Blacklist policies at Twitter

cnav.news/2022/12/09/accountability/news-media/blacklist-policies-twitter/

By Terry A. Hurlbut December 9, 2022

The second Twitter Files thread has finally dropped, after a delay that included flushing out and firing Jim Baker. This thread details a secret blacklist policy, its unwritten rules, who enforced it, and who fell victim to it. Tellingly, unlike the Hollywood Blacklist of the Fifties, those who enforced this blacklist *never informed the targets*. This, then, is the origin of the verb *to shadowban*. Franz Kafka (*The Trial*, *Metamorphosis*) would no doubt have enjoyed putting this into a novel. Except that *we* are the Josef Ks and the Gregor Samsas of this drama.

The Secret Blacklist Policy

Bari Weiss, founder and editor of *The Free Press*, released the story beginning at 7:20 p.m. EST (4:20 p.m. PST).

> 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.
>
> — Bari Weiss (@bariweiss) December 9, 2022

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. pic.twitter.com/qTW22Zh691

— Bari Weiss (@bariweiss) December 9, 2022

5. Twitter set the account of conservative activist Charlie Kirk (@charliekirk11) to "Do Not Amplify." pic.twitter.com/dOyQIVdsW2

— Bari Weiss (@bariweiss) December 9, 2022

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

— Bari Weiss (@bariweiss) December 9, 2022

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.

— Bari Weiss (@bariweiss) December 9, 2022

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.

— Bari Weiss (@bariweiss) December 9, 2022

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."

— Bari Weiss (@bariweiss) December 9, 2022

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."

— Bari Weiss (@bariweiss) December 9, 2022

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.

— Bari Weiss (@bariweiss) December 9, 2022

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: pic.twitter.com/d9FGhrnQFE

— Bari Weiss (@bariweiss) December 9, 2022

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.

— Bari Weiss (@bariweiss) December 9, 2022

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: pic.twitter.com/Li7HDZJtlJ

— Bari Weiss (@bariweiss) December 9, 2022

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.”

— Bari Weiss (@bariweiss) December 9, 2022

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: https://t.co/qYaBJzKcZj.

— Bari Weiss (@bariweiss) December 9, 2022

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

— Bari Weiss (@bariweiss) December 9, 2022

30. Watch @mtaibbi for the next installment.

— Bari Weiss (@bariweiss) December 9, 2022

This thread mentions three prominent target names (Libs of TikTok, Dan Bongino, and Charlie Kirk) and one not-so-prominent one (Dr. Jay Bhattacharya). The tweets mentioning them have embedded screencaps of what look like moderational consoles. Note the settings: “Trends Blacklist,” “Search Blacklist,” and “Do Not Amplify.” And at Libs of TikTok:

> Do Not Take Action on User Without Consulting With SIP-PES.

The particular speech Twitter sought so suppress included criticism of COVID lockdowns, and objections to child grooming.

Note also how high this policy went: to Yoel Roth, Vijaya Gadde’s immediate successor as Officer in Charge of Legal Policy, Trust and Safety. Matt Taibbi quoted this same Yoel Roth as referring to “actual Nazis in the [Trump] White House.” Yesterday the newsletter *NewsHouse* carried this article directly critical of Roth and his open advocacy of censorship.

Roth has also come in for direct criticism on Twitter itself, now that Elon Musk has apparently destroyed his regime. (*Warning.* The tweets below mention some highly family-unfriendly concepts. *Parental judgment and discretion are advised.*)

> This was the guy who 10 years ago, GOP pundits were saying would " get a taste of the real world" and be "shocked back to reality" when he got out of college. Instead he banned the president from twitter because he hated conservatives and white people.
>
> — Märtin (@davidcrockettfa) December 3, 2022

This tweet from Jordan Peterson is somewhat more family-friendly.

> The ex safety chief (and censor) of Twitter is absolutely everything you'd dream he'd be @yoelroth another appalling creation of modern higher ed https://t.co/et92oyl4LA
>
> — Dr Jordan B Peterson (@jordanbpeterson) December 4, 2022

The *NewsHouse* piece reveals that Roth actually said that any criticism of gender switchers threatens. Their. *Lives*.

Reaction

Reactions to this secret blacklist policy vary from:

- Outrage, to
- “I told you so,” to
- “Ho, hum,” to
- “Oh, so it’s *that* again!” to
- “That’s a lie!” to
- “They had every right to do it,” to

- “You/they deserve what you/they got!” to
- “It happened to leftists, too!”

Here’s an exchange in the “You deserve it!” category.

> So pretty much everyone in the CDC, the WHO and most of the government officials in the last 3 years should be blacklisted. Cause just about every one of them lied and misled the public on at least one occasion, and SHOULD have known better at the time they made BS statements
>
> — Rocketman (@rocketman_c) December 9, 2022

Remember: Twitter quietly rescinded and literally deleted its “COVID Misleading Information Policy.” Now that Twitter is reinstating many more accounts (including Laura Loomer), those accounts will not be subject to that policy.

Nor has Elon Musk been lacking in attention. Last night he promised to update a user’s account status display to *tell* a user when such moderational action applies.

> Twitter is working on a software update that will show your true account status, so you know clearly if you’ve been shadowbanned, the reason why and how to appeal
>
> — Elon Musk (@elonmusk) December 9, 2022

He also shared Vijaya Gadde’s denials,

> 😕 pic.twitter.com/CVfoK7Ljch
>
> — Elon Musk (@elonmusk) December 9, 2022

Yoel Roth’s worst excesses,

> Former head of censorship at Twitter was perhaps not entirely unbiased … pic.twitter.com/yynb9whc5S
>
> — Elon Musk (@elonmusk) December 9, 2022

an assurance that Twitter *does not* retain Perkins Coie as outside counsel,

> Twitter isn’t using Perkins Coie. No company should use them until they make amends for Sussman’s attempt to corrupt a Presidential election.
>
> — Elon Musk (@elonmusk) December 9, 2022

and a promise to address an alleged continuation of a Trends Blacklist against Libs of TikTok.

> Looking into it
>
> — Elon Musk (@elonmusk) December 9, 2022

The skeptics

But the reaction also includes skepticism. Why, for instance, does Elon Musk still refer to “hate speech?” What exactly will constitute it, moving forward? And why not disable these moderational tools now? Today? This instant?

> He's said that "hate speech will lead to a user not being viewed as much" or something similar, which I hate because we still don't have a consistent definition of the speech. I understand it can't be a free for all, but the system should at least be transparent.
>
> — Bear Market Enjoyer (@jeffleb89495079) December 9, 2022

Over on Telegram, Andrew Torba naturally weighed in:

> Elon is dropping the Twitter files out one side of his mouth while bragging about how much they have shadowbanned “hate speech” out the other. Centrist “moderates” are unprincipled losers who cannot be trusted.

Recall that Andrew Torba defines “hate speech” as anything that violates U.S. federal or State law. Torba also boasted of new “tools” for Gab users that can detect shadowbanning elsewhere.

Torba has a *business reason* to refuse to give Elon Musk any credit. But perhaps the other skeptics should remember that Twitter is like a battleship, and can’t steer like a PT boat. The changes at Twitter are still an improvement, whether they represent perfection or not. Banned accounts *are* coming back. And the Twitter Files releases are far from ended.