Vox

Social media platforms aren't equipped to handle the negative effects of their algorithms abroad. Neither is the law.

Because of one law, the internet has no legal duty of care when it comes to hate speech. Just take a look at what happened in Myanmar.

by **Tiffany Ng**

Updated May 9, 2024, 8:58 PM GMT+8

Tiffany Ng is a New York-based freelance tech and culture writer, exploring the ever-changing dynamics between capitalism, technology, and everyday life.

Just after the clock struck midnight, a man entered a nightclub in Istanbul, where hundreds of revelers welcomed the first day of 2017. He then swiftly shot and killed 39 people and injured 69 others — all on behalf of the Islamic State of Iraq and Syria (ISIS).

Among those killed was Jordanian citizen Nawras Alassaf. In response, his family filed a civil suit later that year against Facebook, Twitter, and Google, which owns YouTube. They believed that these tech companies knowingly allowed ISIS and its supporters to use each platform's "recommendation" algorithms for recruiting, fundraising, and spreading propaganda, normalizing radicalization and attacks like the one that took their son's life.

Their case, *Twitter v. Taamneh*, argued that tech companies profit from algorithms that selectively surface content based on each user's personal data. While these algorithms neatly package recommendations in newsfeeds and promoted posts, continuously serving hyper-specific entertainment for many, the family's lawyers argued that bad-faith actors have gamed these systems to further extremist campaigns. Noting Twitter's demonstrated history of online radicalization, the suit anchored on this question: If social media platforms are being used to promote terrorist content, does their failure to intervene constitute aiding and abetting?

The answer, decided unanimously by the Supreme Court last year, was no.

The Court insisted that without ample evidence that these tech companies offered explicit special treatment to the terrorist organization, failure to remove harmful content could not constitute "substantial assistance." A similar case in the same Supreme Court term, *Gonzalez v. Google,* detailing a 2015 ISIS attack in Paris, shared the same decision as *Twitter v. Taamneh.*

Both decisions hinged on 26 words, stemming from a nearly three-decades-old law: "[N]o provider or user of an interactive computer service shall be treated as the publisher or speaker of any information provided by another information content provider."

Known as Section 230 of the Communications Decency Act, the law fundamentally encoded the regulation — or lack of it — of speech on the internet. According to the logic of Section 230, which dates back to 1996, the internet is supposed to act something like a bookstore. A bookstore owner isn't responsible for the content of the books they sell. The authors are. It means that while online platforms are free to moderate content as they see fit — just as a bookstore owner can choose whether or not to sell certain books — they are not legally responsible for what users post.

Such legal theory made sense back in 1996, when fewer than 10 million Americans were regularly using the internet and online speech had very little reach, be it a forum post or a direct message on AOL. That's simply not the case today, when more than 5 billion people are online globally and anything on the internet can be surfaced to people who weren't the intended audience, warped, and presented without context.

emerged.

But when a thirst for personalization exacerbates existing social tensions, it can amplify potential harm. It's no surprise that the US, where social media platforms have intensified partisan animosity, has experienced one of the largest rising political polarization levels in a developed democracy over the past four decades. And given how most platforms are based in the US and prioritize English speakers, moderation for other languages tends to be neglected, especially in smaller markets, which can make the situation even worse outside the US.

Investments follow competition. Without it, ignorance and negligence find space to thrive.

Such myopic perspectives end up leaving hate speech and disinformation undetected in most parts of the world. When translation algorithms fail, explicitly hateful speech slips through the cracks, not to mention more indirect and context-dependent forms of inciting content. Recommendation algorithms then surface such content to users with the highest likelihood of engagement, ultimately fueling further polarization of existing tensions.

Speech is not the crux of the issue; where and how it appears is. A post may not directly call for the death of minorities, but its appearance in online groups sharing similar sentiments might insinuate that, if not help identify people who might be interested in enacting such violence. Insular social media communities have played a sizable role in targeted attacks, civil unrest, and ethnic cleansing over the past decade, from the deadly riots that erupted from anti-Muslim online content in Sri Lanka to the targeted killings publicized online in Ethiopia's Tigray War.

Of course, the US Supreme Court doesn't have jurisdiction over what a person in another country posts. But what it has effectively accomplished through Section 230 is a precedent of global immunity for social media companies that, unlike the Court, do act globally. Platforms can't be held responsible for human rights abuses, even if their algorithms seem to play a role in such atrocities.

One notable instance would be Meta's alleged role in the 2017 Rohingya genocide, when Facebook's recommendation algorithms and targeted advertising amplified hateful narratives in Myanmar, playing what the UN later described as a "determining role" in fueling ethnic strife that instigated mass violence against the Rohingya Muslim minority in Myanmar's Rakhine state. While the company has since taken steps to improve the enforcement of its community standards, it continues to escape liability for such disasters under Section 230 protection.

One thing is clear: To see regulation only as an issue of speech or content moderation would mean disregarding any and all technological developments of the past two decades. Considering the past and ongoing social media-fueled atrocities, it is reasonable to assume that companies know their practices are harmful. The question initially posed by *Twitter v. Taamneh* then becomes a two-parter: If companies are aware of how their platforms cause harm, where should we draw the line on immunity?

Myanmar's walled garden and the many lives of online speech

The rapid adoption of Facebook when it entered Myanmar in the 2010s offers a poignant example of the pitfalls of unbridled connectivity.

Until fairly recently, Myanmar was one of the least digitally connected states on the planet. Its telecommunications market was largely state-owned, where government censorship and propaganda were prevalent. But in 2011, the deregulation of telecommunications made phones and internet access much more accessible, and Facebook found instant popularity.

"People were using Facebook because it was well-suited to their needs," anthropologist Matt Schissler said. By 2013, Facebook was Myanmar's de facto internet, Schissler added. In 2016, the Free Basics program, an app that provided "free" internet access via a Facebook-centric version of the internet, was launched.

human rights abuses — and in particular, a record of discrimination against Muslim populations since at least 1948, when Myanmar, then called Burma, gained independence. As a result, the Rohingya — the largest Muslim population in the country — have long been a target of persecution by the Myanmar government.

In the process of connecting millions of people in just a few years, anthropologists and human rights experts say Facebook inadvertently helped exacerbate growing tensions against the Rohingya. It took very little time for hateful posts — often featuring explicit death threats — to proliferate.

Then came the Rohingya genocide that began in 2017 — an ongoing series of military-sanctioned persecutions against the Rohingya that have resulted in over 25,000 deaths and an exodus of over 700,000 refugees. Anti-Rohingya posts on Facebook were gaining traction, and at the same time, reports from the Rohingyas of rape, killings, and arson by security forces grew. Myanmar's military and Buddhist extremist groups like the MaBaTha were among the many anti-Muslim groups posting false rape accusations and calling the Rohingya minority "dogs," among other dehumanizing messages.

In a 2022 report, Amnesty International accused Facebook's newsfeed ranking algorithms of acting to significantly amplify hateful narratives, actively surfacing "some of the most egregious posts advocating violence, discrimination, and genocide against the Rohingya."

The Amnesty International report heavily referenced findings from the UN's Independent International Fact-Finding Mission on Myanmar, outlining how Facebook's features, along with the company's excessive data-mining practices, not only enabled bad-faith actors to target select groups but also created financial incentives for anti-Rohingya clickbait.

"Facebook's signature features played a central role in the creation of an environment of violence," said Pat de Brún, the report's author and the head of big tech accountability and deputy director at Amnesty International. "From the Facebook Files leaked by Frances Haugen, we found that Facebook played a far more active and substantial role in facilitating and contributing to the ethnic cleansing of Rohingya."

Facebook, hosting nearly 15 million active users in Myanmar at the time, also operated with a malfunctioning translation algorithm and only four Burmese-speaking content moderators — a disastrous combination. Drowning in the sheer quantity of posts, moderators more often than not failed to detect or remove the majority of the explicitly anti-Rohingya disinformation and hate speech on its platform.

In one case, a post in Burmese that read: "Kill all the kalars that you see in Myanmar; none of them should be left alive," was translated to "I shouldn't have a rainbow in Myanmar," by Facebook's English translation algorithm. ("Kalar" is a commonly used slur in Myanmar for people with darker skin, including Muslims like the Rohingya.) If a moderator who encountered such a post wasn't one of the company's four Burmese speakers, a post that's equally if not more inflammatory would go undetected, freely circulating.

Facebook's reported failure to detect hate speech was only one small part of the platform's role in the Rohingya genocide, according to the report. Facebook's recommendation algorithms acted to ensure that whatever slipped through the cracks in moderation found an audience. According to Amnesty International's investigation, Facebook reportedly surfaced hateful content to insulated online communities seeking affirmations for their hateful positions — all in the service of engagement. Between Facebook's market entry and the mass atrocities of 2017, the UN's investigation found that some of the most followed users on the platform in Myanmar were military generals posting anti-Rohingya content.

Hate speech was not the only type of speech that engagement-optimizing algorithms amplified. "There's hate speech, but there's also fear speech," said David Simon, director of the genocide studies program at Yale University.

Forcing formerly neutral actors to take sides is a common tactic in genocidal campaigns, Simon said. Core to the Burmese military's information operations was "targeting non-Rohingya Burmese who had relationships with Rohingya people," Simon said. In doing so, militant groups framed violence against the Rohingya as acts of nationalism — and, consequently, inaction as treason. Reuters' 2018 investigation reported that individuals who resisted campaigns of hate were threatened and publicly targeted as traitors. By forcing affiliations, the Burmese military was able to normalize violence against the Rohingya.

“It’s not a matter of making everyone a perpetrator,” Simon told Vox. “It’s making sure bystanders stay bystanders.”

The context-dependent nature of fear speech manifested most notably in private channels, including direct texting and Facebook Messenger groups. In an open letter to CEO Mark Zuckerberg, six Myanmar civil society organizations reported a series of chain messages on Facebook’s messaging platform that were sent to falsely warn Buddhist communities of “jihad” attacks, while simultaneously notifying Muslim groups about anti-Muslim protests.

While hate speech, considered in isolation, explicitly violates Facebook’s community guidelines, fear speech, taken out of context, often does not. “Fear speech would not get picked up by automatic detection systems,” Simon said.

Nor can Meta claim it had no advance notice of what might unfold in Myanmar. Prior to the 2017 military-sanctioned attacks in northern Rakhine state, Meta reportedly received multiple direct warnings from activists and experts flagging ongoing campaigns of hate and cautioning of an emergent mass atrocity in Myanmar. These warnings were made as early as 2012 and persisted until 2017, taking shape in meetings with Meta representatives and conferences with activists and academics at Meta’s Menlo Park headquarters.

Meta, the parent company of Facebook, has published several reports in the years since about current policies and updates in Myanmar, including that it significantly increased investments there to help with moderation, in addition to banning the military (Tatmadaw) and other military-controlled entities from Facebook and Instagram.

The internet is nothing like a bookstore

The Rohingya are not recognized as an official ethnic group and have been denied citizenship since 1982. A majority of stateless Rohingya refugees (98 percent) live in Bangladesh and Malaysia. Being a population with little to no legal protection, the Rohingya have very few pathways for reparations under Myanmar law.

On the international stage, issues of jurisdiction have also complicated Meta's liability. Not only is Myanmar not a signatory of the Rome Statute, the treaty that established the International Criminal Court (ICC) to address acts of genocide, among other war crimes and crimes against humanity, the ICC is not designed to try corporations. Ultimately, the closest anyone can get to corporate accountability is in the US, where most of these platforms are based but are effectively protected under Section 230.

Section 230 was written for an internet that did not have recommendation algorithms or targeting capabilities, and yet, many platform regulation cases today cite Section 230 as their primary defense. The bill grounds itself in the analogy of a bookkeeper and a bookstore, which is now a far cry from the current state of our internet.

In the landmark First Amendment case *Smith v. California*, which involved a man convicted of violating a Los Angeles ordinance against possessing obscene books at a bookstore, the Supreme Court ruled in 1959 that expecting a bookstore owner to be thoroughly knowledgeable about all the contents of their inventory would be unreasonable. The court also ruled that making bookstore owners liable for the material they sell would drive precautionary censorship that ultimately limits the public's access to books.

The internet in 1996, much like a bookstore, had a diverse abundance of content, and then-Reps. Chris Cox and Ron Wyden, of California and Oregon respectively, saw a meaningful parallel. They decided to take the Court's bookstore analogy one step further when they framed Section 230: Not only should online platforms have free rein to moderate, but pitting websites with better, "safer" curations against each other would also create monetary incentives for moderation.

Today, the concentration of users on a handful of social media platforms shows that real competition is long gone. Social media companies, without such competition, lose incentive to maintain safe environments for site visitors. Instead, they're motivated to monetize attention and keep users on the platform for as long as possible, whether via invasive ad targeting or personalizing recommended information.

These developments have complicated the original analogy. If entering a platform like Facebook were akin to entering a bookstore, that bookstore would only have a personalized display shelf available, stocked with selections based on personal reading histories.

Today, the bounds of Section 230 are painfully clear, yet that law still effectively bars activist groups, victims, and even countries from trying to hold Meta accountable for its role in various human rights abuses.

Section 230 has prevented the landscape of platform regulation from expanding beyond a neverending debate on free speech. It continues to treat social media companies as neutral distributors of information, failing to account for the multifaceted threats of data-driven targeted advertising, engagement-based newsfeed rankings, and other threatening emergent features.

Although platforms do voluntarily enforce independently authored community guidelines, legally speaking, there is little to no theory of harm for social media platforms and thus no duty of care framework. In the same way landlords are responsible for providing lead-free water for their tenants, social media platforms should have the legal duty to protect their users from the weaponization of their platforms, alongside disinformation and harmful content — or in the case of Myanmar, military-driven information operations and amplified narratives of hate. Social media companies should be legally obligated to conduct due diligence and institute safeguards — beyond effective content moderation algorithms — before operating anywhere, akin to car manufacturers installing and testing road safety features before putting a car on the market.

“It’s not that companies like Facebook intentionally want to cause harm,” Schissler said. “It’s just that they’re negligent.”

The way forward

What needs to change is both our awareness of how social media companies work and the law’s understanding of how platforms cause harm.

“Human rights due diligence as it is currently practiced focuses narrowly on discrete harms,” said André Dao, a postdoctoral research fellow studying global corporations and international law at Melbourne Law School. He said internationally recognized frameworks designed to prevent and remedy human rights abuses committed in business operations only address direct harms and overlook indirect but equally dire threats.

In a Business for Social Responsibility (BSR) report that Meta commissioned in 2018

about its operations in Myanmar, BSR — a corporate consultancy — narrowly attributed human rights abuses to Meta's limited control over bad actors and

> Myanmar's allegedly low rate of digital literacy. The report recommended better content moderation systems, neglecting a core catalyst of the genocide: Facebook's recommendation algorithms.

Giving users more agency, as Brún notes in the Amnesty report, is also critical in minimizing the effects of personalized echo chambers. He advocates for more stringent data privacy practices, proposing a model where users can choose whether to let companies collect their data and whether the collected data is fed into a recommendation algorithm that curates their newsfeeds. To Brún, the bottom line is effective government regulation: "We cannot leave companies to their own devices. There needs to be oversight on how these platforms work."

Between fueling Russia's propaganda campaigns and amplifying extremist narratives in the Israel-Hamas war, the current lack of social media regulation rewards harmful and exploitative business practices. It leaves victims no clear paths for accountability or remediation.

Since the Rohingya genocide began in 2017, much of the internet has changed: Hyperrealistic deepfakes proliferate, and the internet has started sharing much of its real estate with content generated by artificial intelligence. Technology is developing in ways that make verifying information more difficult, even as social media companies are doubling down on the same engagement-maximizing algorithms and targeting mechanisms that played a role in the genocide in Myanmar.

Then, of course, there's the concern about censorship. As Vox has previously reported in the past, changes to Section 230 might engender an overcorrection: the censorship of millions of social media users who aren't engaging in hate speech. "The likelihood that nine lawyers in black robes, none of whom have any particular expertise on tech policy, will find the solution to this vexing problem in vague statutes that were not written with the modern-day internet in mind is small, to say the least," wrote Vox's Ian Millhiser.

But to an optimistic few, programmable solutions that address the pitfalls of recommendation algorithms can make up for the shortfalls of legal solutions.

"If social media companies can design technology to detect copyright infringement, they can invest in content moderation," said Simon, referencing his research for Yale's program on mass atrocities in the digital era. He said these new technologies shouldn't be limited to removing hate speech, but should also be used in detecting potentially harmful social trends and narratives.

ExTrac, an intelligence organization using AI to detect and map emerging risks online, and Jigsaw, a Google incubator specialized in countering online violent extremism, are among the many initiatives exploring programmable solutions to limit algorithmic polarization.

"Tech isn't our savior, law isn't our savior, we're probably not our own saviors either," Simon said. "But some combination of all three is required to inch toward a healthier and safer internet."