VIEW POINT

BALANCING AI AND HUMAN MODERATION FOR SAFER COMMUNITIES: A STRATEGIC PERSPECTIVE

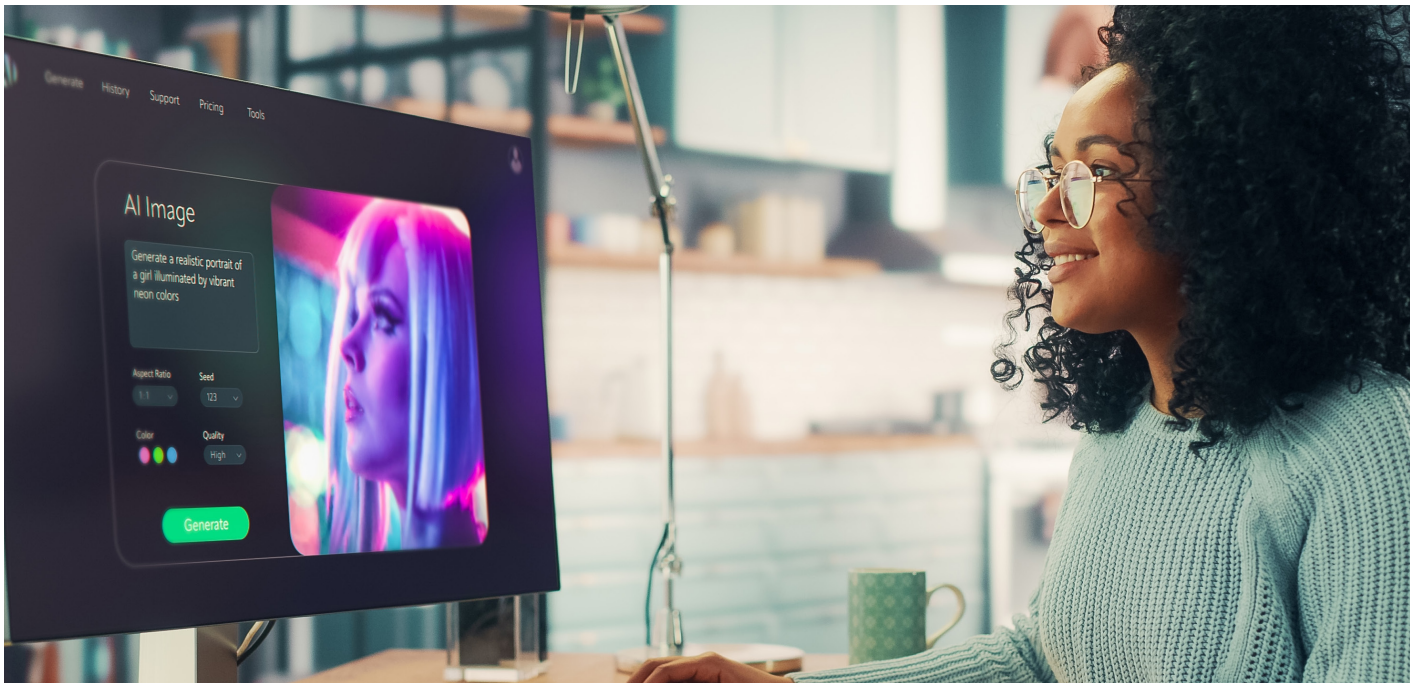
Abstract

The exponential growth of user-generated content poses unprecedented challenges for content moderation, requiring platforms to balance community safety, freedom of expression, and human wellbeing. With platforms like Meta processing 2.5 billion pieces of content daily, only human moderation has become impossible, making AI integration unavoidable. However, AI-driven systems struggle with contextual understanding and nuance. On the other hand, content moderators worldwide experience significant psychological distress from exposure to disturbing material. A hybrid solution combining AI scalability with human judgment is perhaps the best way forward to ensure systems serve both digital communities and their guardians effectively.

Infosys®
Navigate your next

We are bang in the middle of a digital age where we are confronting an unprecedented challenge: the problem of moderating the billions of pieces of User Generated Content (UGC) being churned out across the world daily, while maintaining community safety, protecting free expression, and preserving human wellbeing wherever relevant. Oftentimes, all three of these requirements may come to the fore in a single piece of UGC, depending on the geography.

Social platforms are among the foremost of many "content stewards" who are confronted with numerous challenges during content moderation. Let us examine what they are and how these are being tackled.



The scale challenge

Let's begin with the scale conversation. The sheer volume of online content today makes pure human moderation impossible. Consider Facebook, one of the oldest social media platforms. Facebook's parent company Meta processes 2.5 billion pieces of content daily, amounting to 500 terabytes of data. Between Facebook and Instagram, Meta has over 15,000 human moderators reviewing flagged items. Most content moderation decisions are now made by machines, not human beings, simply because human-only approaches cannot match the required scale and speed. Hop over to Google's video platform, YouTube, where over 500 hours of video are generated every minute. In fact, even text messages, one of the most rudimentary forms of communication today, reach up to 24.04 billion a day. It's quite evident that the scale is not humanly controllable any more, from a moderation perspective. This is where Artificial Intelligence (AI) can come into play, very adroitly.

AI excels at this volume challenge. In the fourth quarter of 2024, Facebook reported removing 5.8 million pieces of hate speech from the platform. Like others, Meta has been using AI-driven automated moderation systems to process content instantaneously, flagging obvious violations and immensely reducing the burden on human reviewers. Amazon Rekognition can identify inappropriate or offensive content at an 80% accuracy rate and remove it from the platform.

The context challenge

AI is not, however, a panacea for all content moderation ills. The greatest challenge with using AI lies in these systems understanding nuance and context. Enforcement that relies only on automation—when using technologies that have a limited ability to understand context—can lead to over-enforcement, which in turn disproportionately interferes with freedom of expression. Research from Youtube shows that AI-driven content moderation resulted in almost 50% of removal appeals being upheld, which is significantly higher when compared to appeals upheld by human moderation, which amounted to less than 25%.

This context deficit becomes particularly problematic in sensitive situations. From "traditional" social platforms, we have entered the era of social Virtual Reality (VR) platforms such as VRchat, AltspaceVR, Bigscreen, Rec Room, and Meta Horizon Worlds. Users in these platforms can participate in virtual reality experiences mimicking the real world — such as walking in a park or throwing a party — in highly realistic and immersive ways in 3D virtual environments, via stratagems such as real-time voice chat and avatars that are partly or fully body tracked. However, such unique digital realities are giving harassers the opportunity to "touch", "grope", and verbally harass other users in such a way that may be felt more strongly than in traditional social and gaming environments.

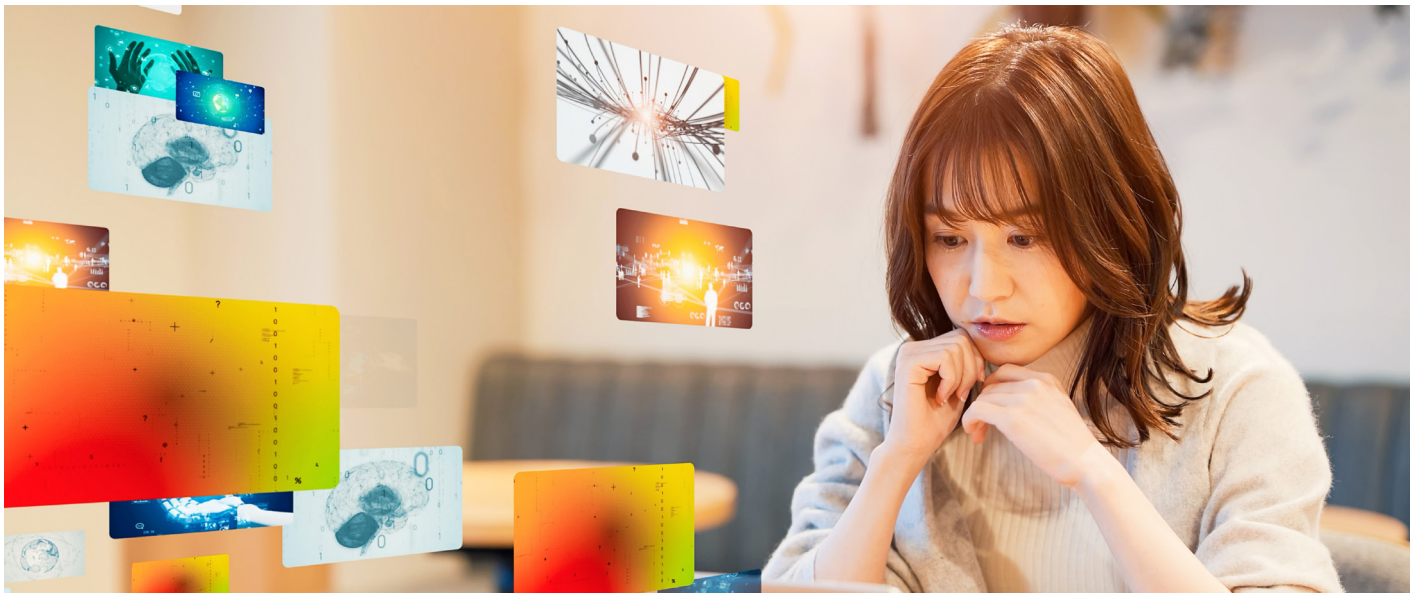
Such nuanced harassment requires a sophisticated understanding of human social dynamics, cultural context, and emotional impact, which many automated systems cannot do.

The human cost

Another, often hidden cost, to all this moderation activity has emerged. While human moderators provide the all-crucial contextual understanding, the psychological toll on these very moderators cannot be ignored. Over a quarter of content moderators demonstrated moderate to severe psychological distress in a survey, and a quarter of them were experiencing low wellbeing. Workers tasked with this labor of moderation endure psychological injury silently without support, and sometimes even face legal threats if they speak about it.

Approximately 100,000 individuals today work in commercial content moderation roles. Such moderators are frequently exposed to disturbing material that can lead to long-term psychological and emotional trauma. In fact, research reveals symptoms consistent with post-traumatic stress, including intrusive thoughts, triggered by situations with similar contexts to those encountered at work, avoidance behaviors and negative cognitive and emotional effects such as cynicism, anxiety, and detachment.

We need to protect the moderators from hate speech and violent content as well.



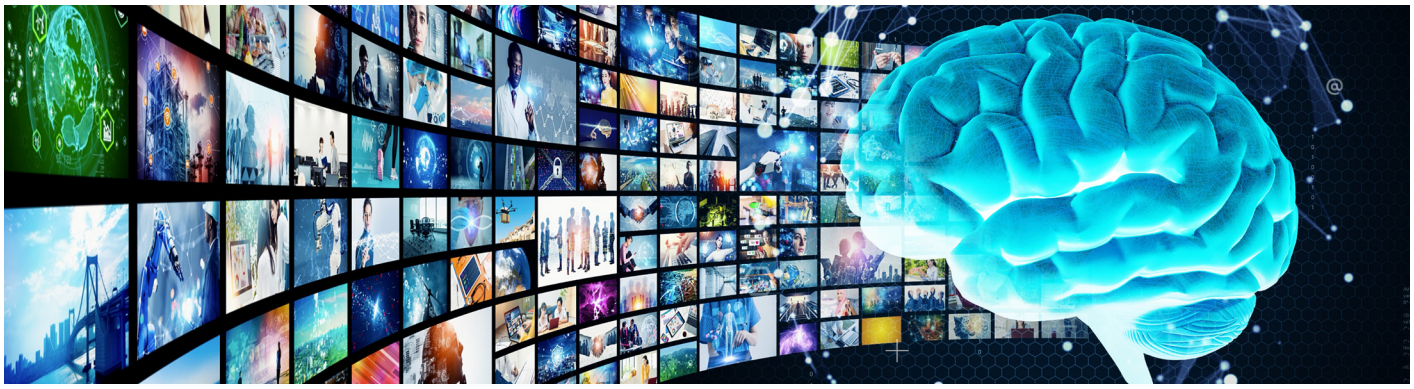
Addressing bias and fairness

Finally, the all-important question of bias. Both AI and human moderation suffer from bias, but in different ways. AI can reduce the subjective interpretation of data that humans tend to make, because machine learning algorithms learn to consider only the variables that improve their predictive accuracy. However, it has been noticed that facial recognition technology was significantly less accurate for people with darker skin tones, leading to higher rates of false positives. This kind of AI bias may creep in due to bias during training of the models.

We all know and recognise that while they bring a valuable contextual understanding of context, humans also introduce their own biases. For instance, the judgements of judges in legal cases can be unconsciously influenced by their own personal biases, while employers have been shown to grant interviews at different rates to candidates with identical resumes but with names considered to reflect different racial groups.

The hybrid solution

It has become rapidly evident that the most effective approach combines the scalability of AI with human judgment through intelligent collaboration paradigms. Modern trust and safety challenges require both technology and human expertise. This hybrid model creates multiple benefits:



First-line AI filtering removes obvious violations at scale, protecting human moderators from the most egregious content. AI technologies can capably blur sensitive images, redact offensive language, and even mask inappropriate audio recordings in real-time. By filtering out the most egregious content, AI can effectively reduce the terrible psychological burdens that are placed on human moderators

Human expertise can come into play for nuanced cases that require contextual understanding, cultural sensitivity, and empathy. Personalized interactions make platform users and content creators feel heard and respected, especially during appeals or sensitive disputes. This kind of empathy fosters stronger connections within communities.

Continuous learning allows AI systems to improve constantly through human feedback. Outcomes from flagged content, whether the moderator validated the flag or not, can be fed into the AI system's database, enabling it to detect similar content automatically in the future, without the need for users to flag it.

A hybrid approach also helps mitigate both types of bias through cross-validation. AI flags potential bias in human decisions and humans provide oversight for algorithmic blind spots that may have crept in while training.

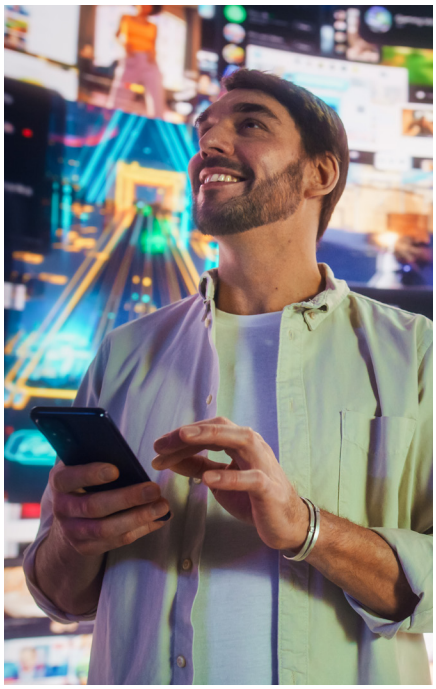
Making hybrid work

For hybrid moderation to succeed long-term, organizations must prioritize several key areas:

Transparency and accountability are essential. Third-party researchers, from around the world, should be given access to data allowing them to assess the impact of algorithmic content moderation. Platforms must give clear communications to users so that they understand how content moderation decisions are made. They must also have meaningful appeal processes.

Worker protection cannot be optional. Companies must invest in comprehensive mental health support, reasonable workload limits, and proper training. Content moderators need to receive sufficient training that focuses on building their resilience. Such training helps build stronger and healthier psychological coping skills.

Collaborative governance involving diverse stakeholders helps ensure systems serve the needs of the community rather than just corporate interests. AI must operate more at the level of expectation of human users. The onus is on the content platform companies to build AI that adapts to the unique environmental considerations of the platform.



The future of online safety depends on thoughtfully designed hybrid systems that leverage the scale and efficiency of AI, while preserving human judgment and empathy. This is not just a technical challenge — it's a business and human one as well, and it is a vital necessity to protect both our digital communities and the people who safeguard them.

Success will be measured not just by the volume of harmful content removed, but by the fairness of decisions, the wellbeing of moderators, and the trust communities place in these systems.

As we build the infrastructure for our digital future, we must ensure it serves humanity rather than diminishing it.

How can Infosys BPM help?

Infosys BPM helps organizations make the shift from reactive threat response to proactively embedding digital safety through human expertise, AI moderation, and strategic frameworks. Our deep Gen AI solutions and digital transformation expertise help build strong trust and safety capabilities across a range of sectors from eCommerce and gaming, to BFSI and healthcare.

For more information, contact infosysbpm@infosys.com

Infosys®
Navigate your next