directly causes harm, we should use remediations that reduce exposure.”

This was, to Roth’s mind, only the beginning. Referring to former Twitter CEO Jack Dorsey, he wrote: “We got Jack on board with implementing this for civic integrity in the near term, but we’re going to need to make a more robust case to get this into our repertoire of policy remediations - especially for other policy domains.”



The revelation of these surreptitious actions left Stanford’s Bhattacharya shaken. After reading [our initial Twitter thread](#), he took to Twitter to share his [thoughts](#). He opined that, in the absence of shadow banning—sorry, visibility filtering—the Covid lockdowns might have been applied differently. There would have been, in his view, less unhappiness, less isolation, less economic despair.

In a follow up tweet, he [said](#): “Still trying to process my emotions on learning that @twitter blacklisted me. The thought that will keep me up tonight: censorship of scientific discussion permitted policies like school closures & a generation of children were hurt.”

The big question now was: Could Musk be trusted? It’s not that the power had been broken up. The only thing that had changed was that it had shifted hands.

When we asked Musk whether the world should be worried about all this power being concentrated in his hands, he laughed. (He laughed a lot.) The old content-moderation teams—SRT-GET, SIP-PES—hadn’t been disbanded; the people on them had just changed.

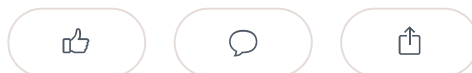
“I’m open to ideas,” Musk said. “I’ve got a lot of work on my plate. I was just worried that Twitter was sending civilization in a bad direction.”



We believe in honest, dogged, independent journalism. If you do, too, become a subscriber to The Free Press today:

Type your email...	Subscribe
--------------------	-----------

Tuesday, March 28, 2023



Comments 89
