Why do companies with unbounded resources still have terrible moderation?

moultano.wordpress.com/2019/10/02/why-do-companies-with-huge-resources-still-have-terrible-moderation/

Ryan Moulton October 2, 2019

Disclaimer: I work at Google, but none of this reflects the opinion of my employer, my experiences at work, or describes any product I've worked on or even have indirect knowledge about. I'm speaking purely in hypotheticals based on my experiences with human labelled data and machine learning, and am not describing any specific event that has happened to me. Do not interpret this as insider information about YouTube, I have never worked on YouTube and have no knowledge of its policies or systems. Also, if you don't want to see some really revolting racism, don't click the links.

Social media companies have moderation SNAFUs on a near daily basis. People get suspended for joking with their friends. People get suspended for clap backs at Nazis. And the Nazis keep on Nazi-ing and no one seems to care. Inevitably, the public and the tech press reacts with incredulity, "This company has *so much money*. They've hired *thousands of PhDs*. How can they still be so bad at this?"

I'm going to tell you how.

Pretend you work at Twitter.

Twitter disallows threats of violence. You want to enforce this as uniformly as you can, so you want to build an automated system that detects threats of violence and suspends users.

First step, you hire a team of PhDs that can build state of the art transformer-based language models. This is what the tech press thinks is the hard part, and that after you've done that, it should be a cakewalk right?

This is the easiest part of the process, because all you have to do is spend money to do it. Language models are a commodity now. You can download the best in the world from GitHub. The expertise to adapt them to a particular task is still rare and expensive, but rapidly becoming more common. But models by themselves are useless. What you need is training data. How do you get training data for what is, and what isn't, a threat of violence?

Typically, you'd write up some guidelines for what is and what isn't a threat, iterate on those with the legal, policy, and PR departments, and then contract with some external company to have human beings read those guidelines and rate lots of examples that you send them.

Writing guidelines for what is and isn't a threat is really hard. The people who write these guidelines often have backgrounds in Law or Philosophy and lots of experience with the subtleties, but governments have hundreds of pages of case law covering this specific thing, and experts still disagree all the time. When governments need to figure out whether something is or isn't a threat, they take months of adjudication and 10s of thousands of dollars. You have to be able to do it at the rate of QPS.

Unsurprisingly, when you get your ratings back from the contractor, they've labelled tons of sarcasm and jokes as threats, and tons of outright threats that are slightly hidden behind euphemisms, memes, or slang as benign. So in the next round of guideline revisions you realize you have to teach people about memes.

Imagine that the archetypal rater you're trying to teach is someone in rural India. I'm using rural India as an example not because they're actually from rural India, but if I describe them in a way that seems less foreign to you, you will instinctively overestimate how much you have in common with them. They are a random English speaking person interested in a flexible part time job. They are from somewhere that is not where you're from and probably have very little cultural context in common with you. If you are reading this, they are probably "less online" than you. They might be a stay at home mom in Minnesota who needs some extra cash while the kids are at school. They might be a security guard working the night shift with a lot of time to stare at a screen. They might be someone in India. These people are different from you, and their understanding of the world and the meme wars is different from yours. Your sub-culture is deeply aware of how 8chan white supremacists talk about the people they're going to swat because you see that shit every day. But theirs isn't. Imagine how well you'd do if you were asked to rate hate speech directed at the Rohingya. There are slurs out there that you couldn't even imagine, and they change from moment to moment.

So now, as you rewrite the guidelines, you're providing a lot of links to urban dictionary and knowyourmeme. You're explaining what 4chan is. You're explaining what 8chan is. Your guidelines are becoming less and less general and more and more extremely online. You realize that they'll all be obsolete in 6 months. 4chan finds new ways to be racist way faster than your legal department can approve new documents, likely faster than you can write them or the raters can read them. Rival communities are now weaponizing your "Report Abuse" button, so you can't even get any indirect signal from what your users think in aggregate without becoming somebody's tool.

Your users expect the system to behave perfectly, or at least non-embarrassingly. Your users also expect everything to be interpreted in a cultural context that they're deeply steeped in, but that even a random person from the US can't figure out without a ton of research and training, let alone someone from the Philippines.

You're realizing that your ideal rater is a multilingual extremely-online expert in first amendment law who takes a morbid interest in the dregs of the internet, but there's only one Popehat and he's already gainfully employed.

Nazis

Let's say you haven't had enough suffering from trying to do something as seemingly unambiguous as getting rid of outright threats of violence. Now you want to go even further to do what lots of Twitter users have been begging for and ban the Nazis. So you hold your nose, roll up your sleeves, and write some guidelines about Nazis.

You get your first ratings back and you realize you need to describe the difference between 1488 and 1619, that one is a white supremacist slogan and the other is an important date in American history. You update the guidelines with policy, PR, and legal, and then you try again.

On the next round you realize you need to teach them the difference between "Le 56%" and "the 1%." That one of them is how European white nationalists make fun of American white nationalists, and the other is how the left talks about rich people in America. So you update the guidelines with policy, PR, and legal, and try again.

On the next round you realize that they haven't quite figured out what antisemitism looks like, and you need to explain the difference between "Jews have infiltrated the government" and "Russians have infiltrated the Trump campaign," that one is a justification for mass murder, and the other is something that kinda, maybe, happened. So you update the guidelines. Legal and policy are all increasingly concerned that if any of the contractors leak the guidelines that you're going to be hauled before congress. PR is worried about angry presidential tweets. You try again.

But there's still a lot of stuff they're getting wrong, so you try to explain the difference between "flashing an OK sign upside down at a gun range" vs "flashing an OK sign right-side up on a ski slope," that one might be a white supremacist showing a "white power" sign, maybe, and the other is a guy enjoying his hobby. Most of your raters have never been to a gun range or to a ski slope, and may have never seen the OK sign. Now PR is sure that if any of this leaks not only will you get hauled before congress, but the general public will think you're a bunch of loons.

This process can go on literally forever, and the ratings may never materially improve. The world of racism is full of euphemism specifically designed to fool onlookers, to allow people to be racist in public without detection, to be joking and not joking at the same time for plausible deniability. Racists have already inoculated their movement against exactly the type of social/technical system you are trying to design.

How do you know it sucks?

You've built the best system you can. You may know it sucks, but you need a solution today, there isn't any obvious way to improve it, so you deploy it and hope for the best. Despite investing 10s of millions of dollars and years of many people's lives, you have a system that still causes public relations nightmares on a continual basis.

It's difficult at an aggregate level to even figure out what the system is getting wrong. Sure, you can find individual examples, they're in your inbox every day, usually as a flood of vitriol and accusations, or worse, in the press. You can try to cut around them, to add them as examples to your guidelines the next time you do a round of rating, but what you really need are aggregate statistics. Maybe these are all just outliers and things are fine on average? But how do you get aggregate statistics? You need humans to rate things again. And then you're back in the same morass of teaching rural Indians about 4chan.

You'd love to explain to the public how difficult this all is, how hopeless it feels, and what an impossible standard they're holding you to. But PR knows that making any comment less tepid than "we're working to continually improve," will kick off another press cycle, and so the train wreck rolls on.

What can be done?

If I had a real solution, I'd be taking it to the bank instead of writing this, but I'll take a stab at back-seat product design anyways.

The bottom line is that people cannot effectively moderate a culture and a context that they are not a part of. Maybe that means that being "extremely online" should be part of the hiring criteria. Maybe that means hiring people specifically from each subculture in your network. Or maybe it means turning the problem around, and giving the power to the users.

Give communities the tools to moderate themselves at scale. Give users (or advertisers) the tools to control their own experience at scale. It may be impossible to teach someone with specificity what is and isn't doxxing or what the early signs of getting targeted by the alt-right are, but once you've experienced them, or watched it happen to someone, you'll never forget. This means that users may need to see more data about other users than you are comfortable exporting. If it's a signal you think your spam or abuse team might need, show it to your users. Let them use it themselves, or even script it themselves, so they can share protection techniques the same way they share everything else. Know that perfect content moderation is unattainable, and your users cannot wait for you to improve. They need the tools now, and they will be better protecting themselves than anything you develop as an outsider.

Because if you're still imagining that if only you had better ratings you could fix this, imagine what would it look like to have a system that could do content moderation perfectly, that understood the nuances of every culture on earth, that could see the intention behind everything that everyone could write and would know whether it was meant in jest or in hate, that could really see into our hearts, that would know if we deserve forgiveness.

It would look like a lot like a God, and I think that's a bit much to expect out of any group of people, let alone a big pile of linear algebra.