Twitter Thread by 12 Foot Tall Giant Alexandra Erin



12 Foot Tall Giant Alexandra Erin
@AlexandraErin

LRT: One of the problems with Twitter moderation - and I'm not suggesting this is an innocent cause that is accidentally enabling abuse, but rather that it's a feature, from their point of view - is that the reporting categories available for us do not match up to the rules.

Now, Twitter's actual policy is that wishes or hopes for death or harm are the same as threats. That policy has been in place for years. But there's no report category for hoping someone dies. You can only report it as a threat.

Which gives the moderator, who doesn't spend long on any individual tweet, mental leeway to go, "Well, there's no threat here." and hit the button for "no violation found".

They have a rule that says persistent misgendering or other dehumanizing language is not tolerated, but again - there is no reporting category for that. We have to report it as hate against a group.

So again, the moderator looks at that tweet, briefly in isolation, and sees what, without context, might look neutral or matter of fact. A series of tweets referring to somebody consistently by the same set of pronouns or a statement that somebody is a man or woman.

I've said this before, but having a rule against misgendering or otherwise dehumanizing trans people and not enforcing it is worse than having no rule.

Because the rule's existence creates the impression that we have protections we don't.

So the people who dehumanize, misgender, and wish death upon us get the best of both worlds - they can freely do it over and over again while proclaiming themselves censored martyrs to free speech. They can use the "power" our supposed "protected status" gives us to foment hate.

There rules, their reporting tools, and their rulings all ultimately feel like they are each created/run by a different group of people who not only don't agree but haven't communicated with each other about what they're doing.

But again, that makes it sound like it's an innocent, well-intentioned mess and even if at one point it started out that way (and I'm not saying that it did, I'm saying *if*) at this point it's been going on so long and has been pointed out to them so many times, it's deliberate.

It is a deliberate choice to keep running their system this way.

Meanwhile, people who aren't acting in good faith can, will, and DO game the automated aspects of the system to suppress and harm their targets.

They can run coordinated and/or bot-assisted mass reporting campaigns to make sure their complaints get escalated or the system automatically steps in and locks accounts.

Another side of this is that, for peoples who have historically been targeted for death, there are all sorts of references that are ready-made for making EXPLICIT DEATH THREATS that to an untrained moderator looking at a tweet in isolation, might just seem like absurdism.

E.g., references to ways people died or had their corpses abused in the Holocaust, in slavery or Jim Crow America. References to lynching, to atomic bombings, to drone strikes.

And then, then we come to the fact that the people making the moderation decisions are making decisions, even on the stuff that "will not be tolerated".

A death threat is not supposed to be allowed on here even if it's a joke. That's Twitter's premise, not mine.

But a sizable chunk of Twitter's moderation pool has a hard time looking at, say, a straight white man threatening violence upon a woman, a gay person, a trans person, etc., and seeing it as serious. It's like background radiation. It's always there. Not alarming.

But anger, even without an explicit threat, from those groups directed against more powerful ones... that's alarming to the same people.

It's the Joker Principle. I know we're all sick of pop culture exegesis but I just can't let go of this one: "Nobody panics when things go according to plan."

A guy going "Haha get raped." is part of the plan. It's normal.

His target replying "FUCK OFF" is not. It's radical.

Things that strike the moderator as unusual, as radical, as alarming are more likely to get moderated.

Things that strike the moderator as "That's just how it is on this bitch of an earth." get a pass.

Helicopter rides. A trip to the lampshade factory.

And then the ultra-modern ones like "Banned from Minecraft in real life."

https://t.co/5xdHZmqLmM

> Helicopter rides.
>
> — azteclady (@HerHandsMyHands) October 3, 2020

And needless to say, all of this "confusion" and subjectivity in what are supposedly objective, zero tolerance rules that apply to everybody... they give people who *want* to protect and promote fascism and violence through moderation a lot of cover.