WHITEPAPER

THE GROWING NEED FOR CONTENT MODERATION

An essential guide for Boards and Management

probecx
NEXT GENERATION DRIVEN

TABLE OF CONTENTS

INTRODUCTION

The past two decades have seen an explosion in online platforms that allow users to upload content for others to view. From its earliest days, the internet was hailed as a revolutionary way for people to share information without relying on traditional gatekeepers but technical skills were still largely needed to do so. That is no longer the case.

Platforms such as YouTube, Facebook and Instagram have created a world where sharing content with a global audience is as simple as clicking a mouse, while comments sections and message boards thrive on people unleashing their opinions, feedback and, as is often the case, criticisms. User-generated content is one of the internet's greatest selling points. Increasingly, it is also cause for concern as users, businesses and policy makers recognise the dangers that come with technology that allows people to post what they want when they want.

At its most extreme, the internet is a hotbed of disturbing content ranging from child pornography, and recruitment for terrorist groups to hate speech, illegal drug sales and persecution of minority groups. Governments and law enforcement agencies are waging a war to identify and eradicate such material, along with prosecuting the people and groups who thrive on its existence.

Then there is the user-generated content that strikes fear into the heart of businesses.

The ability for people to upload material or remarks to corporate platforms is a blessing and a curse. While companies have never been able to more easily engage with the public, the flipside is it has never been easier for consumers to publicly criticise, rebuke and even shame companies if dissatisfied with their efforts. There is also the constant threat of users uploading inappropriate content such as misogynistic, racist and homophobic material on business platforms and the reputational harm that can be caused if not addressed or removed quickly enough.

And so we come to the two most important words businesses need to consider when operating in the online space – Content Moderation.

WHAT IS CONTENT MODERATION?

Social Media Today defines Content Moderation as "the practice of monitoring submissions and applying a set of rules that define what is acceptable and what is not."[1] At first glance, it seems the most simple of concepts – set a standard, monitor what users upload and remove anything that is not acceptable.

If only.

The reality is the sheer volume of content on the internet means it is impossible to monitor every piece of user-generated content.

To put it in perspective, the world wide web contains at least 5.75 billion pages[2], while 300 hours of video are uploaded to YouTube every minute and almost 5 billion videos watched on the platform every day[3].

5.75 billion
pages on the world wide web

300 hours
of video are uploaded to YouTube every minute

5 billion
videos watched on YouTube every day

Everest Group is tipping the Content Moderation market to reach up to US$6 billion by 2022[5].

Uploading content to the internet has become second nature for hundreds of millions of people across the world but the organisations that host that content are continuing to grapple with how best to monitor and moderate what it contains.

More than 30 million videos were removed from YouTube in 2019 for violating its community guidelines, while the same year Google Maps detected and removed more than 75 million policy-violating reviews and 4 million fake business profiles[4]. It is not only big-tech behemoths facing an uphill battle to stay on top of the issue though, with management consulting firm Everest Group tipping the Content Moderation market to reach up to US$6 billion by 2022[5].

Any organisation with an online presence – and there are very few that don't – needs to proactively address illegal or questionable content on its site. It is not good enough to turn the other cheek or hope one's consumers understand if 'a few bad eggs' upload inappropriate material. The risks are simply too high.

Fortunately, ready-made solutions are available for those willing to invest in them.

[1] GRIMES-VIORT, BLAISE. '6 Types of Content Moderation You Need to Know About.' Social Media Today. Social Media Today, 07 Dec. 2010. Web. 25 June 2016

[2] www.worldwidewebsize.com

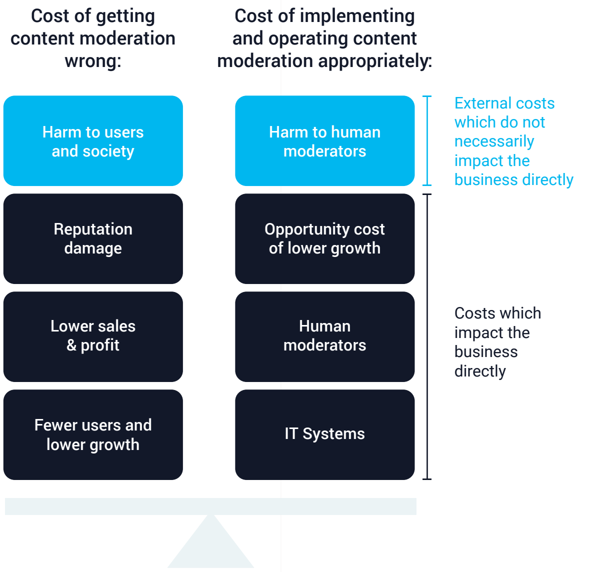
[3] DONCHEV, DANNY. '37 Mind Blowing YouTube Facts, Figures and Statistics – 2021'. FortuneLords. 07 February, 2021. Web.

[4] GOOGLE, YOUTUBE. 'Information Quality and Content Moderation'. Whitepaper.

[5] EVEREST GROUP. 'Content Moderation Thought Starter'. 2020

WHY CONTENT MODERATION IS VITAL

User-generated content is an overwhelmingly positive factor for organisations. Fostering an environment where consumers can engage with your website, share product reviews and upload content to social media platforms is an opportunity to build an ongoing connection and potentially see them act as advocates for one's services. Increased traffic has also been shown to attract more visitors and contribute to improved search engine rankings.



Cost of getting content moderation wrong:
Cost of implementing and operating content moderation appropriately:
Harm to users and society
Reputation damage
Lower sales & profit
Fewer users and lower growth
Harm to human moderators
Opportunity cost of lower growth
Human moderators
IT Systems
External costs which do not necessarily impact the business directly
Costs which impact the business directly

Organisations must balance competing costs to determine their approach to online content moderation (**SOURCE**: Cambridge Consultants)

The problem is user-generated content can also expose brands to risk if it is not moderated adequately. By implementing a strategic content moderation system, businesses can rest easy knowing they are giving themselves every chance of achieving three vital goals.

Protection

Every business will at some point receive user-generated content that violates guidelines. Being prepared for that moment is the key. Content Moderation drastically reduces the potential harm caused by offensive material and defamatory comments and helps reinforce an online environment where users can feel safe and free to contribute in a respectful and increasingly positive manner.

Understanding

Content Moderation is an excellent opportunity for businesses to gain insights into users' behaviours and opinions. This knowledge can be highly valuable in devising strategies and fostering communities that influence future decision-making, along with ensuring businesses can scale campaigns easily and effectively.

Growth

Just as people will rarely forget a negative experience on an organisation's website or social media, they are likely to return or leave favourable feedback when their visit has been a positive one. User-generated content plays a crucial role in attracting more traffic to a site, which flows through to increased customer engagement and improved search engine rankings. It all starts with a good experience and quality content moderation is essential to ensuring satisfaction among site visitors.

[1] GRIMES-VIORT, BLAISE. '6 Types of Content Moderation You Need to Know About.' Social Media Today. Social Media Today, 07 Dec. 2010. Web. 25 June 2016

[2] www.worldwidewebsize.com

[3] DONCHEV, DANNY. '37 Mind Blowing YouTube Facts, Figures and Statistics – 2021'. FortuneLords. 07 February, 2021. Web.

[5] EVEREST GROUP. 'Content Moderation Thought Starter'. 2020

TYPES OF CONTENT MODERATION

Content Moderation is a sure-fire way for businesses to ensure the user-generated content on their websites is a positive force for their brands. Expert moderators are able to quickly identify concerning material or posts and respond accordingly, with essentially five styles of moderation used:

Pre-Moderation

As the name suggests, this type of moderation captures any unfavourable content before it is uploaded to the site. All submitted content is pre-screened by content moderators, with the likes of product reviews, comments and user-generated content requiring a tick of approval before 'going live'. While there is no better safeguard against legal or reputational risk, there are clear disadvantages. It is a time-consuming and laborious process, while delaying the appearance of content or comments can prove frustrating for users and, in turn, reduce traffic numbers.

Post-Moderation

This approach results in all submitted content immediately going live but then being queued for review. Users get the boost of seeing their comments or material instantly appear online and, due to that immediacy, conversation on forums or social media sites can flourish. It is then the job of moderators to assess each item and delete any items that do not meet the site's guidelines. While suitable for sites with less traffic, post-moderation can be difficult when moderating large volumes of material and increases the risk of inappropriate content being missed.

Reactive Moderation

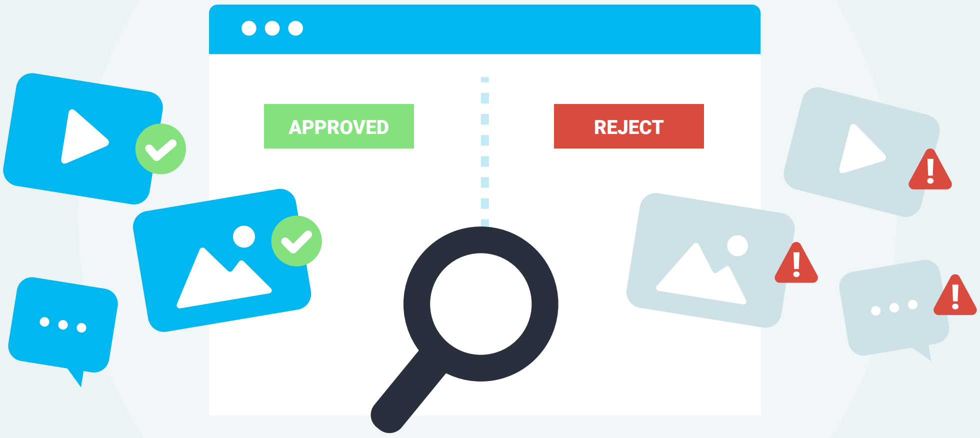
Rather than employing a team of content moderators, this approach asks users to 'flag' any offensive or questionable material for moderator review. One of the main benefits is its cost-effectiveness, with staff no longer required to review every piece of content uploaded to the site. It can also create a sense of site ownership among users, given they are invited to play a key role in fostering a safe online environment. That said, it relies on a proactive audience and the risk of inappropriate material going undetected is a significant concern, particularly for brand-conscious organisations.

Distributed Moderation

This form of moderation does not require specific moderators. Instead, a rating system is employed for the site's online community to score or 'vote' on published content, with material that does not meet a certain standard removed from the site. While this approach can inspire users to become a community, it is a risky strategy as it relies on the honesty of users who may have ulterior motives for targeting certain content and can allow harmful material to remain online far longer than it should. Distributed moderation should only be used by small organisations with a known and trustworthy user base.

Automated Moderation

Instantaneous, automatic and definitive – such are the benefits of employing automated moderation as a safeguard against inappropriate online material. Using digital tools that detect predetermined content, automated moderation can block user-generated content before it even goes live, along with IP addresses belonging to users known to be abusive. The cost benefit of not employing human moderators is obvious but their absence also means a lack of analysis and interpretation, which can lead to worthy content being rejected and vice versa.



THE HUMAN FACTOR

The need for Content Moderation is evidenced by the fact **25% of the internet's search results for the world's 20 largest brands are related to user-generated content**[6]. Up to 4 billion photos are posted on social media every day[7], with the likes of Facebook, Twitter, Instagram and YouTube all entrenched in business models that rely on users supplying the content that makes them among the world's most popular platforms.

But who are the people charged with moderating such content and, given the unsavoury nature of much of the content, what is being done to care for their wellbeing?

In a 2018 article in the Harvard Journal of Law & Technology, it was reported the job "is done by tens of thousands of online content moderators, mostly employed by subcontractors in India and the Philippines" and they "spend their days making split-second decisions on whether to take down questionable content, applying appropriateness criteria that are often ambiguous and culturally specific"[8].

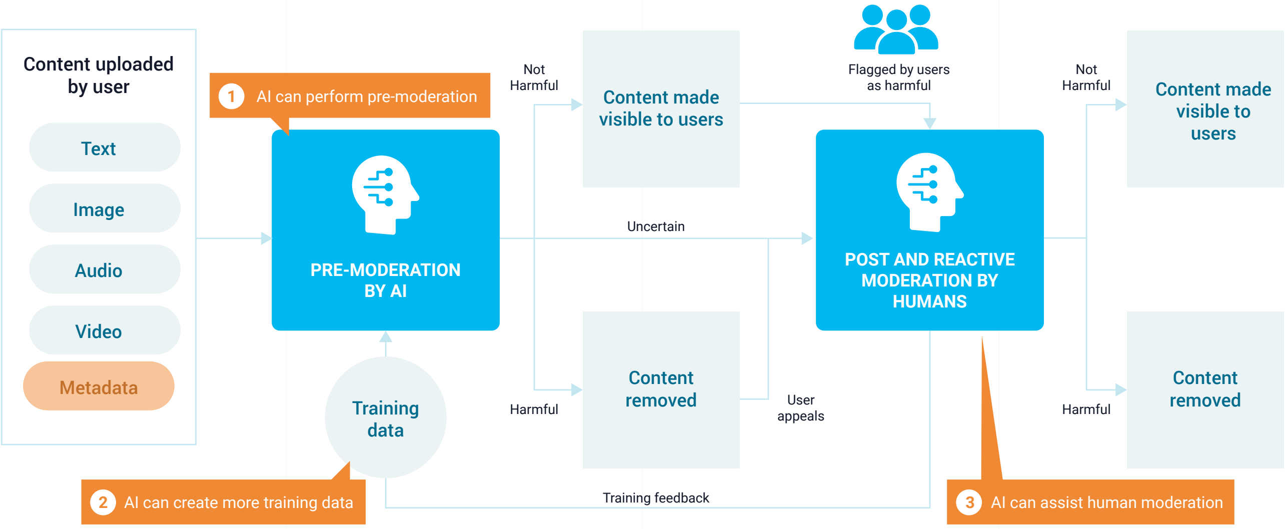
The article went on to highlight that filtering "the worst images and videos the internet has to offer" comes with inherent risks for moderators, with several subjects expressing their job had caused them fatigue, distress and, in one case, depression. As the need for Content Moderation services rise, companies are increasingly exploring steps to care for the people they employ to take care of their online communities.

This includes requiring that the people they hire clearly understand the nature of their roles and have access to learning initiatives or training programs that focus on the content they are likely to encounter on the job. Workplace counselling services are also essential to ensure content moderators are able to discuss and address their mental health and wellbeing as needed.

Given the growing demand for Content Moderation services, outsourcing companies are investing more resources in the sector. Higher pay rates, better rostering and access to on-site nurses and psychologists are just some of the initiatives that are being delivered to staff and playing a key role in promoting positive physical and mental health.

[6] DHAMDHERE, PRASAD. 'The Ultimate List of User Generated Content Statistics.' Social Annex Blog. Social Annex, 1 Sept. 2016. Web. 02 Nov. 2
[7] DHAMDHERE, PRASAD. 'The Ultimate List of User Generated Content Statistics.' Social Annex Blog. Social Annex, 1 Sept. 2016. Web. 02 Nov. 2
[8] ARSHT, ANDREW and ETCOVITCH, DANIEL. 'The Human Cost of Online Content Moderation'. Jolt Digest, 2 Mar. 2018.

THE DIGITAL FUTURE



Content uploaded by user
Text
Image
Audio
Video
Metadata
1 AI can perform pre-moderation
PRE-MODERATION BY AI
Training data
2 AI can create more training data
Not Harmful
Content made visible to users
Uncertain
Harmful
Content removed
User appeals
Training feedback
Flagged by users as harmful
POST AND REACTIVE MODERATION BY HUMANS
Not Harmful
Content made visible to users
Harmful
Content removed
3 AI can assist human moderation

There are three key ways in which AI can improve the effectiveness of the typical online content moderation workflow (**SOURCE**: Cambridge Consultants)

Just as it has created a world where user-generated content flourishes on the internet, technology has a key role to play in policing that material. Capabilities such as robotic process automation, AI-assisted decision support tools and AI-enabled task automation of review steps are helping social media companies protect their communities and scale their contact management operations[9].

While humans will always have a role to play in Content Moderation – particularly in terms of context and understanding – advancements in technology mean businesses are increasingly embracing it as a cost and time-effective means of filtering unacceptable online material.

Relatively simple AI techniques such as 'hash matching' (where a fingerprint of an image is compared with a database of known harmful images) and 'keyword filtering' (in which words that indicate potentially harmful content are used to flag content) are being used to improve the pre-moderation stage and flag content for review by humans. However, limitations such as detecting the nuances of language and scene understanding mean the human touch remains essential.

AI is also assisting human moderators by increasing their productivity and reducing the potentially harmful effects of content moderation. By prioritising content to be reviewed based on the level of harmfulness perceived in the content or the level of uncertainty, it can limit the type of harmful content they are exposed to. An AI technique known as 'visual question answering' is even allowing humans to ask the system questions about content to determine its degree of harmfulness without viewing it directly.

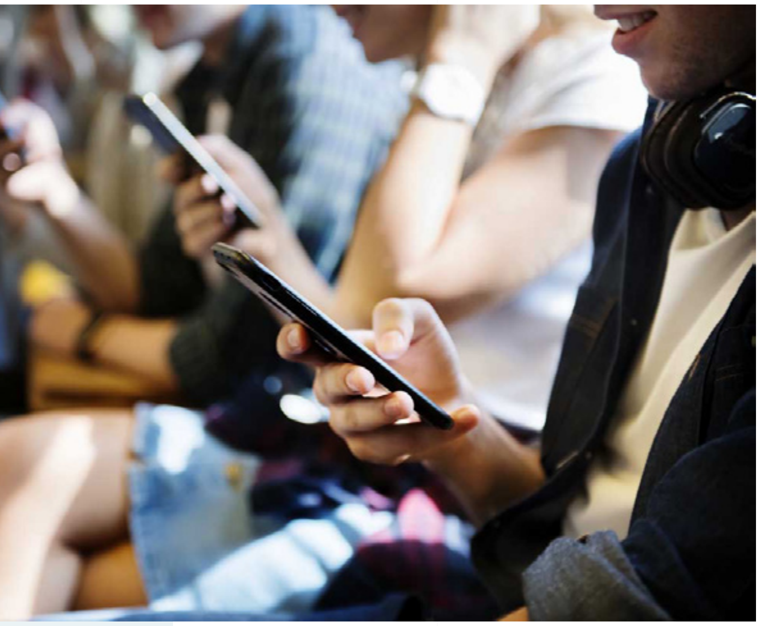
In exploring the "Use of AI in Online Content Moderation Online" in a 2019 report, Cambridge Consultants acknowledged that "AI is not a silver bullet"[10] That is indeed the case but it is also undeniable that any organisations not factoring it into their Content Moderation strategies are doing themselves a disservice.

[9] EVEREST GROUP. 'Four Key Trends in Social Media Content Moderation'. Blog. 27 May 2019.
[10] CAMBRIDGE CONSULTANTS. 'Use of AI in Online Content Moderation'. Report.

THE REGULATORY LANDSCAPE

Given the relative youth of the internet and the fact that user-generated content and online speech is a recent phenomenon, the development of a regulatory framework on Content Moderation is also in its infancy. While many questions are starting to be asked of big-tech and social media platforms about how they plan to control what is uploaded to their sites, many governments appear far from resolving such concerns.

The Australian Government has shown its commitment to tackling the issue via the development of an Online Safety Act, which is aimed at enhancing the powers of the nation's eSafety Commissioner, boosting protections for Australians and - of great significance in the Content Moderation space - increasing the responsibility on industry to keep their users safe online.



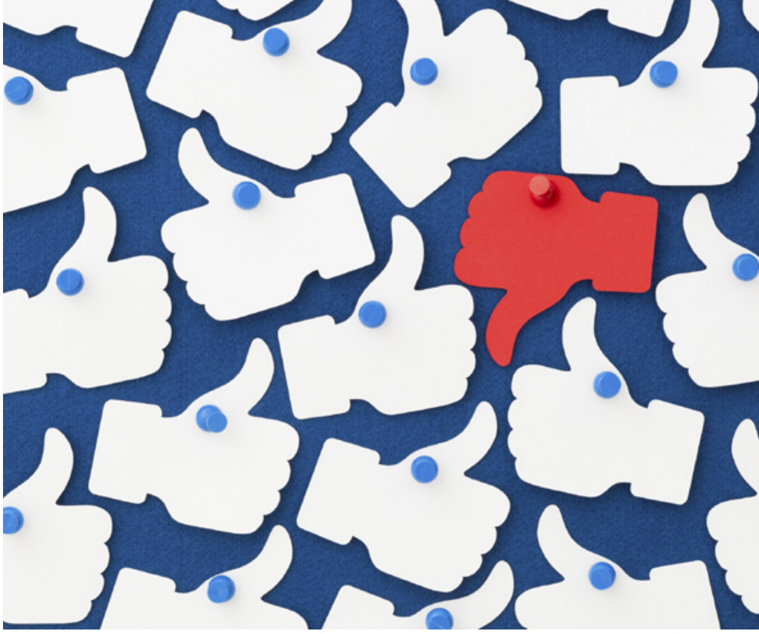
The act was introduced to parliament in February 2021 and includes legislation that will see a world-first cyber-abuse take down scheme that will empower the eSafety Commissioner to order the removal of seriously harmful online abuse. If a website or app systemically ignores take down notices for 'class 1 material', the eSafety Commissioner can require search engines and app stores to remove access to that service and these protections will be backed by civil penalties for service providers who fail to comply.

While the Australian Government is promoting its "world-first" approach, other nations are also working to introduce legislative safeguards regarding internet content. France passed its 'Fighting hate on the Internet' law in May 2020, which was "built in the image" of Germany's Network Enforcement Act Law, which was strengthened further by a law commanding social media platforms to not just remove violent hate speech but also report it to police[11]. Brazil passed its own bill fighting fake news in June 2020 - 'Brazilian Law of Freedom, Liability and Transparency on the Internet'.

[11] www.brookings.edu/blog/techtank/2020/09/21/the-push-for-content-moderation-legislation-around-the-world

In the United States, there have been several failed attempts at reforming legislation that protects social media platforms from facing lawsuits over what users post. In early 2021, the Congressional Research Service released a paper called ‘Social Media: Misinformation and Content Moderation Issues for Congress’ in which it declared: “Congress may wish to consider the roles of the public and private sector in addressing misinformation, including who defines what constitutes misinformation” before noting “if Congress determines that action to address the spread of misinformation through social media is necessary, its options may be limited by the reality that regulation, policies, or incentives to affect one category of information may affect others”[12]

Despite such hurdles, the day is coming when more regulation will be in play and it is essential organisations are ahead of the curve in addressing such concerns. Adhering to local jurisdictional policies and industry codes of practice is an obvious starting point. Google has highlighted that “many laws, from consumer protection to defamation to privacy, already govern content online” and “a smart legal framework for online platforms has been essential to enabling a reasonable approach to illegal content”[13].



At the very least, any Content Moderation strategy needs to be based on a company’s own guidelines for user-generated content, with a key emphasis on ensuring the content it hosts online does not negatively harm the organisation’s reputation or brand image.

SUMMARY

If anyone remains in doubt that Content Moderation is a growing concern for businesses, consider these numbers - by 2024, **30% of large organisations will identify content moderation services for user-generated content as a C-Suite priority**.

In a world where organisations are increasingly reliant on cultivating a positive online presence, the ability to welcome and promote user-generated content is a golden opportunity but can quickly result in tears if left unchecked. Content Moderation is the safety net all businesses need to ensure they are in the best position to decide what should and should not appear on their websites and platforms. Between the legal and reputational risks, it should not be left to chance.

[12] www.crsreports.congress.gov/product/pdf/R/R46662
[13] GOOGLE, YOUTUBE. ‘Information Quality and Content Moderation’. Whitepaper.
[14] GARTNER. ‘Gartner Unveils Top Marketing Predictions for 2021 and Beyond’. Media release.

ABOUT PROBE

Probe is a globally recognised and award-winning customer experience organisation that designs and deploys solutions to bolster and optimise our client operations. Founded more than 40 years ago and with 13,000-plus staff across five countries, the company delivers exceptional customer experiences through its deep knowledge and capabilities in Contact Centre and Customer Management, Digital Consulting, Intelligent Automation and Analytics.

www.probegroup.com.au

© 2021 The Probe Group. All Rights Reserved.

probecx
NEXT GENERATION DRIVEN