MIT Technology Review
Insights

Produced in association with

Teleperformance
each interaction matters

Modern content moderation scales on AI, but succeeds with human judgment.

Humans at the center of effective digital defense

Key takeaways

1. Skilled content moderation – which includes fraud prevention, bot and spam detection, and removal of fake user profiles, in addition to addressing abusive posts and comments – is a necessity for all organizations that rely on user-generated content (UGC).

2. Increasingly sophisticated artificial intelligence (AI) aids content moderation at scale by automatically identifying the most egregious content, but it still requires human assistance to understand situational and cultural nuance.

3. Because content moderation is primarily a social, rather than a technological, problem, its solutions must ultimately center around human judgment and expertise.

Digital information has become so ubiquitous that some scientists now refer to it as the **fifth state of matter**. User-generated content (UGC) is particularly prolific: in April 2022, people shared around 1.7 million pieces of content on Facebook, uploaded 500 hours' worth of video to YouTube, and posted 347,000 tweets **every minute**.

Much of this content is benign – animals in adorable outfits, envy-inspiring vacation photos, or enthusiastic reviews of bath pillows. But some of it is problematic, encompassing violent imagery, mis- and disinformation, harassment, or otherwise harmful material. In the U.S., four in 10 Americans **report** they've been harassed online. In the U.K., 84% of internet users **fear** exposure to harmful content.

Consequently, content moderation – the monitoring of UGC – is essential for usable online experiences. In his book *Custodians of the Internet*, sociologist Tarleton Gillespie writes that effective content moderation is necessary for digital platforms to function, despite the "utopian notion" of an open internet. "There is no platform that does not impose rules, to some degree – not to do so would simply be untenable," he writes. "Platforms must, in some form or another, moderate: both to protect one user from another, or one group from its antagonists, and to remove the offensive, vile, or illegal – as well as to present their best face to new users, to their advertisers and partners, and to the public at large."

Content moderation is used to address a wide range of content, across industries. Skillful content moderation can help organizations keep their users safe, their platforms usable, and their reputations intact. A best practices approach to content moderation draws on increasingly sophisticated and accurate technical solutions while backstopping those efforts with human skill and judgment.

Content moderation is a rapidly growing industry, critical to all organizations and individuals who gather in digital spaces (which is to say, **more than 5 billion people**). According to Abhijnan Dasgupta, practice director specializing in trust and safety (T&S) at Everest Group, the industry was valued at roughly $7.5 billion in 2021 – and experts anticipate that number will double by 2024. **Gartner** research suggests that nearly one-third (30%) of large companies will consider content moderation a top priority by 2024.

Content moderation: More than social media

Content moderators remove hundreds of thousands of pieces of problematic content every day. **Facebook's Community Standards Enforcement Report**, for example, documents that in Q3 2022 alone, the company removed 23.2 million incidences of violent and graphic content and 10.6 million incidences of hate speech –

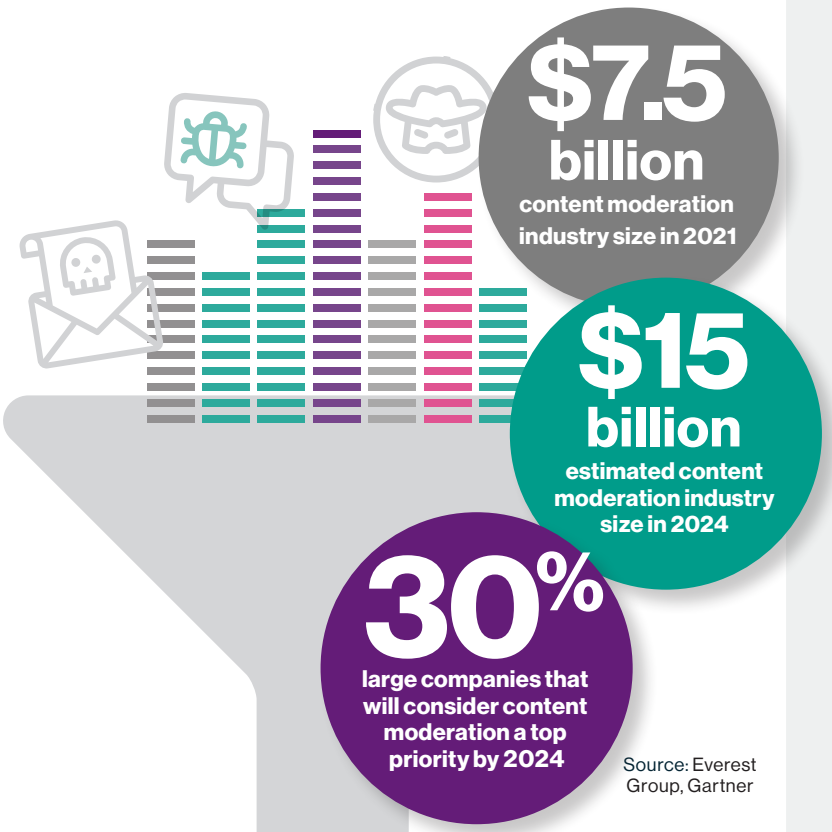
"Any site that allows information to come in that's not internally produced has a need for content moderation."

Mary L. Gray, senior principal researcher, Microsoft Research

in addition to 1.4 billion spam posts and 1.5 billion fake accounts. But though social media may be the most widely reported example, a huge number of industries rely on UGC – everything from product reviews to customer service interactions – and consequently require content moderation.

"Any site that allows information to come in that's not internally produced has a need for content moderation," explains Mary L. Gray, a senior principal researcher at Microsoft Research who also serves on the faculty of the Luddy School of Informatics, Computing, and Engineering at Indiana University. Other sectors that rely heavily on content moderation include telehealth, gaming, e-commerce and retail, and the public sector and government.

In addition to removing offensive content, content moderation can detect and eliminate bots, identify and remove fake user profiles, address phony reviews and ratings, delete spam, police deceptive advertising, mitigate predatory content (especially that which targets minors), and facilitate safe two-way communications in online messaging systems.



Source: Everest Group, Gartner

Different tiers of harmful content

Though there is no universal consensus on what constitutes "problematic" or "harmful" digital content, there are distinctions among what's known in the industry as "egregiousness." Though classifications vary, generally speaking, content may be highly egregious (including such things as violent and graphic content, suicide and self-injury, child nudity); egregious (hate speech, propaganda, bullying and harassing behavior); or non-egregious (brand safety measures, ad moderation and placement, mislabeled or mistagged data).

A recent report by HFS Research and Teleperformance states that the majority of content moderation work performed by humans falls into the non-egregious category. Akash Pugalia, the global president of trust and safety at Teleperformance, also points out that human judgment is often a key differentiator in the egregious category.

Research done at the University of Colorado, Boulder, has identified key dimensions that impact the severity of harmful content. Creating frameworks for understanding which types of content are most harmful can aid in creating both global policy and platform community standards. The finding that more severe content is exponentially more harmful than less severe content may help platforms prioritize their approaches to content moderation and reduce harm to both platform users and human moderators.

As technology continues to evolve and improve – and as the content moderation industry expands – protecting human moderators from exposure to the most egregious content continues to be a top priority.

One area of serious concern is fraud, especially on e-commerce platforms. "There are a lot of bad actors and scammers trying to sell fake products – and there's also a big problem with fake reviews," says Akash Pugalia, the global president of trust and safety at Teleperformance, which provides non-egregious content moderation support for global brands. "Content moderators help ensure products follow the platform's guidelines, and they also remove prohibited goods."

Technology enables content moderation at scale

Most content moderation systems employ artificial intelligence (AI) to tackle the scope and scale of the work involved. The sheer amount of content produced on a minute-by-minute basis is too vast for humans to parse, and machines can dramatically reduce the volume of content that needs human oversight.

Content moderation algorithms are typically trained using machine learning and deep neural networks. Different tools specialize in different types of content: automated, API-enabled natural language processing algorithms detect harmful written content; speech-to-text models handle video and spoken content; and computer vision technology monitors images and visual content. Other content moderation AIs flag anomalies in user identity and account verification scenarios.

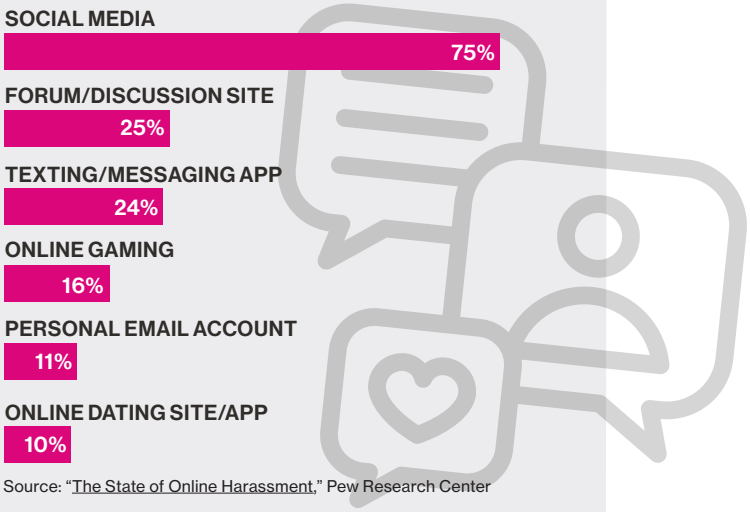
Advances in computational sophistication and power have made this type of AI accessible to more organizations. Other technological improvements mean that the most offensive, violent, or otherwise harmful content – what's called "highly egregious" content – is very likely to be flagged before it ever reaches human eyes.

For example, Facebook's Q3 2022 **Community Standards Enforcement Report** notes that automation enabled it to proactively detect and remove 99.1% of terrorist content, 99.1% of violent and graphic content, and 98.9% of child endangerment content. Similarly, Teleperformance's internal data cites automated proactive removal rates of up to 99.7% for fake accounts and 98.9% for inappropriate comments.

But in other cases AI requires substantial human assistance, especially when it comes to identifying context or navigating cultural, regional, and social intricacies. For example, Facebook reports proactively identifying only 67.8% of bullying and harassment content in Q3 2022. Teleperformance cites automatic detection rates ranging from 75% to 85% in similar categories. Explains Pugalia, "There are challenges to understanding when hate speech, harassment, or bullying is happening. You have to not only understand that it is happening, but also the context of where it happens."

There are other notable limitations and challenges when it comes to the current capabilities of algorithmic content moderation – the risk of bias, for instance.

Places people were most recently harassed online



Source: "The State of Online Harassment," Pew Research Center

"On a daily basis, content moderators try to reduce bias as much as possible so that the data being fed to the AI model is robust, which in turn makes the AI more robust."

Abhijnan Dasgupta, practice director, Everest Group

Dasgupta points out that this is another area in which human guidance is key. “On a daily basis, content moderators try to reduce bias as much as possible so that the data being fed to the AI model is robust, which in turn makes the AI more robust,” he says.

Human judgment provides an essential backstop

Because content moderation is a social rather than technological problem, human content moderators will always be an essential element of the solution. “Human moderators are a force of good,” says Pugalia. “They are the first responders of the internet who remove the threat of bad actors so we can all work, play, and enjoy the internet without being exposed to disturbing content.”

The dynamic that Dasgupta describes, in which humans help refine training data for machine-learning solutions, is sometimes referred to as “human-in-the-loop” or “technology-assisted review.” ML-based AI is iterative, meaning the longer it’s around and the more training data it receives, the better it gets, at least in theory.

“Human moderators are the first responders of the internet who remove the threat of bad actors so we can all work, play, and enjoy the internet without being exposed to disturbing content.”

Akash Pugalia, global president of trust and safety, Teleperformance

Content moderation is just one part of a larger trust and safety strategy

Content moderation is just one aspect of the larger universe of trust and safety (T&S) strategy. “The principle here is very simple: you create a platform that’s safe for whoever is using it and that provides a nice user experience. As part of that, you want to have checks and balances, e.g., policies for that platform in terms of what content can be there, how users should behave, etc.,” says Abhijnan Dasgupta, a practice director at Everest Group.

In addition to content moderation, T&S encompasses the entire set of policies and practices that allow users to feel comfortable interacting online. These might include things like community guidelines, fraud investigation, data labeling, platform safety, ad moderation, and the like. “All of these services combined is probably going to be the moderation of the future,” explains Dasgupta.

He sees a paradigm shift occurring right now. Enterprises are increasingly turning their attention toward a holistic view of T&S versus being laser-focused on content moderation in a silo. This trend is only likely to continue as more formats for digital engagement emerge.

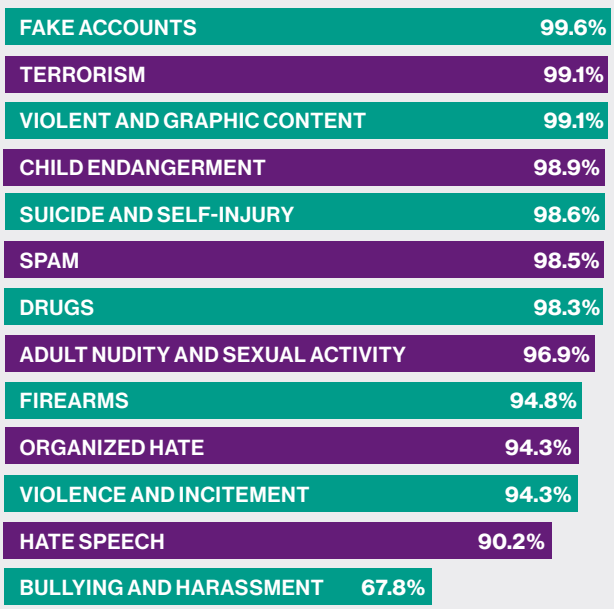
“As new kinds of content come in – let’s say, when the metaverse kicks in – all the stakeholders will need to get on the same page,” says Dasgupta, elaborating that those stakeholders may include platform owners, regulators, agencies who advertise on platforms, service providers who provide support, and payment compliance companies.

“All of those entities need to come together to create a holistic user experience,” he says.

In the context of content moderation, human workers help close the "machine learning feedback loop" by flagging content that escapes the algorithm. The AI then uses that data to make more accurate decisions in the future. "As moderators flag egregious content, it feeds back into a mechanism to ensure that the machine learning algorithms get better over time," says Pugalia, "so that next time that content doesn't even come for human moderation."

Facebook's proactive content removal rates, Q3 2022

Content removed before being reported by users



Source: Facebook Community Standards Enforcement Report Q3 2022

Human input is also critical when it comes to gray areas – content that's not obviously or overtly harmful but may be inappropriate in a more arbitrary way. Pugalia notes that human intervention in the "egregious" category is common. For instance, AI may be excellent at identifying videos that depict extremist violence, but less adept at identifying subtle recruitment techniques. It may bleep out or blur curse words with nearly 100% accuracy but struggle with a gesture that's innocuous in one culture yet taboo in another. "Machines and AI can't make that judgment call – should this content be on the platform or not?" says Pugalia.

Gray also points out that it's difficult for machines to identify "dog whistles," or instances in which political or otherwise inflammatory messages are encoded with specific, seemingly innocuous words or phrases understood only by the intended recipients. "Put plainly, there isn't going to be a technical identification that can keep up with how quickly an end user or communities of end users will come up with ways of coding their language," she says. "It's a persistent problem that technically doesn't have a solution because socially, we don't have a solution."

Human intervention is also useful for emerging types of digital interactions. The rise in live-streamed video, for example, often requires real-time monitoring, or "live moderation." This is an area where both human and technological skill is necessary for a timely response. "Live moderation essentially means moderating large amounts of data in real time. So you need speed, but at the same time you also need a lot of accuracy and context," says Dasgupta.

Pugalia adds, "In significant real-world incidents, content moderators have had a crucial impact in quickly and effectively flagging violent terrorist material." He cites Twitch's human-centered content moderation process that enables a quick response to live-streamed violence. "While we use technology, like any other service, to help tell us proactively what's going on in our service, we always keep a human in the loop of all our decisions," Rob Lewington, vice president of safety operations at Twitch, **told *The Washington Post***.

Improving the human-AI dynamic

Because the future of content moderation will rely on both humans and machines, creating a more effective

and safer human-AI dynamic is necessary to ensure both the welfare of human moderators and the online communities they protect.

One major concern, of course, is the impact that exposure to harmful content has on human content moderators' mental health and well-being. This type of work is emotionally taxing – and while technology helps absorb and deflect some of that burden, there's increasing attention being paid to employment practices and working conditions.

"One thing that's changed in recent years is public awareness that these are people's livelihoods and lives – I think that's the first step," says Gray. "There are more people with a hand in what we take as automation than seems the case."

Given the importance of human content moderators, more companies find that supporting their well-being is essential. Although many organizations still have a long way to go, there is an increasing industry understanding that improving human moderators' working conditions, trust and safety training, and access to support resources are critical to success in this line of work.

Teleperformance is investing in wellness research and partnering with startups and academics to identify ways to improve its human moderators' experiences. Pugalia attributes Teleperformance's success to this focus: "We take care of our content moderators. Best practices show the importance of investing in employee well-being with mandatory wellness hours, access to psychological services, taking breaks, and the like. We take this responsibility very seriously."

Another major difficulty for the human-AI dynamic is a lack of globally unified standards around what constitutes "harmful" content. Because there is also no worldwide consortium to dictate trust and safety best practices, each company or platform is left to its own devices to define them. They typically do this in their terms and services, which then serve as the blueprints moderators follow when flagging or removing content.

Julie Owono, executive director of Internet Sans Frontières (Internet Without Borders) and affiliate at the Berkman Klein Center for Internet & Society at Harvard, predicts that we may eventually see more alignment on content governance practices – but developing them will require significant collaboration. "Content moderation rules and practices, which I include under the umbrella of content governance, will evolve," she says. "Under increasing pressure from users, advertisers, and governments for safe online spaces, we may see the emergence of more common standards and, perhaps, common procedures. This will require a multistakeholder approach through which industry, civil society, governments, and academia collaborate."

The challenges facing platforms looking to moderate a vast and ever-expanding amount of online content are enormous. Still, the effort and investment involved in improving both the technology that underpins these systems and the working conditions for human moderators are worthwhile, especially as content becomes ever more central to digital interactions.

"The future of content moderation lies in a synergy between humans and technology – there are specific roles that each play," says Dasgupta. "While there's a lot of obvious opportunity to improve that dynamic, neither element is going away anytime soon."

Pugalia adds that because content moderation is a human problem, its solutions must also necessarily be human-centered: "There is a real need for humans to be part of the moderation process. These first-line responders moderate the worst internet content to not only protect us online, but their expertise helps make

"The future of content moderation lies in a synergy between humans and technology – there are specific roles that each play."

Abhijnan Dasgupta, practice director, Everest Group

"Humans at the center of effective digital defense" is an executive briefing paper by MIT Technology Review Insights. We would like to thank all participants as well as the sponsor, Teleperformance. MIT Technology Review Insights has collected and reported on all findings contained in this paper independently, regardless of participation or sponsorship. Laurel Ruma was the editor of this report, and Nicola Crepaldi was the publisher.

About MIT Technology Review Insights

MIT Technology Review Insights is the custom publishing division of MIT Technology Review, the world's longest-running technology magazine, backed by the world's foremost technology institution – producing live events and research on the leading technology and business challenges of the day. Insights conducts qualitative and quantitative research and analysis in the U.S. and abroad and publishes a wide variety of content, including articles, reports, infographics, videos, and podcasts. And through its growing MIT Technology Review Global Insights Panel, Insights has unparalleled access to senior-level executives, innovators, and entrepreneurs worldwide for surveys and in-depth interviews.

From the sponsor

Teleperformance, the global leader in outsourced customer and citizen experience management and advanced business services, serves as a strategic partner to the world's most respected brands across all industry types. It offers advanced business consulting services and end-to-end digital solutions that optimize both customer engagement and business efficiency. With the largest global workforce in the industry, and over four decades of deep, industry-specific expertise, Teleperformance supports billions of interactions each year by applying advanced innovations in automation, AI, machine learning, data protection, fraud prevention, and online safety services. For the next frontier of CX, Teleperformance is also leading the charge into the metaverse by designing and delivering immersive experiences between consumers and their favorite brands.

Teleperformance
each interaction matters

Illustrations
Cover art and spot illustrations created by Chandra Tallman Design LLC, compiled from The Noun Project.

While every effort has been taken to verify the accuracy of this information, MIT Technology Review Insights cannot accept any responsibility or liability for reliance on any person in this report or any of the information, opinions, or conclusions set out in this report.

© Copyright MIT Technology Review Insights, 2023. All rights reserved.

MIT Technology Review Insights

www.technologyreview.com

@techreview @mit_insights

insights@technologyreview.com

MIT Technology Review Insights

Produced in association with

Teleperformance
each interaction matters

A robust AI-human dynamic is necessary to ensure the internet stays safe for everyone.

Humans are essential to content moderation