wipro

Content Moderation

The art and science of making the internet safe and useful again

Foreword

Anees Ali Khan

362 likes

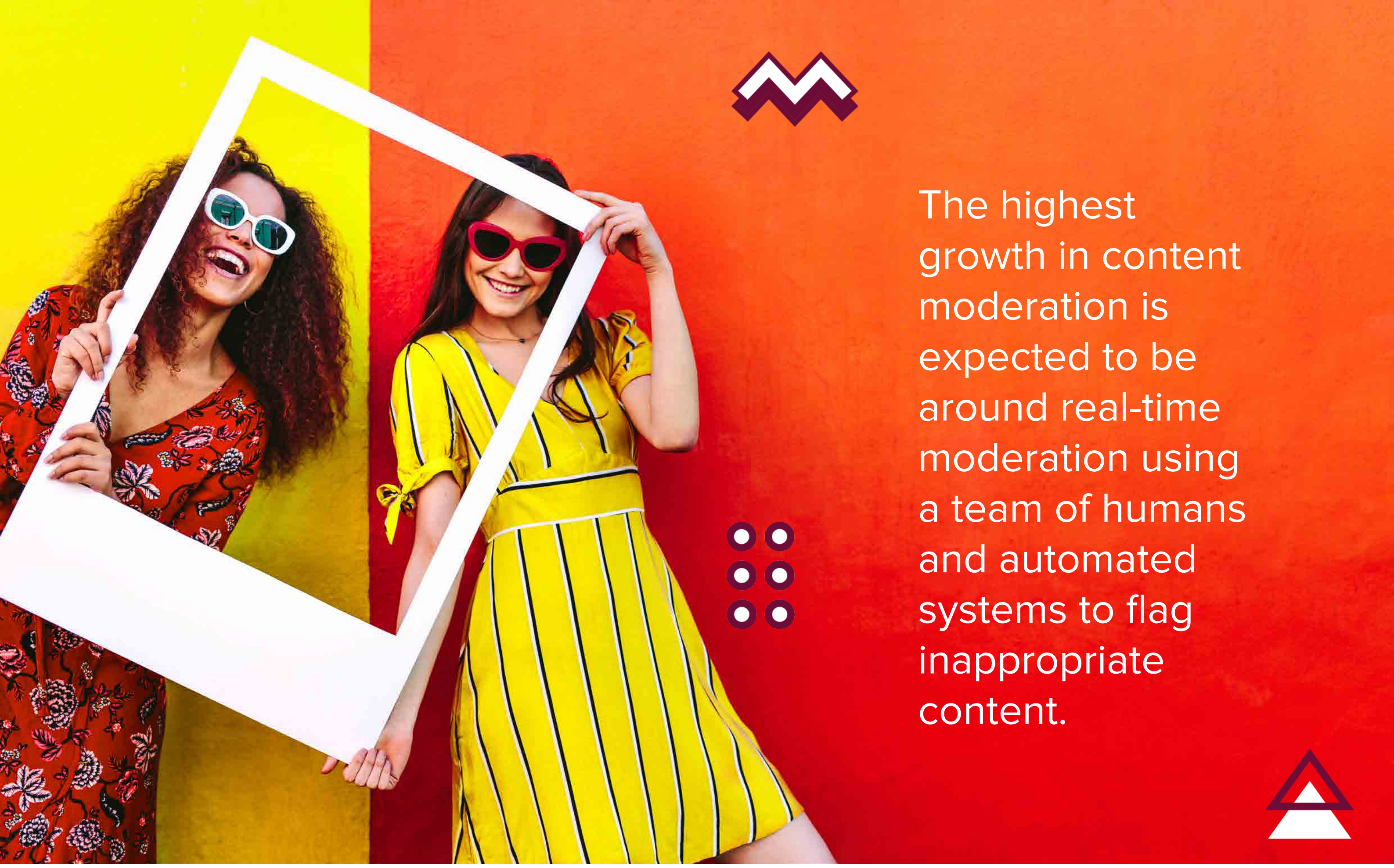
Vice President –
Consumer Business, Wipro-iCORE

#hashtag

Studies show that users on the Internet publish three times the amount of content available in all the libraries of the world and all archived movies and television broadcasts daily. We currently have 4.66 billion internet users, with 4.20 billion active social media users. Services like Substack, Medium, Wikipedia, Linkedin, TripAdvisor, Flickr, Mailchimp, Soundcloud, BuzzSprout, YouTube, Instagram, Twitter, TikTok and Facebook are leading the UGC revolution and are household names. Facebook has 243,000 photos published every hour; over 500 hours of video are published every minute on YouTube alone. A content tsunami is sweeping the world.

A substantial amount of this content is racists, designed to deceive, hate-filled and violates social and regulatory norms. This is why CMOs are worried. Globally, organizations are spending millions to moderate the content on their sites and services and keep the content safe, useful and trustworthy. These spends are going to grow several folds as the UGC trend becomes stronger, bringing content moderation under the ambit of an important business necessity. Which also means that a deeper understanding of UGC is required.

One reason for the growing need of moderation is the changing profile of online users—a larger proportion of the younger generation is spending more time online. For long, studies have consistently indicated that age makes a difference to who engages in content creation activities. Research show those under 35 have a greater propensity to create content. In 2019, 50 percent of internet users worldwide were in the 18 to 34 age group. In early 2021, 84 percent of those in the 18 to 29 age group and 81 percent in the 30 to 49 age group had used at least one social media site. These two groups form the bulk of customers for major businesses. What they say makes or breaks brands. It influences voters. It destroys reputations. It spreads fear and divides society. But with trustworthy content, they can also help communities find support and guide other to the right answers to questions when required. They can save lives and build better neighbors. Maintaining the quality of information and its relevance is critical to keeping the internet safe and useful.



The highest growth in content moderation is expected to be around real-time moderation using a team of humans and automated systems to flag inappropriate content.

Creating content is so simple that even a fifth grader can do it at will. Not surprisingly, content across types, geographies, and industries is growing in every possible language and format. This content has the power to make or break reputations, win customers, drive sales, win friends and influence governments.

Today, video content is growing the fastest – perhaps propelled by the fact that all it needs is a smart phone and an internet connection to take a message to the rest of the world within minutes. In addition, there is UGC in the form of comments on blogs, news sites, ecommerce platforms, networking platforms, etc.

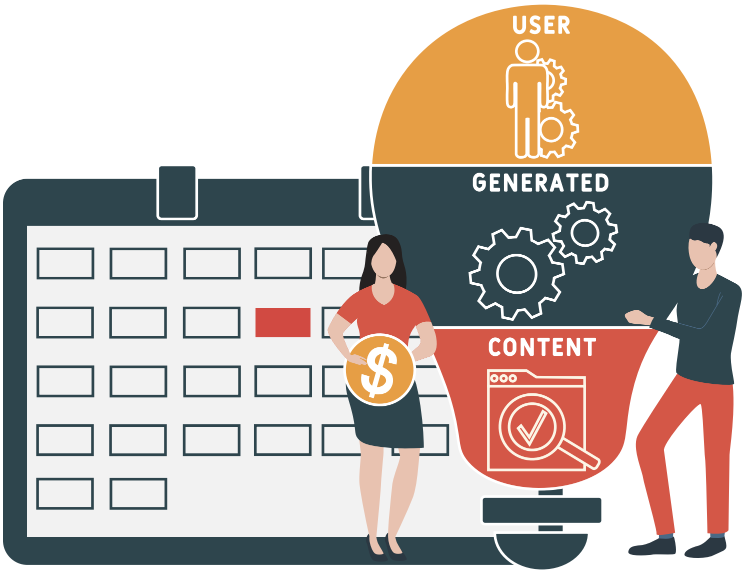
The urgent need for moderation

The challenge to moderate UGC is immense. Moderators are working 24X7, and using an array of technological tools to decide what stays and what goes from platform that publish user content. Yet, many of them see error rates of 10 percent and more in taking down content. As an example, if a platform that has two million posts a day, 200,000 errors slip past the moderation net, each with the potential to spread misinformation, trigger chaos, contribute to violence, impact national security, lead to poor decision making and increase psychological anxiety. The quest to distill good from bad is complex. It relies on the doggedness with which each business and government pursues the values it believes in. This also spells out the central problem that UGC presents: It is difficult to moderate content with consistency across the globe. There are cultural differences, a broad spectrum of language nuances, economic priorities, socio-political distinctions and organizational rules and varying ethical standards that make the task challenging.

Regional differences have led to Germany having its Network Enforcement Act, Brazil passing its Brazilian Law of Freedom, Responsibility and Transparency on the Internet bill last June, and the US having its Eliminating Abusive and Rampant Neglect of Interactive Technologies Act of 2020 or the EARN IT Act of 2020, and so on. Central to these positive efforts is a lack of global standards, defamation laws and legislation around UGC, making moderation inordinately sensitive and difficult.

As digital markets grow and penetrate deeper into every aspect of life, a close understanding of regional cultures, histories and legislation, while considering the local permissible limits of expression, has become central to the success of content moderation initiatives. Without localized expertise, content moderation will suck valuable dollars without creating the intended impact. And as the content grows in variety and volumes, moderation without automated and intelligent technology cannot keep pace.

Moderation is a tough job – but technology is coming to its rescue



Within reasonable limits, it is possible to have human powered moderation. With adequate training, humans can apply well-defined rules, local laws, sensibilities and contextual inferences to cull out what is acceptable and what is not.

Humans alone, as history has shown, cannot provide the level of moderation required. In 2017, a massive rash of crude, low budget videos aimed at children broke out on YouTube. They contained badly behaved Disney characters with nursery rhymes music, depicting sex and violence, obscene language, alcohol and drug abuse. Elsagate – as the incident came to be known—was traumatizing children. For months the videos went unnoticed by adults. When the videos found media attention, they were hastily mass deleted.

It is obvious from these examples that it is difficult for humans to moderate UGC because

of how it is positioned and its real-time nature. It is a losing battle. And this is just one reason an increasing amount of technology must support humans in taking swift and accurate decisions. There are other reasons technology will play a major role in content moderation: Laws change frequently, businesses have their own set of guidelines they want applied and new platforms appear that almost defy moderation at scale (example: Clubhouse, a social media app where users communicate in voice chat rooms).

To keep pace with the volumes, it is necessary to deploy Artificial Intelligence (AI), Machine Learning (ML) and automation to identify and quarantine unwanted UGC. Technology is also required to flag and escalate content instantly for humans to take moderation decisions. Most of all, the science of moderation needs to move from preventing to predicting and from updates to alerts in real time. The goal should be to reduce the burden of manual reviews without impacting the time taken for users to publish legitimate content.

The ideal solution is a hybrid approach that combines humans and technology to deal with moderation at scale, optimizes results and keeps the cost of moderation in check.

Wipro's experience in moderating UGC, ads, social pages, apps and shopping content, identifying and processing fraudulent transactions and patterns, reviewing suspicious users and performing background checks, monitoring social media for brand health and customer sentiment has led to a "Trust and Safety Service" (see table for details). The service is based on having carried out 52 million UGC reviews/filtering annually (for posts, comments, pictures, videos), 15 million ads and landing pages reviewed for content, quality and compliance annually, and 2.5 million posts analyzed annually for relevance and sentiment.

Wipro's Trust and Safety Service

Wipro in the Trust and Safety space – **What we do**

8,700+
associates

Working for the best media organizations in the world:
Largest search engine giant, Largest social media platform

Across
more than 6 countries

In more than
12 languages

15 Million reviewed annually

UGC review & filtering (Posts, comments, pictures & video

Dedicated employee wellness team

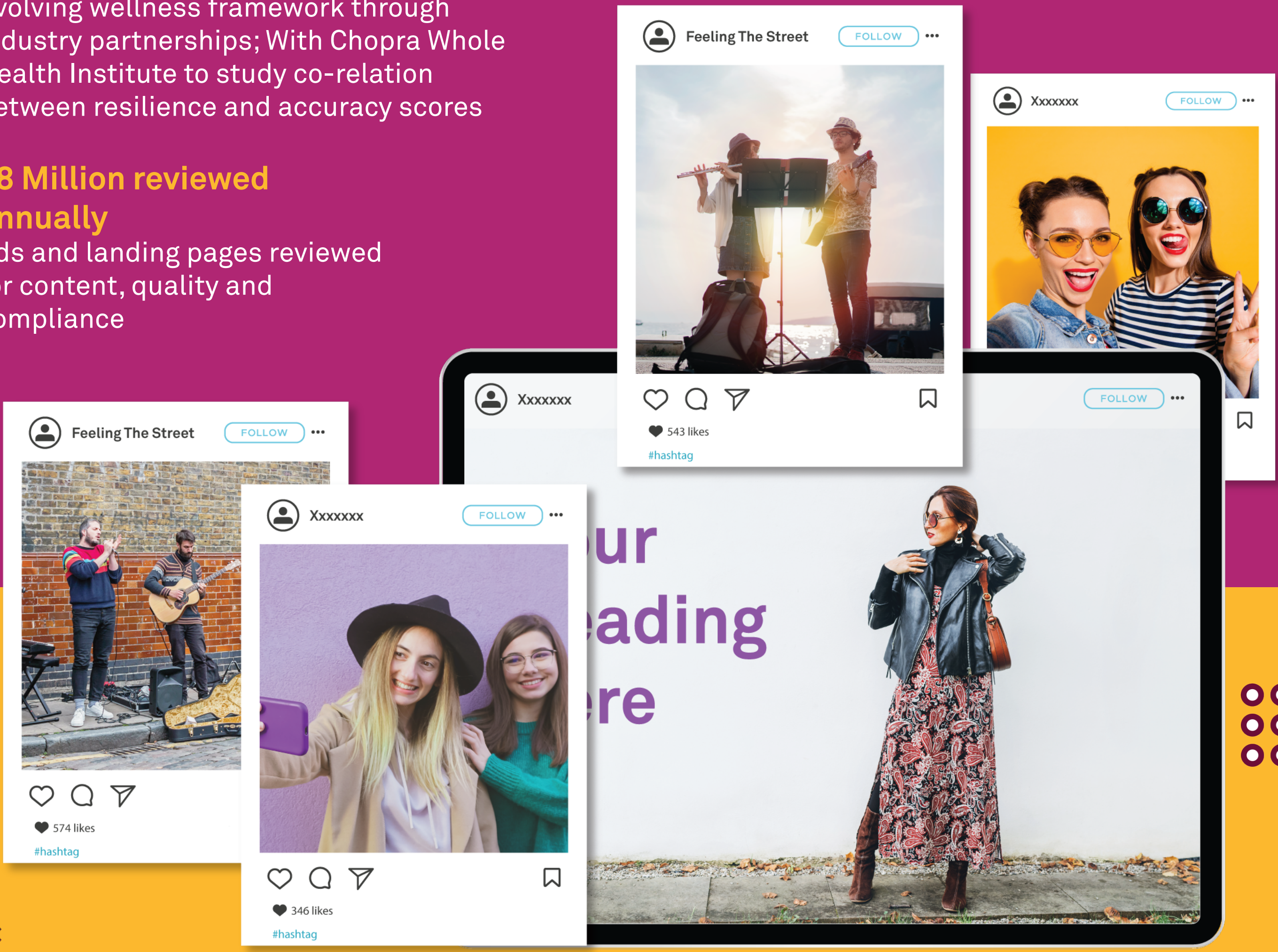
Evolving wellness framework through industry partnerships; With Chopra Whole Health Institute to study co-relation between resilience and accuracy scores

18 Million reviewed annually

Ads and landing pages reviewed for content, quality and compliance

Intervention tool

Veritas – Wipro fake news verification platform
Partnership with Community Sift and Emailage



Service	Service Details
Content Monitoring and Monetization Services	• Comment moderation • Pre-publish and post publish • Review moderation • In game chat and public forum moderation
Cyber Security and Risk Management	• Triage service to Federal services • Monitor and engage the online community • Protect the user base
Fraud and Payment Risk	• Run checks and detect anomalies • Classify and process fraud case • Analyze, identify pattern and update rules • Screening users for identity, background checks, online searches, high risk screening, data labeling • Validate payment URLs
App Review	• Monitoring and certifying apps, wallets and content within marketplace and online stores • End to end app certification • In app experience and validation • Shopping content moderation and app classification
Account Integrity	• Review suspicious user accounts, performing background verification checks • Analyze, identify pattern and update rules
Fake News and Event Management	• Triage support for reporting fake news • Anomaly detection during events • Database validation • Fact checking and data validation • Report writing and summary

Our content moderation solution is keyword driven (it matches keywords using Trigger Means Target model), and has a truth data store updated with facts after every interaction. It also leverages an automated verification engine based on NLP and intelligent pattern matching, Machine Learning to discern patterns for early warnings and alerts and a cloud-native workflow that allows systems to scale. Finally, it depends on human intervention at the right touch points to ensure nuance and context are captured and understood.

We also have two purpose-built applications designed for today's sophisticated UGC moderation needs:

Vantage, an AI-powered content intelligence platform that analyzes video and extracts key intelligence/meta data. It supports 125 languages. Brands, newsrooms, social networks and fact checking organizations can leverage Vantage to monitor unauthorized and unrelated usage of their brand and monitor for unauthentic news articles. Vantage has several application areas. It has delivered a 50% reduction of time in spotting content that does not meet the standards set by a leading agency. Another client, a leading global news agency, has used Vantage to reduce manual moderation by 75%.

AI-based Fake News Detector that verifies information, identifies the main source of the content, checks the reliability of the source, and determines if the content can be shared and published. With the usage of social media platforms and mobile devices, information is being created and turned viral in an instant. Automated classification of a text article as misinformation or disinformation can be flagged rapidly by our detector.

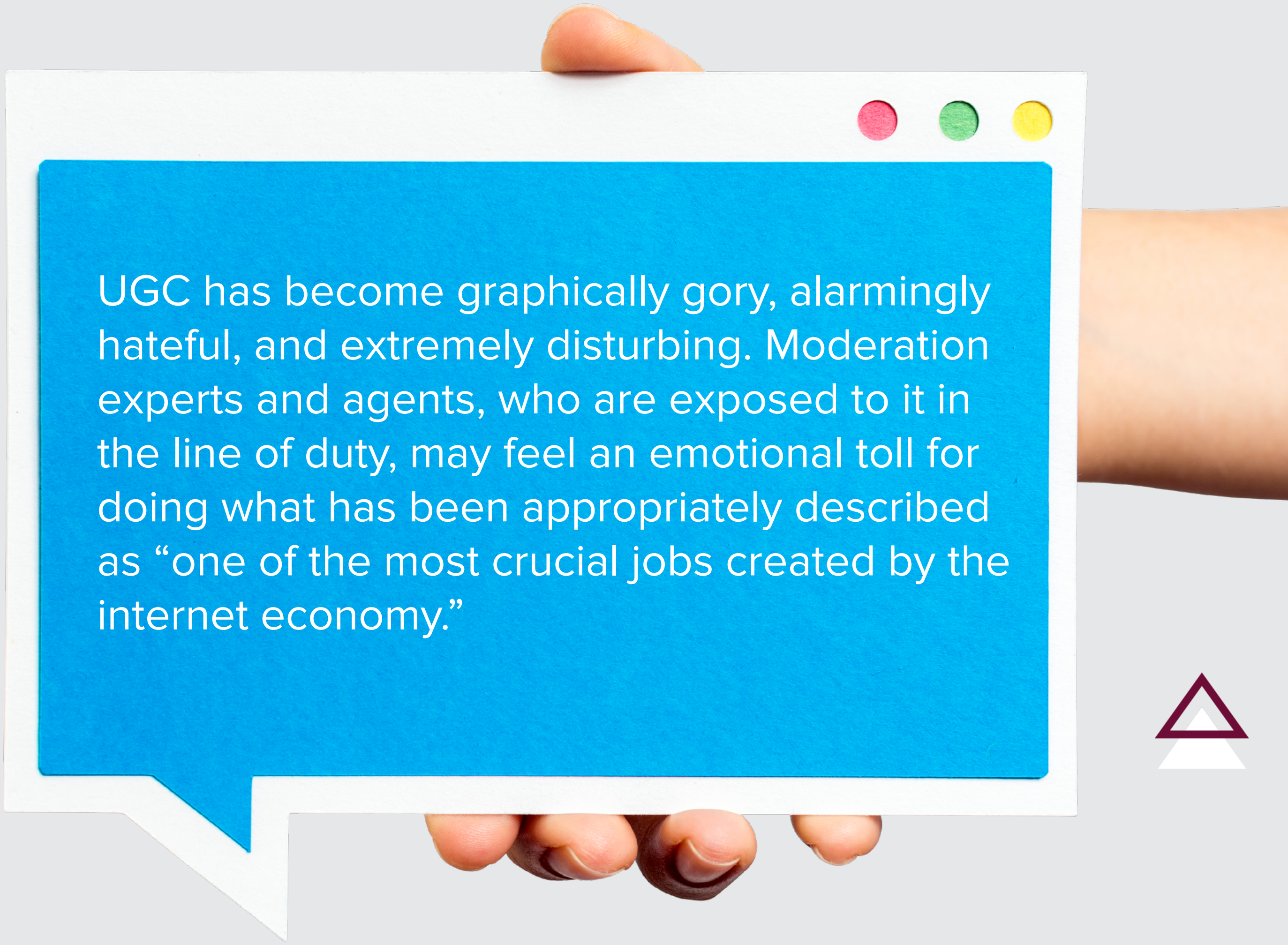
Above and beyond the technology, we rely on human intelligence. This continues to be at the core of fact checking and our related services. Our moderators and subject matter experts provide services around context detection, stance detection, bias detection and building taxonomies / ontologies for automation.

The results that our moderation services deliver testify to the industry experience we have:

An American online social media giant and social networking service company based in Menlo Park, California, wanted to differentiate offensive content related to graphic violence, nudity and pornography, and hate speech in two different languages. We worked with the company to bring a data driven approach and leveraged Machine Learning to fill the gaps in quality, going down to the agent level. The engagement saw an increase in quality of labeling across products by 2% and an 800% impact was created by validating a ML model which was later deployed

The world's leading search engine leverages 2,000+ resources from Wipro to ensure content is monitored and moderated to keep internet safe for its billions of users. The engagement includes 400+ FTEs with 3+ years of experience in Payment and Risk operations who monitor transactions on its online store for fraud.

An American online social media giant and social networking service company based in California wanted content moderation services for their Ad operations, E commerce and social media platforms. We reviewed 277Mn Ads annually to remove over 45,000 Child Abuse links from the social networking platform.



> UGC has become graphically gory, alarmingly hateful, and extremely disturbing. Moderation experts and agents, who are exposed to it in the line of duty, may feel an emotional toll for doing what has been appropriately described as “one of the most crucial jobs created by the internet economy.”

Professional hazards – the mental health challenge

Content moderation has its professional hazards. Content moderation is still at a nascent stage and there is limited scientific data and research published on the impact of this work on the wellbeing of employees. We can, however, draw on the extensive research that is available in other occupations that deal with secondary stress or vicarious trauma. Vicarious Trauma is triggered by direct or indirect exposure to someone else’s traumatic event--for example internet child exploitation investigators, emergency dispatchers and journalists covering traumatic events like wars, homicide, or a natural disaster. In one instance, a content moderator filed a case in a US court after suffering Post Traumatic Stress Disorder (PTSD) from exposure to videos of horrifying brutality, bestiality, murder and pornography. He told courts that “he had trouble with sleep disturbance, nightmares. He suffered from an internal video screen in his head and could see disturbing images, he suffered from irritability, increased startle, anticipatory anxiety, and was easily distractible.” This is not an isolated case. A 2018 documentary, called The Cleaners, reveals the trauma of human moderators working 8 hours a day, 5 days a week. The need to address the long-term impact on the mental wellness of the UGC moderation community is urgent. Automation and AI are one way of not only improving moderation outcomes at scale and at speed, but also a way of reducing the exposure to traumatic content for moderators.

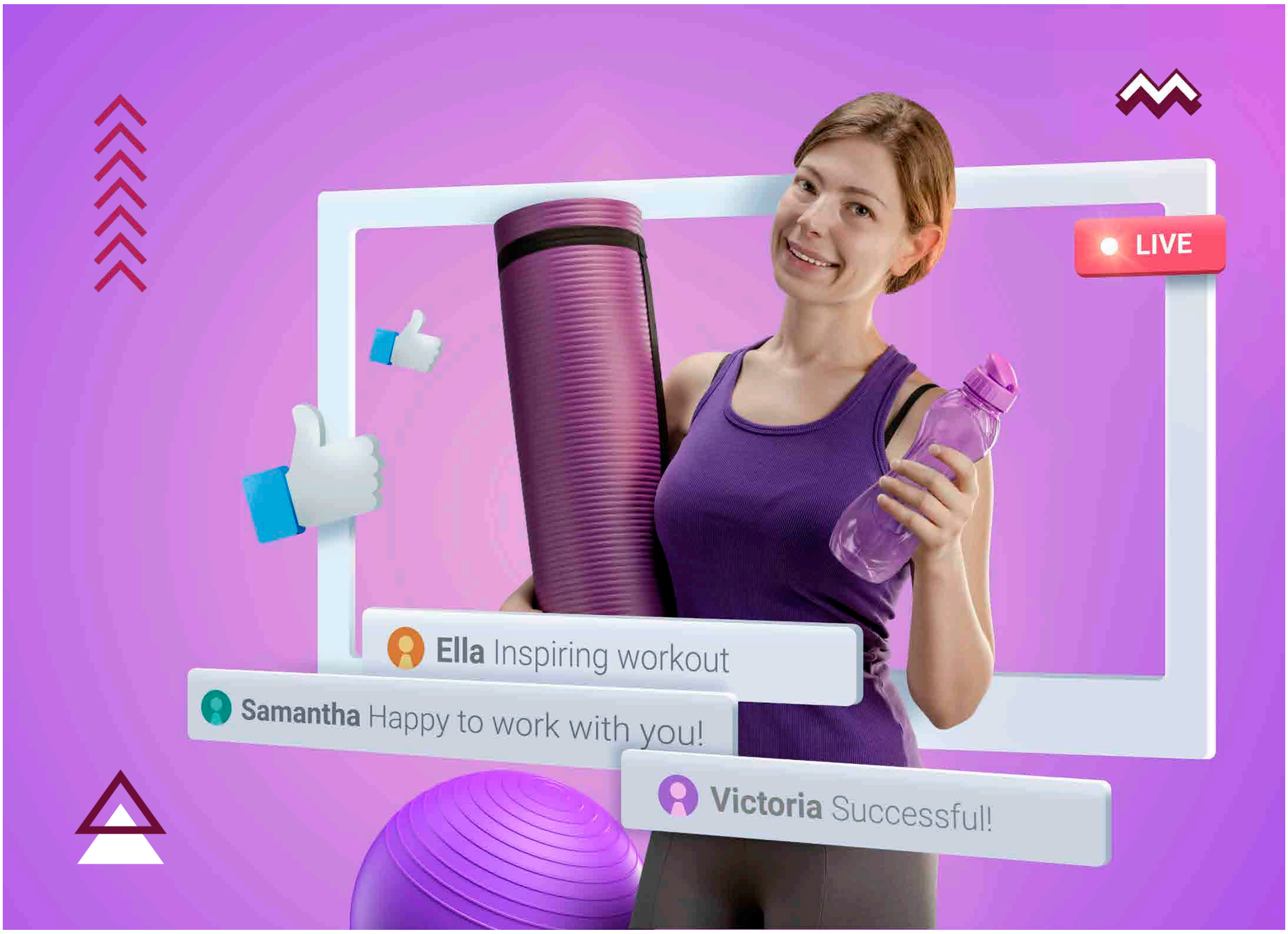
Given the impact UGC can have on moderators, it is vital that service providers such as Wipro draw a careful balance between moderator productivity and their mental wellness. Our approach is to place the wellness of our team at the center of recruitment, training and operations.

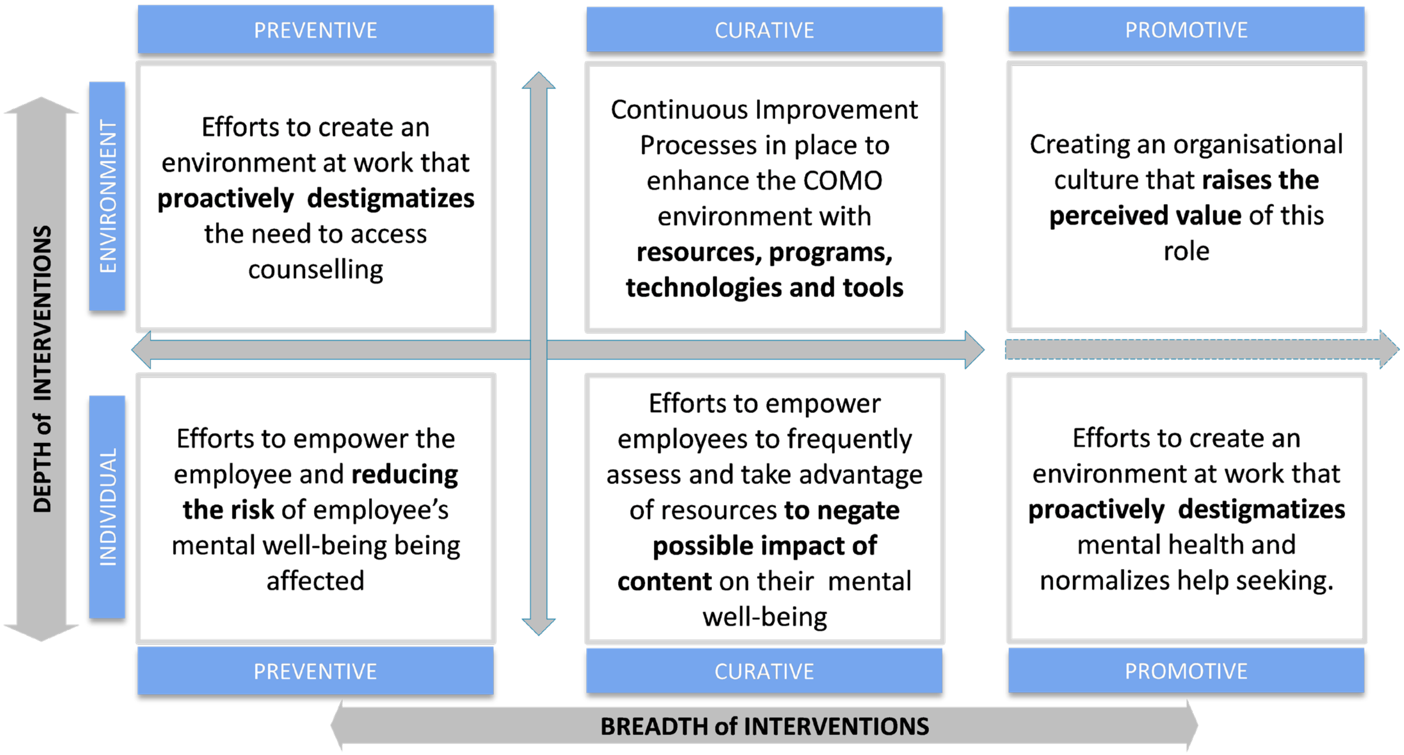
We have been intentional about the wellbeing of our people in the content operation space and have designed a wellbeing and resilience framework that is informed by evidence-based research and practice knowledge. This occupational and psychological health and safety model seeks to provide services at a preventative, curative, and promotive level, covering the individual, group, and organization as a larger ecosystem.

Preventive: The aim of preventive programs is to mitigate the risk of emotional distress. Preventive programs like psychoeducational groups that teach skills to enhance resilience are proactive in nature, as they equip a person to manage any potential distress even before it might take place. Overall, they provide a supportive environment that encourages psychological safety. Gear Up, our Resilience and Psychological Wellbeing Virtual Training and Onboarding Toolkit for New Hires is a classic example. We also identify any possible red flags so that corrective action is implemented to strengthen the protective factors of a person's state of wellbeing. An example of that would be the use of the pulse survey, DASS-21 (Depression, Anxiety and Stress Scale – 21) and the wellbeing check-in calls.

Promotive: The aim of these programs is to nudge content moderators to adopt a healthier lifestyle by taking increased emotional ownership for their wellbeing. Our 1 in 4 Global Wellbeing Virtual Summit organized at an organizational level aimed to destigmatize mental health and build awareness on mental health was one such initiative. These also include regular programs designed on sleep hygiene, healthier eating habits, emotional regulation, physical wellbeing etc.

Curative: The aim here is to reduce negative symptoms after their onset, with the intent to help a person regain a state of balance and emotional stability. Active steps are taken to mitigate further risk like providing individual coaching sessions and group sessions on coping and stress management.





DEPTH of INTERVENTIONS

	PREVENTIVE	CURATIVE	PROMOTIVE
ENVIRONMENT	Efforts to create an environment at work that **proactively destigmatizes** the need to access counselling	Continuous Improvement Processes in place to enhance the COMO environment with **resources, programs, technologies and tools**	Creating an organisational culture that **raises the perceived value** of this role
INDIVIDUAL	Efforts to empower the employee and **reducing the risk** of employee's mental well-being being affected	Efforts to empower employees to frequently assess and take advantage of resources **to negate possible impact of content** on their mental well-being	Efforts to create an environment at work that **proactively destigmatizes** mental health and normalizes help seeking.
	PREVENTIVE	CURATIVE	PROMOTIVE

BREADTH of INTERVENTIONS

The practical application of these dimension is evident when overlaid across the entire lifecycle of an employee — from pre-hiring to exit. Some of these have been captured below:

- Hiring process to find the right candidate with behavioral competencies
- Simulations to test the candidates on job aptitude, information gathering and synthesis ability before recruitment
- Use of psychometric testing tools like Connor-Davidson resilience scales to check the resilience scores of candidates
- Minimum of 6 to 8 hours on the concepts of wellbeing and resilience during employee onboarding
- A dedicated team of wellness leads at all locations to ensure wellness programs are sustainable
- Round-the-clock onsite counselling service for employees
- Dedicated HR team and psychiatrists conducting monthly 1X1 meetings and counselling with all agents
- Onboarding market leading wellness organization to support our operations
- Involving our leads and managers to understand risk patterns and behavioral changes for employees that can help us red flag issues in advance
- Increased governance and sessions for sensitive moderation workflows:
  - Consciously limiting the shifts of moderators involved in egregious and sensitive workflows to 5 hours a day (4 hours for extreme graphic videos) and for non-egregious workflows limiting the shifts to 6-7 hours a day
- 24 Hour employee helpline
- Monthly group theme-based workshops to create awareness of the risk of mental health

- Quarterly feedback/survey mechanism
- Resilience training programs
- Gamification of wellness and counselling activities
- Introduction of AI/ML driven wellness bots that provides wellness information along with the opportunity to have conversations and interactions in complete privacy. The bot:
  - o Acts as the first level of screening by proactively identifying patterns in mood/ behavior over time
  - o Provides behavioral change nudges in small steps and meaningful outcomes
  - o Provides solutions to moderators based on their mental and physiological profiles
  - o Helps create typified nutritional plans catered to reducing the stress level and overall wellness
  - o Provides a constant feedback loop that helps customize wellbeing programs and delivers insights from responses and data points to help teams/ counsellors
  - o Streamlines scheduling and calendar of the wellbeing programs
- Fun at work committee, events, R&R, yoga, meditation and Zumba classes
- Initiatives for differently abled individuals who may not be able to attend the physical fitness sessions

Using the latest technologies — AI, ML, AR/ VR, automation, NLP, and analytics — to aid content moderation is essential to success. These tools can adjust rapidly to volumes and changes in moderation rules and regulatory requirements. They can also improve outcomes with every iteration, eventually reducing the burden on human moderators and lowering the cost of moderation.

However, it is human moderators alone who can take decisions in grey areas where machines are bound to fail. To do this effectively, it is necessary to have a team that is mentally fit and can stay focused on its mission to keep the internet safe and useful.

https://www.hootsuite.com/pages/digital-trends-2021

https://www.statista.com/statistics/259477/hours-of-video-uploaded-to-youtube-every-minute/

https://www.sciencedirect.com/topics/psychology/user-generated-content

https://www.statista.com/statistics/272365/age-distribution-of-internet-users-worldwide/

https://www.pewresearch.org/internet/fact-sheet/social-media/

https://germanlawarchive.iuscomp.org/?p=1245

https://www25.senado.leg.br/web/atividade/materias/-/materia/141944

https://www.congress.gov/bill/116th-congress/senate-bill/3398

https://en.wikipedia.org/wiki/Elsagate

https://www.joinclubhouse.com/

https://www.thedailybeast.com/microsoft-anti-porn-workers-sue-over-ptsd

https://www.theverge.com/2018/1/21/16916380/sundance-2018-the-cleaners-movie-review-facebook-google-twitter

Wipro Limited
Doddakannelli,
Sarjapur Road,
Bangalore-560 035,
India
Tel: +91 (80) 28440011
FAX: +91 (80) 2844 0256
wipro.com

Wipro Limited (NYSE: WIT, BSE: 507685, NSE: WIPRO) is a leading global information technology, consulting and business process services company. We harness the power of cognitive computing, hyper-automation, robotics, cloud, analytics and emerging technologies to help our clients adapt to the digital world and make them successful. A company recognized globally for its comprehensive portfolio of services, strong commitment to sustainability and good corporate citizenship, we have over 200,000 dedicated employees serving clients across six continents. Together, we discover ideas and connect the dots to build a better and a bold new future.

For more information,
please write to us at info@wipro.com

IND/INT-CL/JULY2021-JUNE2022