THREAD BY BARI WEISS

‹ TWITTER FILES ›

#2 TWITTER'S SECRET BLACKLIST

THREAD: THE TWITTER FILES PART TWO.

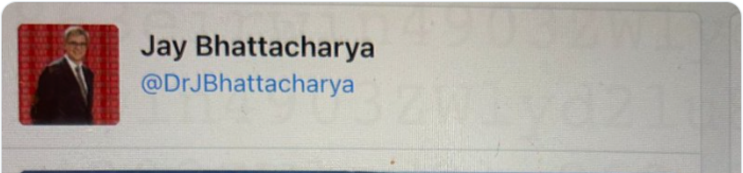
TWITTER'S SECRET BLACKLISTS.

@UNGATHEGREAT

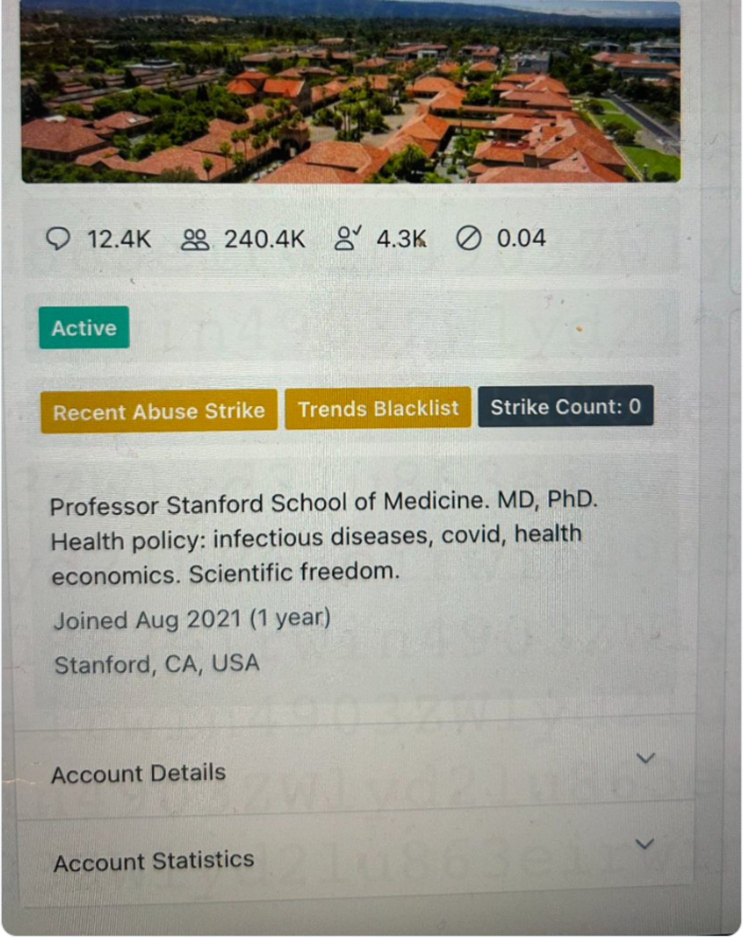
1. A new #TwitterFiles investigation reveals that teams of Twitter employees build blacklists, prevent disfavored tweets from trending, and actively limit the visibility of entire accounts or even trending topics—all in secret, without informing users.

2. Twitter once had a mission "to give everyone the power to create and share ideas and information instantly, without barriers." Along the way, barriers nevertheless were erected.

3. Take, for example, Stanford's Dr. Jay Bhattacharya (@DrJBhattacharya) who argued that Covid lockdowns would harm children. Twitter secretly placed him on a "Trends Blacklist," which prevented his tweets from trending.



Jay Bhattacharya
@DrJBhattacharya



12.4K 240.4K 4.3K 0.04

Active

Recent Abuse Strike

Trends Blacklist

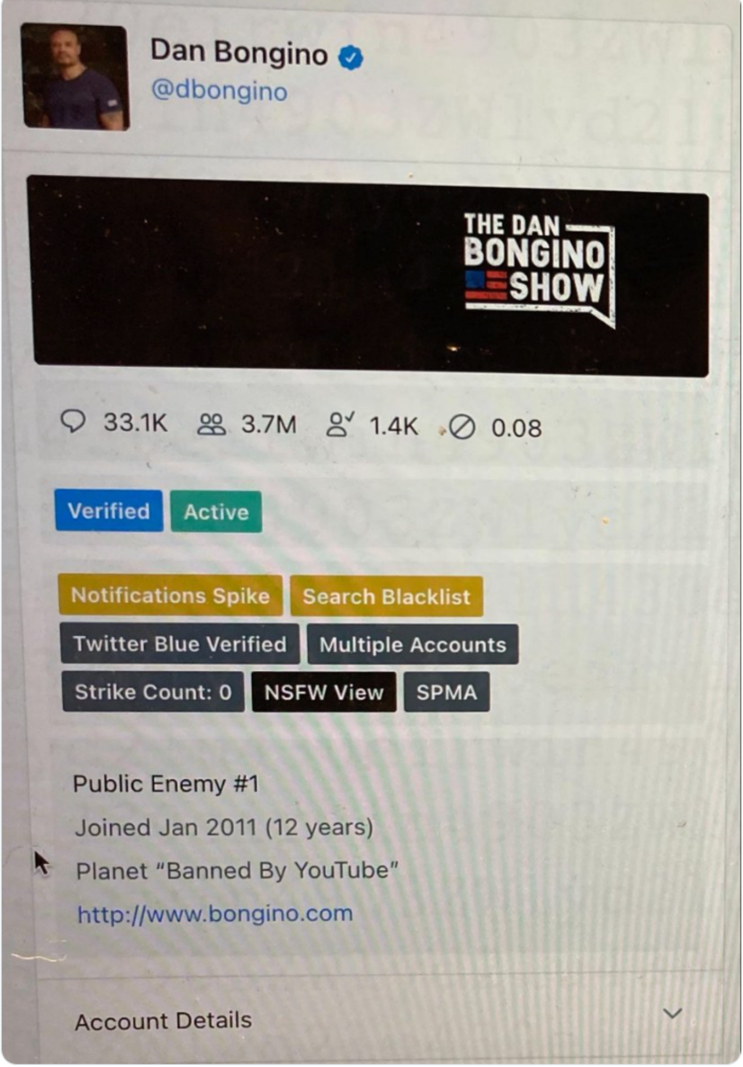
Strike Count: 0

Professor Stanford School of Medicine. MD, PhD. Health policy: infectious diseases, covid, health economics. Scientific freedom.

Joined Aug 2021 (1 year)

Stanford, CA, USA

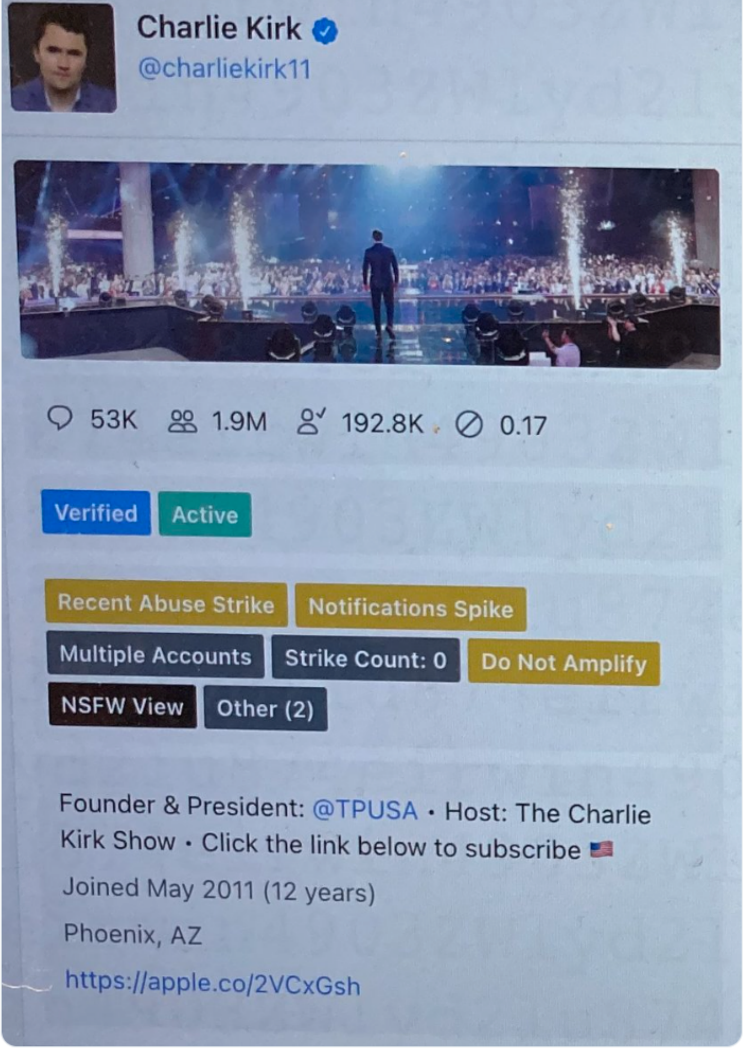
4. Or consider the popular right-wing talk show host, Dan Bongino (@dbongino), who at one point was slapped with a “Search Blacklist.”



Dan Bongino
@dbongino

Public Enemy #1
Joined Jan 2011 (12 years)
Planet “Banned By YouTube”
http://www.bongino.com

5. Twitter set the account of conservative activist Charlie Kirk (@charliekirk11) to “Do Not Amplify.”



Charlie Kirk

@charliekirk11

53K 1.9M 192.8K 0.17

Verified Active

Recent Abuse Strike Notifications Spike

Multiple Accounts Strike Count: 0 Do Not Amplify

NSFW View Other (2)

Founder & President: @TPUSA • Host: The Charlie Kirk Show • Click the link below to subscribe 🇺🇸

Joined May 2011 (12 years)

Phoenix, AZ

https://apple.co/2VCxGsh

6. Twitter denied that it does such things. In 2018, Twitter's Vijaya Gadde (then Head of Legal Policy and Trust) and Kayvon Beykpour (Head of Product) said: "We do not shadow ban." They added: "And we certainly don't shadow ban based on

political viewpoints or ideology."

7. What many people call "shadow banning," Twitter executives and employees call "Visibility Filtering" or "VF." Multiple high-level sources confirmed its meaning.

8. "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.

9. "VF" refers to Twitter's control over user visibility. It used VF to block searches of individual users; to limit the scope of a particular tweet's discoverability; to block select users' posts from ever appearing on the "trending" page; and from inclusion in hashtag searches.

10. All without users' knowledge.

11. "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 engineer told us. Two additional Twitter employees confirmed.

12. The group that decided whether to limit the reach of certain users was the Strategic Response Team - Global Escalation Team, or SRT-GET. It often handled up to 200 "cases" a day.

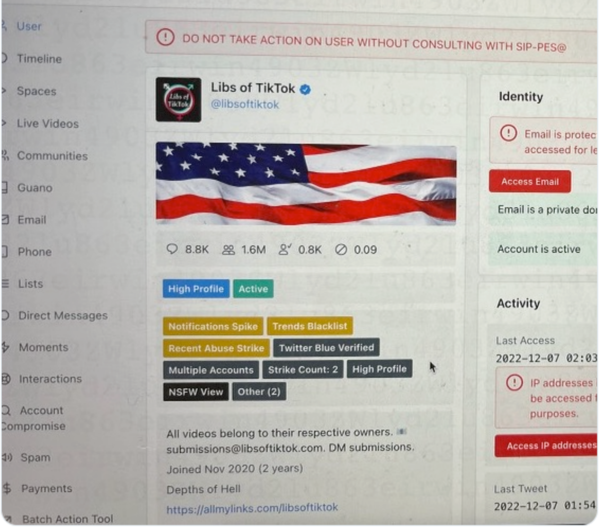
13. But there existed a level beyond official ticketing, beyond the rank-and-file moderators following the company's policy on paper. That is the "Site Integrity Policy, Policy Escalation Support," known as "SIP-PES."

14. This secret group included Head of Legal, Policy, and Trust (Vijaya Gadde), the Global Head of Trust & Safety (Yoel Roth), subsequent CEOs Jack Dorsey and Parag Agrawal, and others.

15. This is where the biggest, most politically sensitive decisions got made. "Think high follower account, controversial," another Twitter employee told us. For these "there would be no ticket or anything."

16. One of the accounts that rose to this level of scrutiny was @libsoftiktok—an account that was on the "Trends Blacklist" and was designated as "Do Not Take Action on User Without Consulting With SIP-PES."

ATEGORY
Testing



All videos belong to their respective owners. submissions@libsoftiktok.com. DM submissions.

Joined Nov 2020 (2 years)

Depths of Hell

https://allmylinks.com/libsoftiktok

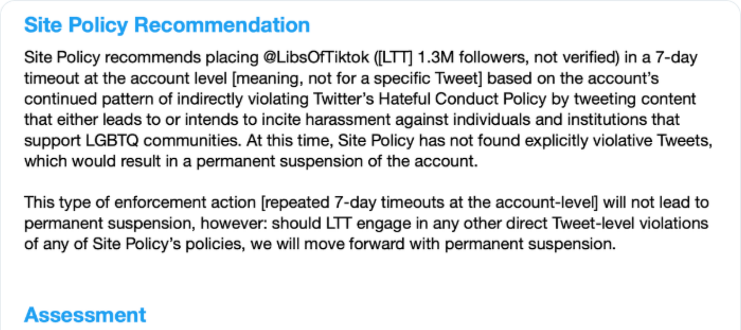
Last Access 2022-12-07 02:03

Last Tweet 2022-12-07 01:54

17. The account—which Chaya Raichik began in November 2020 and now boasts over 1.4 million followers—was subjected to six suspensions in 2022 alone, Raichik says. Each time, Raichik was blocked from posting for as long as a week.

18. Twitter repeatedly informed Raichik that she had been suspended for violating Twitter's policy against "hateful conduct."

19. But in an internal SIP-PES memo from October 2022, after her seventh suspension, the committee acknowledged that "LTT has not directly engaged in behavior violative of the Hateful Conduct policy." See here:

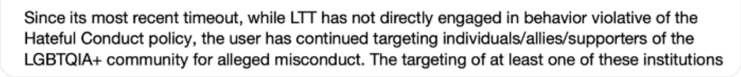


Site Policy Recommendation

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.

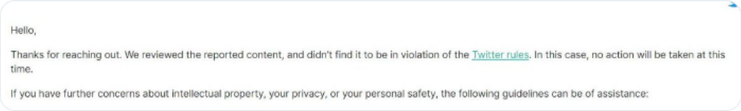
Assessment


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

20. The committee justified her suspensions internally by claiming her posts encouraged online harassment of “hospitals and medical providers” by insinuating “that gender-affirming healthcare is equivalent to child abuse or grooming.”

21. 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.

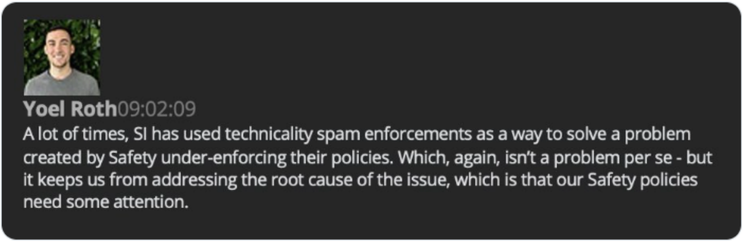
22. When Raichik told Twitter that her address had been disseminated she says Twitter Support responded with this message: "We reviewed the reported content, and didn't find it to be in violation of the Twitter rules." No action was taken. The doxxing tweet is still up.


Hello,

Thanks for reaching out. We reviewed the reported content, and didn't find it to be in violation of the Twitter rules. In this case, no action will be taken at this time.

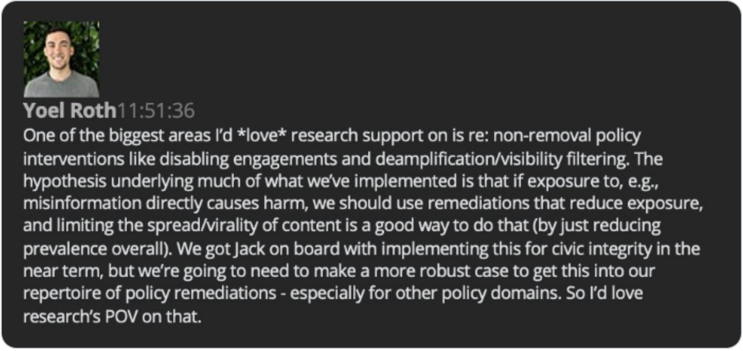
If you have further concerns about intellectual property, your privacy, or your personal safety, the following guidelines can be of assistance:

23. In internal Slack messages, Twitter employees spoke of using technicalities to restrict the visibility of tweets and subjects. Here’s Yoel Roth, Twitter’s then Global Head of Trust & Safety, in a direct message to a colleague in early 2021:


Yoel Roth09:02:09

A lot of times, SI has used technicality spam enforcements as a way to solve a problem created by Safety under-enforcing their policies. Which, again, isn’t a problem per se - but it keeps us from addressing the root cause of the issue, which is that our Safety policies need some attention.

24. Six days later, in a direct message with an employee on the Health, Misinformation, Privacy, and Identity research team, Roth requested more research to support expanding “non-removal policy interventions like disabling engagements and deamplification/visibility filtering.”



Yoel Roth 11:51:36

One of the biggest areas I'd *love* research support on is re: non-removal policy interventions like disabling engagements and deamplification/visibility filtering. The hypothesis underlying much of what we've implemented is that if exposure to, e.g., misinformation directly causes harm, we should use remediations that reduce exposure, and limiting the spread/virality of content is a good way to do that (by just reducing prevalence overall). 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. So I'd love research's POV on that.

25. Roth wrote: "The hypothesis underlying much of what we've implemented is that if exposure to, e.g., misinformation directly causes harm, we should use remediations that reduce exposure, and limiting the spread/virality of content is a good way to do that."

26. He added: "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."

27. There is more to come on this story, which was reported by @AbigailShrier @ShellenbergerMD @NellieBowles @IsaacGrafstein and the team The Free Press @TheFP.

Keep up with this unfolding story here and at our brand new website: thefp.com.

THE FREE PRESS

The Free Press

A new media company built on the ideals that were once the bedrock of American journalism.

http://thefp.com

28. The authors have broad and expanding access to Twitter's files. The only condition we agreed to was that the material would first be published on Twitter.

29. We're just getting started on our reporting. Documents cannot tell the whole story here. A big thank you to everyone who has spoken to us so far. If you are a current or former Twitter employee, we'd love to hear from you. Please write to: tips@thefp.com

30. Watch @mtaibbi for the next installment.