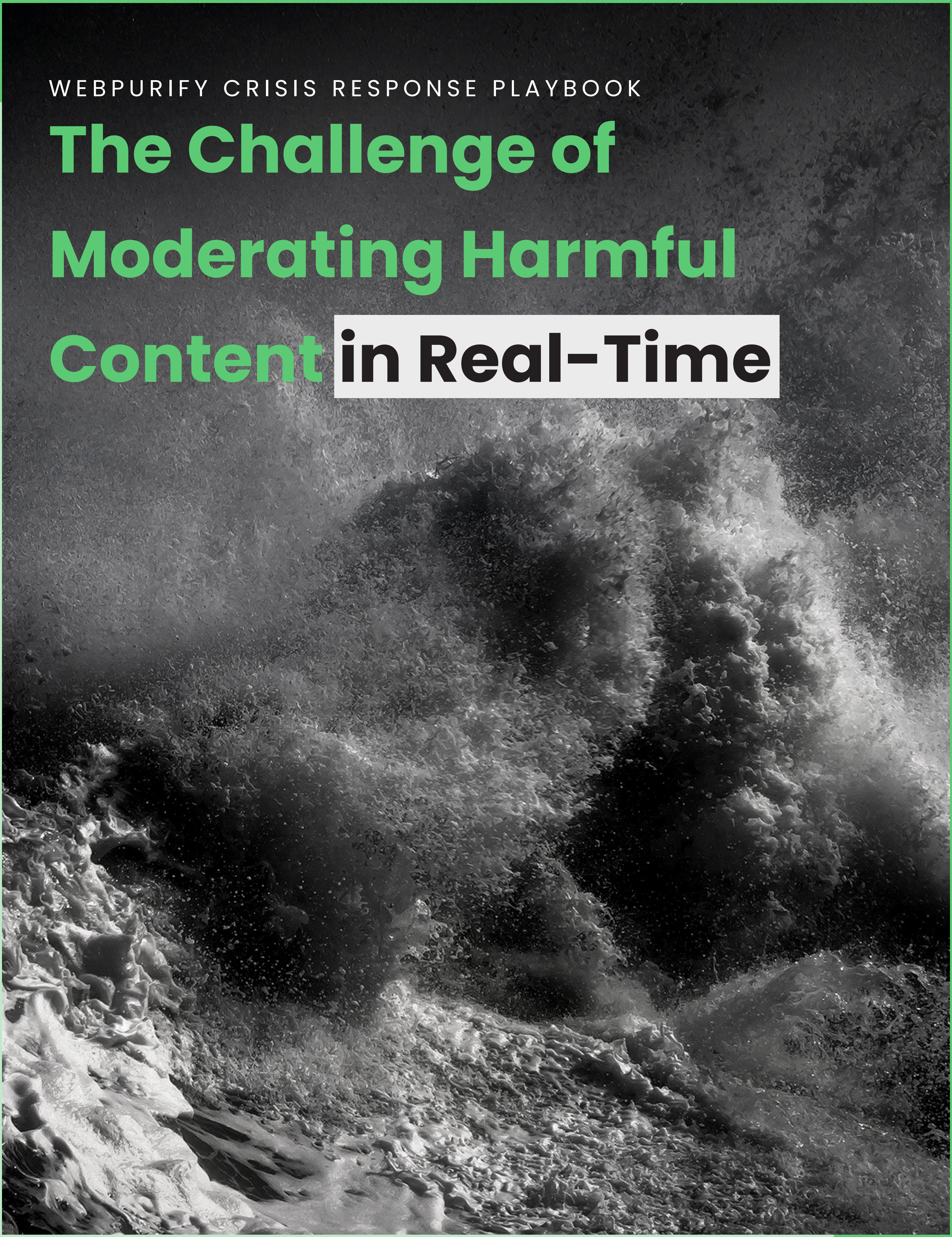
WEBPURIFY CRISIS RESPONSE PLAYBOOK

The Challenge of Moderating Harmful Content in Real-Time



In an increasingly interconnected world, breaking news events often ricochet across the globe in real-time. But with the rush of information comes the specter of harmful content, from gory images to graphic videos that threaten to flood user-generated content (UGC) platforms. Tragic events like the 2019 mass shooting at mosques in Christchurch, New Zealand, which was live-streamed, pose considerable challenges for content moderation teams worldwide.

Meanwhile, data collated by WebPurify casts an alarming light on the pervasiveness of this trend. As revealed in our recent Nefarious UGC report, research found that particularly egregious high-profile events – ones demanding immediate response via rapidly trained human moderators and AI – are being live-streamed and shared by users in real-time more frequently and persistently with every passing year. An analysis of approximately 38 widely shared violent events over the past decade reveals that close to three-quarters (71%) of these traumatic incidents took place in the last four years.

This stark escalation underscores the ever-intensifying challenge of managing crisis content on UGC platforms. The imperative to act is immediate and urgent, raising pressing questions about the responsibility of these platforms, their crisis response mechanisms, the mental health impact on users and content moderators, and potential strategies for mitigation.

This rise in harmful content creates a watershed moment for UGC platforms, forcing them to grapple with their roles and responsibilities in a digital age where users can become broadcasters of real-time, unfiltered events to a global audience. Content moderation in these instances is not just about removing explicit images or shutting down harmful accounts - it involves a complex interplay of ethical considerations, user safety, mental health implications, brand protection, and legal requirements.

How these platforms respond in the face of a crisis can define their very identity and shape the trust users place in them. This is where understanding the risks and responsibilities, as well as developing comprehensive crisis response strategies, become pivotal to maintaining a healthy digital ecosystem. We'll explore these factors in the following sections, including the immediate and long-term repercussions of these crisis events for UGC platforms.

The risks and responsibilities of UGC platforms in a crisis

”As the virtual meeting grounds of today, social media platforms are woven into our societal fabric, shaping our understanding and experience of the world,” says WebPurify co-founder Josh Buxbaum. “However, this also makes them conduits for harmful content, amplifying violence, terror, and despair. The unchecked spread of violent imagery can trigger widespread fear, possibly inspire emulation, and perpetuate harm to the victims involved by not respecting their dignity.”

UGC platforms need to pay particular attention to the sharing of harmful, violent content, especially during high-profile events, Josh says, and there are multiple reasons for this. First, there’s the very direct and immediate concern for user safety and emotional wellbeing. Violent or disturbing content can lead to various psychological issues for viewers, including post-traumatic stress disorder (PTSD), anxiety, and depression. “For many users, especially younger ones, exposure to real-world violence in such a raw form can have long-term detrimental effects,” explains Alex Popken, WebPurify’s VP of Trust & Safety.

Beyond the welfare of their users, platforms must also consider the impact on their brand’s reputation. Allowing such content to circulate can damage the image of a platform, associating it with violence and harm, which can drive away users, advertisers, and partners. As Josh notes, “In the long run, this could even impact the platform’s overall valuation.”

Regulatory risk is another important factor. Countries around the globe have stringent laws around the distribution of violent and harmful content. Platforms that do not adequately moderate this type of material can find themselves facing hefty fines, legal battles, and even enforced changes in platform operation or design. "It's not just about fines," says Josh. "It's about the potential restructuring of your platform forced by regulatory bodies."

Then there's the question of community trust. A platform's handling, or mishandling, of violent content can swiftly erode trust among its user base. Alex stresses the importance of this trust, noting, "If users feel unsafe or unsupported, they will inevitably look for alternatives."

Finally, the unchecked circulation of graphic and violent content can lead to the spread of misinformation, the fanning of social divisions, and potentially the incitement of further violence. "Uncontrolled distribution of violent content can contribute to a warped narrative, which can lead to further societal polarization and real-world harm," warns Josh.

Given these considerations, it's clear that platforms have both a social and a business imperative to take the sharing of harmful content seriously. It's not just about crisis management - it's about protecting users, preserving the brand, and promoting a safer, healthier online environment.

"Uncontrolled distribution of violent content can contribute to a warped narrative, which can lead to further societal polarization and real-world harm."

JOSH BUXBAUM

WebPurify’s Crisis Response Playbook

With 17 years in the business of content moderation, WebPurify is no stranger to the inherent challenges of managing user-generated content during such crisis situations. Over the years, WebPurify has honed a strategic crisis management playbook tailored to UGC platforms. This model, developed through constant learning, experience, and adjustment, is designed to enable platforms to respond effectively and ethically when tragedy strikes. Our playbook’s tactics and principles are built around proactive planning, immediate response, and diligent follow-through.

We know that when these sensitive events happen, they can go viral instantly. Because they’re so sensational and shocking, people are inclined to share them and within minutes this content can reach millions of people.

The moment something like this happens, typically the client will share it with the moderation team - or if not, the WebPurify team will hear about it and go searching for the content. The bottom line is: our team needs to get hold of this content quickly.

At the same time, we activate our training team and alert them that there is concerning content going viral and our moderators need to be trained on what to look for straightaway.

"As soon as a crisis emerges, we mobilize our teams to dissect the content minutely. We start by identifying distinctive features in the video and imagery that can then be used to train our detection systems and human moderators," explains Josh. "WebPurify's focus is on being alert and adaptable, preparing our quality control teams concurrently to handle the surge of content, making our crisis response a dynamic, real-time process."

Once the disturbing content has been found, the next step is to break it down to find the key frames within a video. Sadly, these are often the most violent frames. These scenes are then screenshot and placed into a training guide that is also sent to the Quality Control (QC) team, so they can start feeding it to the moderators.

Beyond training, WebPurify's system further integrates its crisis response by enabling the QC team to insert the newly identified content into the live moderation queue as a form of real-time simulation, even before the actual offensive content begins to surface. This test content, identical in appearance to the live content, helps to maintain the alertness of the moderators, as they are now seeing the content appearing in their queues. The objective is twofold: to keep the moderators vigilant for the disturbing content, and to give them an opportunity to acclimate to the unique identifiers of the new content, improving their ability to recognize it swiftly once it genuinely starts appearing in the live streams.

The key here is to identify the most distinguishable elements of a scene so that moderators can immediately tell if an incoming video is from the tragic viral event. These 'markers' might be the clothing the perpetrator is wearing, the perspective of the camera (often it might be an overhead CCTV), and signs or objects in the shot that make it distinctive. These guides serve to alert the team at a moment's notice so that they're on the lookout for certain visual cues.

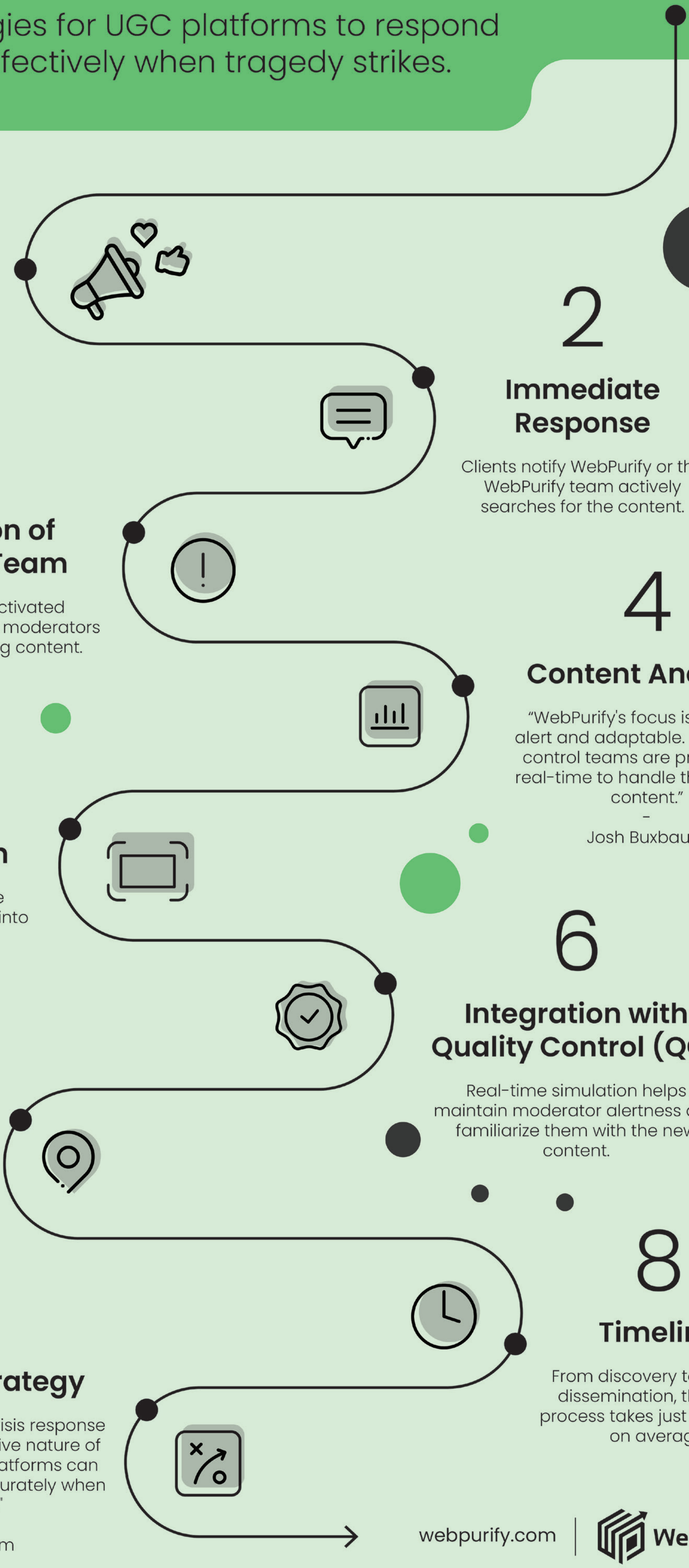
This whole process, from finding the graphic content to getting a training guide disseminated to the moderation teams, takes an average of 15 minutes.

"WebPurify's model of crisis response emphasizes the proactive nature of content moderation," Josh says. "By having strategies in place ahead of time, platforms can respond with speed and precision when a crisis hits, rather than scrambling in the aftermath."

WebPurify’s Crisis Response Playbook

–

Tailored strategies for UGC platforms to respond ethically and effectively when tragedy strikes.



1

Event Trigger

Sensitive events can go viral in minutes, reaching millions.

2

Immediate Response

Clients notify WebPurify or the WebPurify team actively searches for the content.

3

Activation of Training Team

Training team activated immediately to train moderators on new, concerning content.

4

Content Analysis

“WebPurify's focus is on being alert and adaptable. Our quality control teams are prepared in real-time to handle the surge of content.”

–

Josh Buxbaum

5

Key Frame Identification

Disturbing scenes are screenshot and placed into a training guide.

6

Integration with Quality Control (QC)

Real-time simulation helps maintain moderator alertness and familiarize them with the new content.

7

Unique Identifiers

Moderators are trained to spot unique visual cues for immediate recognition.

8

Timeline

From discovery to training dissemination, the entire process takes just 15 minutes on average.

9

Proactive Strategy

“WebPurify's model of crisis response emphasizes the proactive nature of content moderation. Platforms can respond swiftly and accurately when a crisis hits.”

–

Josh Buxbaum

Tips for Platforms and Content Moderation Teams on Managing Crises

Managing crises on UGC platforms involves a multi-pronged approach. It's not just about swift detection and removal of harmful content, but also about anticipating potential crises in the future - particularly around the anniversaries of past events - managing real-time incidents, and fostering a supportive post-crisis environment for affected users.

Navigating these complex situations requires a precise and thoughtful approach. Below, we'll explore some of our best, actionable tips for platforms and content moderation teams that draw upon WebPurify's extensive experience and insights in effectively managing crises, while prioritizing user safety and platform integrity in the face of real-time, high-stakes challenges.

Clear Community Guidelines

Foundational to any crisis response plan is a set of platform policies that establish clear boundaries around content that is prohibited on a platform and corresponding enforcement actions. "For sensitive events, a platform should evaluate whether or not they have clear guidelines around things like sensitive media which may include graphic and gruesome content that's upsetting and harmful for users," advises Alex.

> **"It's essential to have a pre-established chain of command"**
>
> ALEX POPKEN

Crisis Anticipation

"A good crisis management plan starts with anticipation," Alex says. This involves monitoring global events and understanding potential triggers for harmful content. Platforms can set up automated alerts to warn them of breaking news or establish a team whose responsibility it is to monitor emerging trends in key markets. AI algorithms should also be continually updated to identify new vectors of harmful content and moderation teams made aware of what to be on the lookout for.

Establishing a Chain of Command

Clear communication lines and quick decision-making are crucial during a crisis. "It's essential to have a pre-established chain of command," says Alex. "Decisions must be made swiftly, and everybody needs to know

their role. It reduces confusion and enhances the team's efficiency in crisis situations." Frameworks that can help to establish clear roles and responsibilities include the DACI matrix – Driver, Approver, Consulted, and Informed.

Real-Time Incident Management

During a crisis, platforms need to react quickly. This involves both automated and human responses. As Josh notes, "AI has a crucial role to play in real-time incident management, but it's equally essential to have a well-trained human moderation team ready to step in when necessary."

Transparency and Communication

Transparency and communication are key during a crisis. As Alex explains, "Users should be informed about what's happening and what measures are being taken to protect them from harmful content." This could involve releasing public statements or notifying users directly within the platform.

> **"During a crisis, it's vital for platforms to not only remove violent conent but also to provide resources for users who may be traumatized."**
>
> ALEX POPKEN

Navigating News Coverage

Managing a crisis on UGC platforms also involves careful navigation of news coverage. "News outlets have a responsibility to report, but these platforms need to find a balance between allowing essential information and preventing the spread of harmful content," Alex explains. "It's a delicate balance that requires careful content moderation, discerning between journalistic content and explicit or gratuitously violent material."

Providing Supportive Spaces

Platforms should also think beyond the immediate crisis. As Alex points out, "During a crisis, it's vital for platforms to not only remove violent content but also to provide resources for users who may be traumatized. This could include links to mental health resources, helplines, or even just providing a space for users to discuss their feelings."

Debriefing and Learnings

After a crisis has passed, debriefing and drawing lessons from the event is crucial. "What worked well? What didn't? What can we improve for the next incident?" Josh says, underlining the importance of continuous improvement and evolution in crisis response strategies.

"Clients also like us to apply a unique label when we encounter this content so they know how much of it has surfaced on their platform in, say, the last 48 hours," Josh adds.

Managing Mental Health

Beyond managing the content, platforms must also contend with the mental health repercussions on both their users and content moderators. "The repeated exposure to violent content can lead to long-term psychological harm. Users can become traumatized, and moderators often experience extreme stress, burnout, and sometimes even vicarious traumatization," Josh states.

Platforms, therefore, must prioritize the mental health of their communities. As Alex suggests, "Establishing robust support structures, including access to professional mental health resources, is essential."

At the core of WebPurify's operation is our Employee Wellbeing and Assistance Program (EWAP). EWAP is purpose-built to tackle the unique mental health strains inherent in the content moderation field and provides a comprehensive ecosystem of care with features such as round-the-clock counseling services, stress management initiatives, and a library of resources, reading materials, and interactive workshops.

Mindfulness training is also a key element of our approach. Alongside these offerings, we continually monitor our team's wellbeing, enabling us to proactively address stress and provide onsite counseling, complemented by follow-up sessions as needed. To us, the wellbeing of our team is not an afterthought, but a fundamental priority. For more on WebPurify's approach, see our eBook on Best Practices for Mental Wellness in Content Moderation.

In the face of rising global volatility, the challenge for UGC platforms is to navigate the double-edged sword of real-time, user-generated content. But by employing comprehensive crisis response strategies, prioritizing mental health, and acknowledging their role as guardians of shared digital spaces, platforms can rise to meet this challenge, protecting their users and, ultimately, themselves.