Why moderating content without censoring users demands consistent, transparent policies

besedo

Free speech and censorship attract a diverse range of views, depending on who you talk to. Across governments, businesses, and citizens, we find stances on the issue which are vastly different and often contradict one another. For online platforms – social media sites, dating apps and marketplaces – this is a difficult issue to grapple with. Having been handed the keys to decide the definition of harmful and abusive content, online communities are left navigating the requirement to defend and protect their users from hate speech, inappropriate sexual content and fake users. However, to do this they are using their own tools to decipher exactly how to classify and take down harmful content.

This has, at times, drawn negative press. Facebook, for example, came under fire for removing a Pulitzer Prize-winning photo titled 'Napalm Girl'. The photo, strikingly depicting the horror of war, hit their moderation filters for contravening Facebook's strict no nudity policy. That policy is so strict, in fact, that the company has also been criticised for removing photos of breastfeeding women. Artists and mothers alike have been in uproar about their rights being restricted and clearly valuable content being labelled as harmful. Such offence is not as rare as it ideally should be.

Images are not the only controversial content flagged as censorship. Twitter's ban stopping users from sharing a New York Post exposé article on some of Hunter Biden's supposed emails led to commentators decrying the 'end of free speech' in the run up to the US Presidential election. Such was the uproar that the press articles which circulated on the topic saw much more readership reach than Twitter users would ever have generated by sharing the content in the first place.

Dating apps too have had to adjust to the times. Attempting to tackle the fact that its 62% of its users reported that they are likely to receive unsolicited comments about their appearance online, Bumble officially banned body shaming on the platform. According to ReNew Houston, it updated its terms and conditions to 'explicitly ban unsolicited and derogatory comments made about someone's appearance, body shape, size or health'. The announcement created heated debate in the Twittersphere, with some saying that the policy violates freedom of speech.

In the US, Amazon too has been accused of limiting free speech because it halted the sale of a book seen as attacking transgender people, with the Independent reporting that this decision outraged some prominent conservatives.

As social media continues to expand its scope to include commerce, dating, gaming, and other online community enablement, these issues will become more important to every online industry. In this eBook, we will explore what lies at the heart of the tension between moderation and censorship – and how we can resolve it.

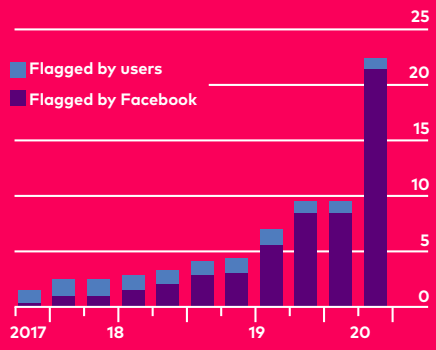
“As social media continues to expand its scope to include commerce, dating, gaming, and other online community enablement, these issues will become more important to every online industry.

Content moderation is not going away

Removal of content from community platforms is growing – and growing more quickly than user numbers, demonstrating that big tech is more determined than ever to crack down on what it, and the law, deems inappropriate. For example, an Economist analysis of Facebook data demonstrated that its removal of hate speech has risen tenfold in two years. It now disables some 17m fake accounts every single day – more than twice as many as three years ago.

Yet, Mark Zuckerberg has defended Facebook as a bastion of free expression, reports the Guardian. He denies charges of censorship levied against the platform, going as far as to defend the platform's decision to allow misinformation in political ads, given the sensitive nature of the content. Zuckerberg said he was against the banning of ads altogether because that would favour political incumbents and whoever the media chooses to cover.

Growing moderation
Facebook, hate-speech content acted upon, millions



Source: The Economist

So, is this censorship?

It's clear that online communities who set out to be neutral platforms are struggling to standardise, support, and justify their content moderation policies and choices. But, when it acts in line with the law and upcoming regulations, is content moderation censorship? In the strictest sense of the definition the answer would be yes. A censor, according to Merriam-Webster, as referenced in our blog, is 'a person who examines books, movies, letters, etc., and removes things that are considered to be offensive, immoral, harmful to society, etc.'.

But does content moderation have to feel like censorship when it's used in everyday life? When applied correctly, using the right processes and technology, does it necessarily impinge on a community's free speech?

We need content moderation

Yes, there are examples where online communities have had to rethink their moderation, and it is true that platforms can go beyond what that the law requires to enforce their own policy on what content is safe for users. Yet, imagine a world without content moderation. There are very few who would agree that we should allow our children to be exposed to sexual content or hateful slurs.

And the wrong type of content can not only be upsetting, but dangerous too. This is clear when we look to other areas where content moderation is used online: few would disagree with the idea of keeping users on dating apps safe by blocking threatening messages, or keeping shoppers safe by eliminating fraudulent listings on marketplaces. On social media, misinformation risks lives. Conspiracy theories about the global pandemic and vaccine availability, for instance, have abounded online, and if left unchecked this misleading information would lead to confusion, scams, and deaths.

> **There are very few who would agree that we should allow our children to be exposed to sexual content or hateful slurs.**

Platforms are curators

A different way of approaching the question, as outlined in one of our recent blogs, is to see how online communities actually have a lot in common with the content in art galleries and museums. The items and artworks in these public spaces are not created by the museum owners themselves – they're curated for the viewing public and given contextualising information. In enabling the sharing of content, they also have a responsibility. They need to make sure that what is being shown does not violate their values as an organisation and community. And, like online platforms, art curators should have the right to take down material deemed to be objectionable.

For galleries and museums, this is often difficult and sensitive work, partly because the collections they house are unique items that must be available to the world, so the interests of their users must be balanced with any judgments around objectionability. For the largest social networks, this problem is magnified: network effects mean that it is difficult (although not impossible) for people to vote with their feet by going to another platform. Moderating content, then, is a power that must be exercised responsibly.

These actions have invariably impacted individual users because that's the intent: to mitigate content which breaks the platform's community standards. In fact, removing hateful and harmful speech can make it significantly safer for other people to discuss their opinions.

The content moderation being enacted by platforms based on their established community standards typically involves:

- **Blocking harmful or hate-related content**
- **Fact-checking**
- **Labelling content correctly**
- **Removing potentially damaging disinformation**
- **Demonetising pages by removing paid ads and content**

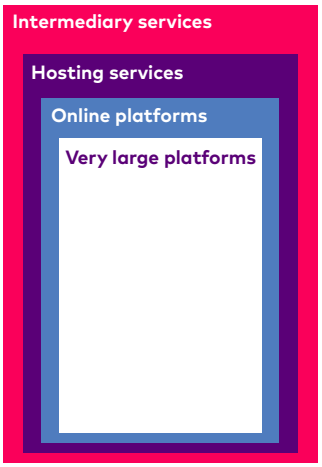
Recently, Twitter has gone one step further than its legal obligations and attempted to remove misinformation on vaccines through a combination of AI and human moderators to determine whether flagged tweets should be labelled as questionable or removed entirely. Repeat violators can expect to have their accounts suspended or deleted.

In fact, removing hateful and harmful speech can make it significantly safer for other people to discuss their opinions.

Content moderation and the regulatory environment

Content moderation involves discretion at an organisational level – not a personal one. Direction is also set by Governments and the law. While legislation varies by country, content moderation policy is often applied by platforms globally so that they can demonstrate consistency in their values.

In Europe, the Digital Services Act will introduce a new set of rules to create a safer digital environment across the EU. These will apply broadly to social media networks, but also e-commerce, dating platforms, and, in fact, all providers of online intermediary services to remove illegal content.



- **Intermediary services, such as ISPs**
- **Hosting services, such as cloud providers**
- **Online platforms, such as marketplaces**
- **Very large online platforms, with more than 45m EU users**

The definition of illegal content, however, is still under development: many propose that this will relate not only to hate speech but also content that is fraudulent, which offers counterfeit goods, or even content that seeks to mislead consumers, such as misinformation. This means that platforms may become directly liable if they do not correct the wrongdoings of third-party traders such as those in their marketplaces.

Last year, the European Court of Justice ruled that European countries could order Facebook to remove content worldwide, not just for users within their borders. The European Audiovisual Media Services Directive, meanwhile, requires online video services to take 'appropriate measures' to protect viewers from harmful or illegal content, including setting up age checks.

Ultimately, these laws could see platforms go further and faster with their content moderation than ever before.

Context aware: AI and human moderation

A major difficulty for many providers is tackling the sheer volume of content to ensure that content moderation policy is applied consistently. Consistency is key: without it platforms can face charges of bias and unfairness, which quickly becomes an experience of censorship. Artificial intelligence has enabled platforms to remove offending content in line with set content guidelines. A good AI model or solution that is built around the platform's specific requirements should capture and automatically refuse the bulk of harmful content.

Some content, such as nudity or pornographic content, is easily identified by machines. AI, however, is not a panacea for content moderation – it's just the starting line. Machines can find it hard to moderate content where the subtext or context is unclear, especially if the content relates to the fast-paced news agenda where it is tricky to make the distinction between malice and opinion.

Consider a comment under a news article about a new videogame, for instance, featuring the phrase 'I'll kill you'; is this a threat, or someone looking forward to playing the game? A real-world example is Facebook's removal of the US Declaration of Independence for violating 'hate speech' standards. Uploaded in small chunks by a local newspaper, AI flagged and took down the phrase 'Indian Savages' that violated their standards.

Some of this can be managed by helping AI to accurately read long threaded conversations rather than singular posts. However, training AI takes time, and filters will remain an important tool to quickly respond to emerging issues, and some posts will always need a trained human moderator to review the content that AI cannot classify.

> **...Some posts will always need a trained human moderator to review the content that AI cannot classify.**

How can platforms get it right?

In order to help AI and human moderators consistently apply content moderation policies, platforms need to establish community guidelines for their sites and explain that users' expression must be within those guidelines. This is about establishing trust with users. In fact, according to a survey by Gallup and the Knight Foundation, 84% of Americans don't trust social media companies to decide what content they should allow on their respective platforms – but they still trust them more than they do the government.

Transparency, communication, consistency, and adaptability are foundational to getting this right.

Transparency

User education is vital to empowering choice. Platforms are like curators with a duty of care for the community, and the internal processes behind this should be transparent so that users can decide whether the actions taken on their behalf match their own values. Ultimately, platforms don't want to lose users, and the suspicion of bad faith on the platform's part is more damaging than clearly stating why some forms of content are not welcome.

Communication

Listening to, and collecting data from, users on their own values to decide what should and shouldn't be allowed on the platform is key. Decision makers should be open to gathering community feedback to inform content moderation policies. This helps create community buy-in on tough decisions. A data-driven approach will alleviate the tension between an earnest ambition for fair policymaking and the common scenario where the CEO is really the ultimate arbiter on high-profile issues.

Consistency

Twitter has pledged to tackle misinformation, alerting users when information in tweets from figures such as Donald Trump is contested. It is also trying to slow the spread of vaccine misinformation through similar methods. To succeed with this type of mission, platforms need to apply measures consistently wherever misinformation is spreading, not on a case-by-case basis. This helps users to appreciate a clear direction for moderation and steers opinion away from accusations of bias.

Adaptability

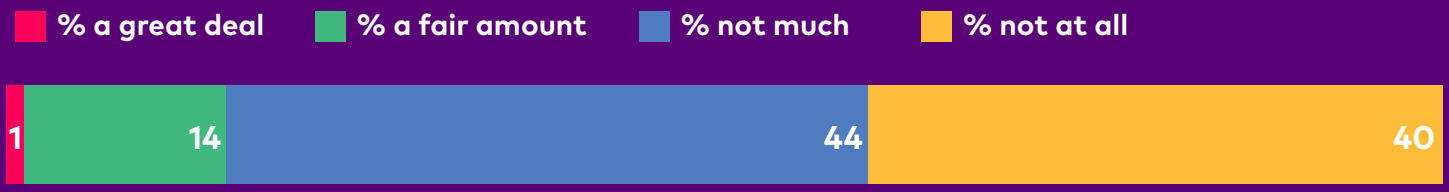
Changing regulations such as the Digital Services Act are creating a framework online harms reduction. Preparing now to ensure that the platform, people, and processes will be able to comply is crucial. Cross-industry collaboration can also help – where one platform is struggling to apply legislation, others will have the same challenge. Coming together to share advice and debate some of the trickier areas will help the industry to lay the foundation for nimble, effective content moderation.

Track technology, its evolution, and how it can help: AI is evolving, and content moderation specialists are using machine learning to help systems understand data and augment that information to account for context-providing signals. For example, AI can now consider a user's background and what they are responding to in order to make a more nuanced judgement. These developments are exciting, and platforms should keep pace with them in order to deploy AI for maximum impact.

A big future for content moderation

Levels of trust in social media companies content decisions

how much do you trust social media companies to make the right decisons about what people can say or poston their websites and apps



note: due to rounding, percentages may total 100% +/-1%
Knight Foundation/Gallup Dec. 3-15, 2019

Content moderation will continue to rise in importance. Gartner predicts that, by 2024, 30% of large organisations will identify content moderation services for user-generated content as a C-suite priority. For large platforms, this is already topping the list of priorities. For smaller platforms, content moderation best practices should also remain high on the agenda: it is better to create a strong foundation from the start than to retrofit policies through learning from costly hindsight.

In the longer term, we'll see platforms working more with external providers and independent specialists, opening up their policies for scrutiny. It was announced at the start of this year that the Facebook Oversight Board, established to independently review and rule on moderation decisions, overruled the platform in four out of five key cases, finding that it was wrong to ban nipples in breast cancer posts and that it had taken down posts which were, according to the Board, incorrectly flagged as hate speech. A movement towards such an approach, as part of a mission to listen and adapt, will likely continue as a way of building greater trust and transparency.

In the dating world, Bumble has thought about making its site more accessible to its users by ensuring that the majority of the company's 11 board members are women, which is perhaps why it has led the charge on listening to its users about how to deal with unsolicited body shaming content. This is particularly in line with the platform's purpose and the values, which is to create a more comfortable and empowering experience for women by, in opposite-gender matches, allowing only female users to make the first contact. This is good moderation because the platform is acting in coherence with its values and policies and being transparent about what those are.

Increasingly, we will see that platforms start to better listen to the content nuances of people who come from all genders, identities, race and backgrounds, so that all in the community feel welcome and understood.

Adaptive policy responds to and encourages change

So, where does this leave us on the question of moderation and censorship? For museums and galleries, as for marketplaces and dating apps, content moderation is applied but accusations of censorship are rare. To achieve that kind of relationship with society, social platforms need to build trust. Other platforms also need to work to maintain the trust they have generated: Bumble, for instance, recently had to back-track on a decision to remove political alignment from profiles.

Whichever yet-untrodden content moderation path is taken, adaptive policies must be supported by a set of clear guiding principles. Platforms will need to continue to provide transparency around guidelines and policies to inform users and also to help promote good behaviour. Either opaque policies which users don't understand or slow responses which only catch some content will only feed the feeling that content moderation is biased, and therefore counts as censorship.

Content moderation which uses AI, filters, and manual moderation will be needed to execute on guidelines providing both content moderation accuracy and quality. AI should also be able to help platforms scale content moderation to manage the volume of user-generated content and ensure users can have a better, safer experience.

However, this execution will need to be easily adapted, and this can be done by using manual and filter solutions to respond, quickly, to new trends and content formats which content breaks policies. These can do so in a way that AI often can't because, currently, machines still need weeks to be trained to understand deeply nuanced context. Often this hybrid model of human and machine content moderation empowers platforms with the flexibility to put resources and concentrate efforts where they're needed.

This should go hand-in-hand with platforms having their finger on the pulse of the next AI evolution: helping machines to better learn context so that content moderation is better applied. With AI becoming more powerful every day, in the future it will take an ever-larger stake in getting content moderation right. Done right, it's a route to content moderation which users understand and trust to keep them safe.

“Either opaque policies which users don't understand or slow responses which only catch some content will only feed the feeling that content moderation is biased, and therefore counts as censorship.

besedo.com

hi@besedo.com

besedo