Human vs. Machine: The Future of Content Moderation & What it Means For Your Business

The evolution of moderation, how AI is changing the game, and how businesses can prepare for the unknown...

The global population produce 4.6 billion pieces of content every day, with user-generated content (UGC) playing a key role in this proliferation of digital media.

In tandem with this burgeoning amount of UGC, the challenges of moderation have grown exponentially as brands strive to both preserve the integrity of their user experience (UX) and protect their reputation, without forfeiting user engagement in a digital wild west. It can be a tough line to tread, but a balance is being found with increasingly sophisticated AI content moderation solutions, along with tougher regulations and legislation globally. But can these measures keep up with the pace of online innovation, and what's to come next?

In an age when authenticity rules, and communities are king, UGC plays a powerful role for brands. In a survey, technology company Stackla found that 80% of customers reported UGC "highly impacts their purchasing decisions," while 88% said authenticity was important to them, and influenced their brand loyalty. It's easy to see why: unbiased information is the consumer gold standard for assessing vendors, and perceived corporate transparency can win over lifetime customers. But the warm and fuzzy feelings of an authentic, organic community easily dissipate if the very content that makes it feel sincere is left unregulated, with 40% of people online saying they disengage with a brand's community if they're exposed to just one piece of offensive content. Peering into 2023 and beyond, the outlook remains fraught with threats to companies who aren't on top of content review.

Over the past decade, the landscape has changed significantly. "People are always going to find ways to exploit technology," says Alex Popken, WebPurify's VP of Trust & Safety, and former Head of Trust & Safety Operations at Twitter. "The future of content moderation for UGC is really about seeing what new technologies are emerging and understanding what known and unknown risks they pose. It's a question of how we will evolve content moderation practices in line with these new technologies and the ways in which humans will exploit them."

So what does the next decade look like, from mental health in moderation to the metaverse and AI? Let's find out.

Contributors

ALEX POPKEN

VP of Trust and Safety, WebPurify

Over the past decade, Alex has led and made contributions to key milestones in content moderation for digital platforms. Prior to WebPurify, she led the Trust & Safety Operations organization at Twitter, where her remit included scaling the enforcement of community guidelines using technology and human review solutions.

JOSH BUXBAUM

Co-founder, WebPurify

Josh has over 17 years of experience working with brands of all sizes to solve their content moderation needs. As a co-founder of WebPurify, he immersed himself in the world of trust and safety while it was still a nascent concept, helping to bring what are now some of WebPurify's best-known solutions to market, from AI models for text, image and video moderation to dedicated, custom human teams - plus a unique, industry-leading combination of the two.

ALEXA KOENIG JD, PHD,

Co-Director, Human Rights Center & Adjunct Professor, UC Berkeley School of Law

Alexa is a leader and professor specializing on the impact of emerging technologies on journalism and international legal practice. She is also co-founder of the Investigations Lab, which trains students and professionals to use digital information to strengthen human rights documentation and reporting. Alexa has won multiple awards for her work, including the UNA-SF's Global Human Rights Award and the Mark Bingham Award for Excellence.

LAUREN KOESTER

VP of Marketing, ForeVR Games

Lauren Koester is a senior marketer passionate about building gaming communities and elevating VR. She has worked at the forefront of gaming technology with brands including Amazon, Microsoft, Xbox and Unity Technologies.

HUMAN MODERATION VS AI MODERATION

Content moderation has long been a necessary but often thankless task that historically fell to human moderators. However, with the sheer volume of content being produced every day, relying solely on human moderators has become increasingly challenging. As a result, many businesses are turning to AI and machine learning to manage the scale and reduce costs – with the added benefit of protecting human moderators, who otherwise would be called upon to routinely view what can be extremely upsetting content.

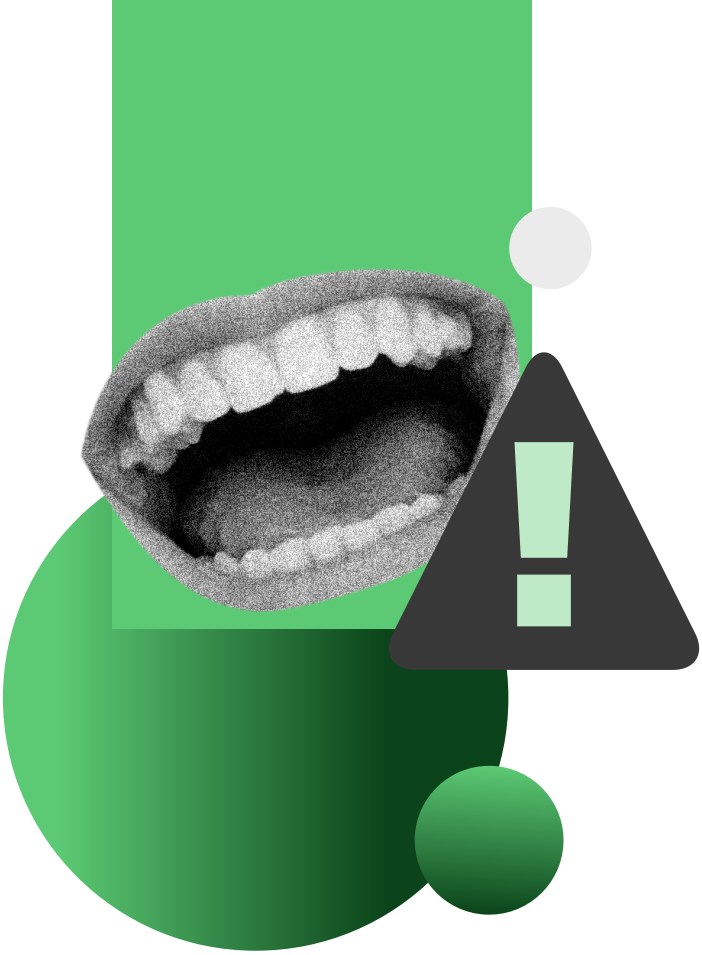
"Artificial Intelligence will play an even larger role over the next 10 years," predicts Alex. "AI has allowed us to moderate at scale for some time, but the explosion of large language models will have significant implications on content moderation. We'll see AI become a powerful tool to help scale and improve precision in content review. On the flipside, we'll see these models be abused by adversaries and require moderation themselves."

Alex anticipates that a hybrid approach of AI paired with human moderators will continue to be the most accurate way to moderate content, on the majority of platforms. "Human moderators will always have a role, in a few key ways."

1. Cultural Context & Complexity

"There will always be workflows that are too complex, nuanced, or context-dependent to rely upon machines. This might change in the distant future but we're nowhere near that," says Alex. AI models simply can't understand all cultural, geographic, or niche nuances. For example, how you accurately moderate content in the Catalan region of Spain versus the rest of Spain. There are regional linguistic and cultural differences. Humans are especially adept at understanding these subtleties. "There's a lot of languages on the planet with millions of speakers that simply do not have AI models, or at least not good ones."

Things move fast, and information spreads like wildfire. For example, the infamous 'Let's Go Brandon' catchphrase and imagery and captions associated with it will be missed by AI models. "Yes, you can add 'let's go Brandon'



to a block list, but will AI understand if a meme is mocking President Biden's age or his son? Unlikely," says Josh. "Another example is moderating forums discussing the Ukraine-Russian war. This is delicate, and lots of offensive things are shown and said but in the context of debate or journalism or denouncing atrocities. AI doesn't differentiate this stuff well."

Alexa Koenig, co-executive director of the Human Rights Center, UC Berkeley School of Law, agrees. "I think content moderation at its best is a partnership between machines and humans," she says. "We really need the humans to be determining the framework around what should stay up and what should come down. But then we also need to make sure that the tools that automate these processes, at scale, are really doing what they're designed to do."

At the same time as the machines become more efficient, the work of humans needs to become more subtle. "There's an increasing need to bring in more of a cultural perspective to content moderation policies," she explains. "That means thinking about the coded language that specific communities use; who in terms of age, gender and geography is using the tools, and in what ways."

2. Training models

"Machine learning relies upon vast amounts of data, and that data is trained by humans. Humans create 'golden set' data that both builds and audits machine learning models."

3. Ferreting out weaknesses

"Humans are able to identify gaps in machines. With the rise of large language models, we're seeing new job postings for 'AI prompt engineers', or individuals whose job it is to test the output of these models and expose weaknesses." Humans will also have a governance role, ensuring that the machines are built and maintained responsibly.

Part of that, she adds, is about "creating incredibly diverse teams. Companies need to be thinking about the diversity of not only the skill sets they need on their teams, but the insights that are contributed by having people from different geographic regions, with different lived experience and so on. I think that can be incredibly helpful."

For example, she notes: "What a woman from one part of the world will find potentially harmful may differ tremendously from a man in a very different part of the world. So thinking about how you get as many insights into the design of your content moderation policies as possible will provide a valuable set of insights for companies to work with. And that will hopefully help us become increasingly fine-grained in thinking about what we leave up and what we take down."

3 examples of nuance currently beyond AI moderation

1. Imminent danger or sexually suggestive

AI cannot detect scenes with impending danger – for example, a person holding a brick, about to drop it off a building onto a person's head. In the same way, it can't detect sexually suggestive images that don't actually show at least partial nudity. For example: the grabbing of breasts, while clothed, or somebody inappropriately gesticulating.

2. Benign words and images that, when combined, become inappropriate

One real life example, widely publicized, was an apparel company that absentmindedly modeled a shirt that read "cutest monkey in the jungle" on a black child. Humans quickly see where this might be offensive; AI misses it entirely.

3. Socially acceptable nudity and sexual imagery

Breast cancer awareness or breastfeeding imagery a platform may allow, but can be difficult for AI to tell apart from sexually explicit imagery – as AI simply sees nudity, and not the context. The same goes for photos of indigenous people, for whom partial nudity is accepted, non-sexual and part of their culture.

FUTURE TRENDS AND TECHNOLOGY

Here we explore the key trends and technologies dominating conversation and concerns for the next decade in content moderation.

Adoption of the metaverse (Web3) technologies

We are in the early days of metaverse technology, but improving graphics processing units, photorealistic 3D engines, content generation through volumetric video and AI are quickly shaping its future. And with it, extended reality (XR) isn't yet widely adopted – but predictions are that virtual reality (VR) and augmented reality (AR) headsets will surpass global game console shipments as early as 2024. Beyond that, it's likely to revolutionize human digital experiences across a myriad of applications. "As VR gets bigger, the challenges for moderating it are going to get bigger because of the scale," says Josh. "Now it's popular, but not everybody has a headset and not everybody's in it 24/7. When this technology is in every home, which it will be, because it's probably the coolest technology I've seen in the last 10 years, the problems scale with it."

Haptics

The haptic market is estimated to grow at a compounded growth rate of 12% by 2026 alone. What started as video game controls that vibrated to create tactile feedback is now advancing with haptics (also known as 3D touch or kinaesthetic communication) – the use of tactile sensations to stimulate the sense of touch. "With haptic suits, the idea is for it to become more real. Now, it feels real when someone encroaches on your space, but you're not actually feeling anything, in the tactile sense," says Josh. The technology emerged from the gaming industry, but promises great opportunities within the healthcare industry, workplace training and beyond. "Reaching a point wherein you're able to engage physically is a whole new level of reality. That comes with risks and considerations for content moderation."

Artificial Intelligence Generated Content (AIGC)

"We don't know how tomorrow's technology will be abused, but we do know Artificial Intelligence Generated Content (AIGC) is the next big threat in content moderation," says Josh. AI is already being used to create convincing deepfakes circulating online – with famous instances including Donald Trump's viral arrest images – and highlight how underprepared businesses and the public are against its abuse, as detection becomes increasingly difficult. While companies that enable users to create art with AI have tried to ensure that they can't produce offensive content such as pornography, users have easily found ways around this. "Inappropriate content, however it's made, can be detected, but having a human in the loop to spot complex and nuanced content like misinformation and deepfakes is likely to remain crucial deep into the future," says Josh.

Regulation and legislation

Another area that's seeing significant change is the emergence of rules to hold digital platforms accountable for online harms. The regulatory and legislative landscape that seeks to hold companies accountable for their online business practices has evolved significantly in the past ten years and will continue to expand worldwide. We will see new standards for the removal of illegal content, data protection and privacy for users, and antitrust guardrails. This means that companies will need to double-down on their investment in content moderation.

Mental health

As human capital continues to play a central role in moderation, and as a growing share of society spends more time online, another predicted trend within the industry is focusing on mental health, and rightly so.

"Any company concerned about their moral obligation to society, and their brand's reputation, should be focused on how the content they're hosting online can create offline harms including adversely impacting users' mental health," says Alex. "I believe we will see increased research from academics who are studying and shedding light on these impacts, and as a result, regulation. It also behooves any company working with moderators to ensure that there are wellness

standards in place for those who are exposed to sensitive content via moderation."

User controls and preferences

Users are demanding greater control over their experiences, from privacy to how their data is used and what content they're shown. Companies have taken heed, and cater to this with more robust user controls, highly tailored recommendation algorithms, and better tools for reporting abuse.

"More control will be put in the hands of users," predicts Alex. "Companies are learning that users want to curate their own experiences. A company's user base, collectively, wields ultimate power, and their opinions regarding how they want to engage online matter."

There is a shift away from the traditional 'keep up, take down' model of content moderation, towards more nuanced and contextualized forms of culling. "It's about providing more options than a binary like that," says Alexa. "More people in the Trust and Safety industry are asking, 'When can we give the user control over what they have exposure to, and options to minimize some of the stuff that they find upsetting, while still allowing them to get the information they need?'"

Does that mean moderation could become obsolete? "No," says Alex. "There will always be a need to remove the most egregious, illegal content out there, so moderation is here to stay."

HOW CAN BUSINESSES PREPARE FOR THE UNKNOWN?

Given the pace with which the landscape is changing, how can businesses be prepared? "Be nimble," says Josh. "With users coming up with new and creative ways to bypass moderation systems and technology, providing them with new avenues for producing offensive UGC, companies need to consistently adapt their approaches and guidelines."

There is a tendency for brands to implement content moderation in reaction, rather than proactively. "They are playing defense against the latest trust and safety crisis on their hands," says Alex. "It's not their fault – UGC risks can be difficult to predict, and technology evolves so rapidly. However, the brands who emerge ahead of the pack will be those who have a proactive strategy in place to address key challenges and anticipate trends. We recommend that part of this plan involves content moderation and partnering with solutions like WebPurify to bolster protections and mitigate risk to one's users and brand."

"If you are new to this area and you're not quite sure how to moderate content, make sure that you are consulting with experts in this domain who can help guide you towards the policies, processes and technology you need to ensure that you are maintaining a safe online space," says Alex.

1. Bake it into your development

"It's important to make sure that you are baking trust and safety practices into your company from day one," says Alex. "Don't assume you're going to be lucky and avoid the risks of UGC. Inevitably, there is going to be a crisis, and if you don't have the appropriate safeguards in place to moderate user-generated content, you're going to fall behind."

2. Stay aware of new technology

Being aware of the potential looming threats starts with understanding the technology of tomorrow: "Generative AI is creating a whole new type of content called AIGC. Imagine someone prompting a Generative AI to create text or an image to bypass most moderation systems?" says Josh. Business leaders need to stay current. "You can't be prepared to defend against something you didn't see coming."

4 Questions to evaluate your content moderation needs

Do you have a moderation model in place that could scale safely with your business, and content volumes?
AI plays a key role in speeding up review processes while minimizing costs, but human moderators remain the most reliable and adaptable form of content moderation. The right choice for your business depends on the content, platform, scale and budget. Consider the investment you're making today, and whether it's ready for your needs tomorrow.

Do you want to invest in-house or with an external content moderation partner?
Choosing between in-house or external content moderation depends on various factors such as the volume of content to be reviewed, timelines, budget, and required language competencies.

Do you have privacy and security measures in place?
Ensuring privacy and security of user data is essential. Data should be stored responsibly and measures should be taken to safeguard privacy. For example, internal software and hardware that prevents unauthorized screen grabs.

Do you and your content moderation partners have adequate mental health support in place?
Support systems that safeguard moderators' mental health create conditions that allow for consistent content moderation, and lower staff turnover. Retaining skills results in higher levels of expertise. This consideration is relevant from the recruitment process for personnel, to mandated breaks, training, and professional onsite mental health support.

Don't know where to start, or take your business next? Book a free consultation with WebPurify's VP of Trust & Safety, Alex Popken.

MODERATING THE METAVERSE

"The metaverse needs moderation to succeed," says ForeVR Games VP of Marketing, Lauren Koester. Like any online community and culture, gaming harbors a subset of disruptive and negative players. "There's always going to be a level of moderation required in video games. But as the metaverse has expanded, a difference has emerged that distinguishes it from its 2D counterparts."

"Your avatar is an extension of your physical self. In VR it's not just the standard moderation of bad words, toxicity," says Lauren. "It's a question of solving this 'physical' abuse that can make you feel violated."

As the metaverse and its occupants continue to expand, forward-thinking game developers are examining the potential solutions to ensure a safe and enjoyable virtual experience for all.

"Having moderation tools and ensuring we provide a safe place for everyone to play is our responsibility as game developers," says Lauren.

WebPurify has partnered with ForeVR to form world-first solutions for metaverse moderation. "We have dedicated teams, specializing in VR moderation, working on this project to provide 24/7 moderation in ForeVR's games," says WebPurify co-founder Josh Buxbaum. "And this is an area of content moderation we're expecting to see expand exponentially as adoption rises over the next decade."

ABOUT WEBPURIFY

WebPurify empowers communities to be their best with scalable hybrid AI and human content moderation solutions for the world's leading brands. Ensuring positive user experiences for millions of customers, from marketplaces to the metaverse, with multimedia content seamlessly filtered to any brand's specifications.

Learn more at www.webpurify.com