moz://a

When content moderation hurts

BY MOZILLA INSIGHTS | MAY 4, 2020

Six stories to inspire better regulation.

Numbers alone do little to illustrate the human impact of a bad content decision on the world's biggest internet platforms. Whether it's content that should be taken down, or content that was unjustly removed, seeking a clear explanation or reversal can be endlessly frustrating.

In this article, we share six brief stories that highlight examples of key challenges and opportunities to improve platform regulation: the limitations of automation and filtering, the gaps in transparency and consistency of rules, and the chance for engagement with an ecosystem of people and groups exploring thoughtful social, technical and legal alternatives.

Public pressure to reduce online hate speech, disinformation and illegal content is mounting along with calls to introduce more regulation to hold platforms accountable. Yet very often, attempts by lawmakers to respond to these real grievances cause more harm than good, and fail to address the root problems. At the same time, some problems remain acute and underappreciated in the policymaking process, including the impact of content moderation on the physical and mental health of human moderators.

By sharing these stories we wish to inspire inclusive policymaking that is grounded in evidence and avoids mistakes of the past. Too often, laws incentivize blunt enforcement as a hasty measure in direct response to scandals and conflicts (or pandemics!) rather than as a sustained and transparent process towards a healthier internet for all. The complexities underscored by these stories show that effective regulation will not be easily achieved, and certainly not everywhere at once, but also how important it is to work with allies to continue to do better.

Index

Automation and filtering

Content-focused regulation often privileges automation and filtering as a universal remedy for content moderation on large platforms. Unfortunately, the technology is still underdeveloped, and often based on biased or incomplete data sets. Far too often, it leads to situations where ‘bad’ content remains online and ‘good’ content is taken down. Today, artificial intelligence is a prominent component of content moderation at scale, but it is a blunt tool when it comes to moderating public interest expression that must be understood within a particular human context. As such, policies that incentivize or mandate its use can give rise to real harm.

Story #1

Misinfodemics and COVID-19

When content moderators of major platforms including Facebook and YouTube were sent home at the beginning of the global pandemic, many people reported high frequencies of automated takedowns of posts about COVID-19. “Users and creators may see increased video removals, including some videos that may not violate policies,” said YouTube’s Creator Blog on March 16, explaining why machine learning systems would begin removing content without human review. Meanwhile, viral disinformation about the coronavirus is still spreading worldwide. Platforms are engaged in a constant scurry to remove, deemphasize or demonetize falsehoods in spam, ads, videos and links. But ‘misinfodemics’ are not a new phenomenon, says Nat Gyenes of Meedan’s Digital Health Lab. “The major platforms need to develop stronger global public health response teams to implement effective content moderation strategies,” she says. Even for protracted issues like cancer misinformation, Gyenes says platform engagement with public health experts and researchers from around the world is insufficient. “Pandemics are global but they are experienced locally, and that’s the challenge with content moderation, too,” she says. People everywhere ask health questions in emails, online forums and message groups informed by local vocabulary and context. Human reviewers are essential, she says. In rapid response to the COVID-19 pandemic, Meedan is developing an expert health source database that can be used by the international fact checking community. It will contain expert-verified health information from around the world, including India, Brazil, Kenya and Senegal. It will also loop information back to health workers and eventually to platforms, too. Ultimately, automated solutions for content moderation have limits and this is a hard truth that has been laid bare by the COVID-19 pandemic. Moderation problems are complex, and there are no silver bullets.

Further reading

- Health Equity Through Health Fact-checking: A Primer, Meedan Digital Health Lab, Nat Gyenes & Megan Marrelli, 2019
- How Coronavirus Disinformation Gets Past Social Media Moderators, Bellingcat, Robert Evans, April 2020
- Coronavirus: How are the social media platforms responding to the ‘infodemic’?, Clea Skopeliti, Bethan John, First Draft, March 19, 2020

Story #2

Social media as evidence and memory

Millions of videos of human rights violations can vanish in the blink of an eye if an automated filter deems it to be 'terrorist' content. Syrian Archive is an organization that collects, verifies and preserves visual evidence to document human rights violations. For years, they have urged platforms to preserve video evidence and make it accessible to investigators, even if they chose to hide it from public view. They regret that social media platforms have become "accidental archives" and aim to offer an alternative archival solution themselves. "It's frustrating," says lead researcher, Jeff Deutch. "In many cases, this type of documentation might offer the only evidence that a certain crime has been committed. When it's removed, it really makes it hard for journalists or investigators to report what's been going on, or for prosecutors to pursue justice," he says, adding that the appeals process to reinstate content is not sustainable, especially when suspended accounts may belong to people who are detained or have been killed. "To give one example, when we approached YouTube about reinstating about 300 videos documenting attacks on hospitals and medical facilities, it took about six months to get the content reinstated," he says. Unfortunately, many countries and regions have enacted problematic counter-terrorism laws that result in takedowns of human rights content. For instance, the European Union is drafting new regulations on terrorist content online with strict takedown timelines and obligations to filter content that activists fear will mean platforms take down even more content that needs to remain available to the public.

Further reading

- Caught in the Net: The impact of "extremist" speech regulations on human rights content Syrian Archive, WITNESS, EFF (May 2019)
- 'Envision a new war': the Syrian Archive, corporate censorship and the struggle to preserve public history online, Global Voices, Ellery Biddle, 2019
- YouTube admits 'wrong call' over deletion of Syrian war crime videos, Middle East Eye, Alex MacDonald, 2017

Public spaces, private rules

Large platforms face pressure from governments around the world to identify and act on harmful and illegal content. But even when laws are on the books, decisions about what stays up and what comes down tend to rest solely with the platforms. In other words, in practice, a platform's terms of service and content moderation operate like law, but without the safeguards of oversight, due process, or accountability that we expect from laws. This can result in important public interest expression being removed without explanation or a meaningful chance to appeal. It also means that platforms are incentivized to prioritize content moderation efforts in regions where governments are most empowered to pressure them, meaning that people and groups in less commercially-important regions are are often neglected in trust and safety efforts.

Story #3

Politics of online attention

The difference in care and attention to users of different countries is something digital researcher Rosemary Ajayi discovered firsthand during the 2019 election in Nigeria, as she and a team of eight monitors collected evidence of hundreds of tweets spreading disinformation, hate speech and impersonating candidates. Not only was it initially difficult to reach anyone from Twitter, she says, but she realized that certain content reporting tools to protect election integrity in the United Kingdom and elsewhere were not available in Nigeria. Even seeking a verified checkmark for the Twitter accounts of Nigerian political candidates proved impossible. “There is really no justification for why existing tools that could protect users in Nigeria aren’t made available. After all, I’m more afraid of violence happening during elections in Nigeria than in the UK.” Ajayi systematically flagged and reported harmful content, documenting how long it took to receive replies. “Some acknowledgements would come after an hour, some after 24 hours, and sometimes it would be even longer. If you report something serious on election day, and they get back to you a week later, what’s the point?” she asks. Her team, under the banner of the Digital Africa Research Lab continued collecting thousands of tweets through 2019. Regulation is necessary, she says, but she worries about a new social media bill in Nigeria (modeled after a Singapore law) that would introduce severe penalties for infractions. “There is a lot of harm on social media but they should be tackling the platforms with regulation not individuals,” says Ajayi. Together with researchers from other countries she has informally volunteered advice to social media companies on local concerns. “The platforms don’t necessarily understand what the issues are, but then even when they do, I don’t know that they have the will to actually address them,” she says.

Further reading

- Nigerian Twitter has an impersonation problem — and the platform is failing to take action, Global Voices, Rosemary Ajayi, September 2018
- Facebook India - Towards a Tipping Point of ViolenceCaste and Religious Hate Speech, Equality Labs, Soundararajan, T., Kumar, A., Nair, P., Greely, J, 2019
- Facebook Says It’s Removing More Hate Speech Than Ever Before. But There’s a Catch, Time Magazine, Billy Perrigo, October 2019

Story #4

Drug harm reduction and education

Reversing a content moderation decision can be especially frustrating when there is no human response or even a semblance of due process. In Poland, an organization named SIN that educates youth about dangerously high potency levels of recreational drugs, say they tried everything to seek an explanation from Facebook when their pages and group of 4,000 members were banned for "violation of community guidelines" in 2018. "One day I woke up and it was all gone," says Jerzy Afanasjew of SIN. "We thought maybe it's an automated decision and clicked the button to seek reevaluation. We tried reaching out to personal connections. We even printed petition signatures and sent them to Facebook in Warsaw, Dublin and California." After a year of waiting for a reversal, SIN launched a court case against Facebook in May 2019 represented by the Polish digital rights group Panoptykon Foundation. They are demanding a public apology and that SIN's pages and accounts be restored. Drug harm reduction groups are often impacted by content moderation systems designed to curb sales of illegal substances, but they are sometimes also reinstated. "I personally know groups in the United States that have had similar issues but were able to resolve them. It's like we're second rate users to them unless we spend lots of money on ads," says Afanasjew. "I feel like five years of my work went down the drain."

Further reading

- SIN vs Facebook, Panoptykon Foundation, 2020
- Facebook Is Censoring Posts That Could Save Opioid Users' Lives, Maia Szalavitz, Vice, July, 2019.

Innovating for future alternatives

Complementary to effective regulation, there are countless examples of community initiatives to fight back against harmful side effects of content moderation practices or alternately of insufficient action, including through software development and direct engagement with platforms. Such efforts are crucial to exploring alternatives to the status quo and need to be encouraged and protected. Lawmakers should learn from successful approaches and seek to engage with the wider ecosystem of people and organizations who are researching thoughtful solutions to the root causes.

Story #5

Coding feminist alternatives

Women and nonbinary people bear the brunt of online harassment, but misogyny is rarely called out as a priority for content moderation in the way that hate speech is generally. What possibilities exist for users to control their own experience? Teresa Ingram worked as the only female software developer at a bank in the United Kingdom when she learned about the online abuse faced by female politicians from her girlfriend who worked for the Danish parliament. “I just thought to myself, if I don’t do something no one else will,” she says, describing her idea to create a browser plugin called Opt Out that hides misogynistic tweets from an individual’s social media feed using a machine learning model. Opt Out is now one project of the tech activism group Opt Out Tools which has seen volunteers from different disciplines rally around the experimental and outspokenly feminist software since the start. “The problem we’re trying to solve isn’t even really recognized by the platforms,” says Ingram, noting that reporting offensive tweets through regular channels doesn’t spare women the discomfort of viewing abusive messages in the first place. Their goal is to develop the browser extension so that the individual can give it feedback and tailor it to their online personal experience. “We've taken a ‘Big Sister’ instead of a ‘Big Brother’ approach to content moderation. The user gets to decide what they do and don't want to see. There is never going to be a successful model or moderation system that is going to be perfect for every single person in the world,” she says, “so we want to help people make their own.”

Further reading

- Opt Out Tools
- Toxic Twitter – A Toxic Place for Women, Amnesty International, 2018
- Online Abuse 101, Women’s Media Center Speech Project

Story #6

Coordinating civic principles for user rights

It is no accident that transparency reports of major platforms have evolved over the years to include more of the information that internet users, organizations and governments are calling for. One influential community effort to create a baseline of standards is the Santa Clara Principles, developed in 2018 on the sidelines of a major conference in California about content moderation at scale. Over two years, numerous platforms have publicly endorsed the principles, and promised varying degrees of due process and opportunities for appeal. “We recognize the need to take a global approach to protect users from vague and unfair content moderation practices,” says EFF’s director of international freedom of expression, Jillian York, while launching a global call for proposals (until June 2020) for how to revise and develop the standards with emphasis on groups representing marginalized voices that are heavily impacted. Organizations on the frontlines of content moderation and corporate transparency, including Ranking Digital Rights, also question how the targeted advertising business model incentivizes platforms to amplify harmful content in ways that content moderation alone can never resolve. “We think it’s time for regulators to hold digital platforms accountable for how content is amplified and targeted to users rather than for the content that their users post,” says Nathalie Maréchal. Both in collaboration with, and sometimes in opposition to platforms, advocates are eager to engage with lawmakers on how to approach platform regulation thoughtfully with respect to human rights and free speech, building on years of work with affected individuals and groups.

Further reading

- The Santa Clara Principles: Call for proposals
- Who Has Your Back? Censorship Edition 2019, Andrew Crocker, Gennie Gebhart, Aaron Mackey, Kurt Opsahl, Hayley Tsukayama, Jamie Lee Williams, and Jillian C. York, EFF, 2019
- It's Not Just the Content, It's the Business Model: Democracy’s Online Speech Challenge, Nathalie Maréchal & Ellery Roberts Biddle, Ranking Digital Rights, March 2020

What can be done?

We often hear the Web referred to as an unregulated 'Wild West' where anything goes. The stories above illustrate that this is far from the reality. Around the clock, decisions are made about online content that have real and significant impact on the human rights and lived experience of citizens everywhere.

While much content moderation is mediated by the terms of service of a platform, the impact of government regulation and pressure on how these decisions are taken should not be underestimated. Today, too much regulation incentivizes or mandates the at-scale deployment of flawed content filtering technology, and too little focus is placed on ensuring transparency, accountability and due process in how platforms act on those government mandates.

We hope these stories have illuminated the human cost of that reality, and highlighted what is at stake when policymakers craft content regulation frameworks. At the same time, while we acknowledge how difficult it will be to strike a balance that works for people around the world, we hope they have also provided a marker for what is possible if regulation is done right.

Primary contributors to this article include: Solana Larsen, Owen Bennett, Brandi Geurkink, Eeva Moore, Stefan Baack and Kasia Odrozek. Thanks to all reviewers! Design: Kristina Shu. Illustrations: Xenia Latii.