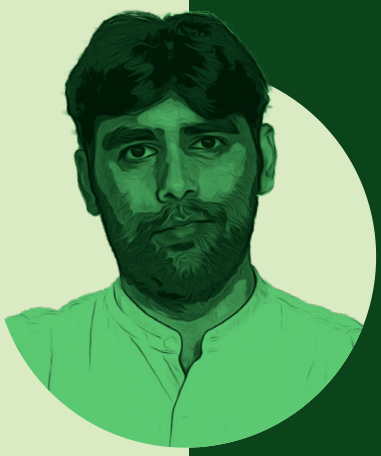
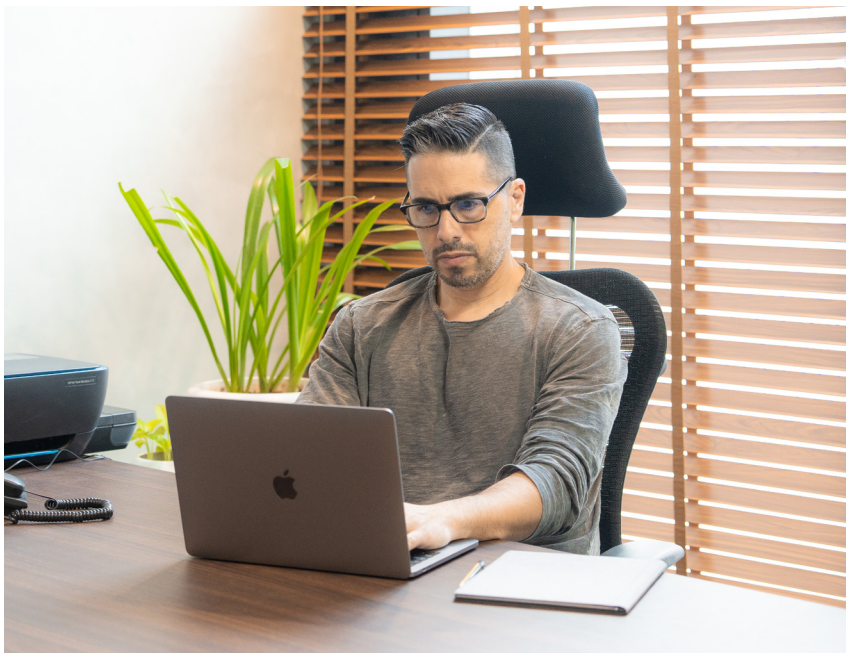


Faces *behind* the Filters

Meet WebPurify's Expert CSAM Moderation Team

"The amount we review daily can be discouraging." Josh Buxbaum co-founded WebPurify nearly two decades ago, knowing that seeing disturbing or offensive material is inherent to content moderation work. Here, "disturbing or offensive" might mean references to topics prohibited by community guidelines, slurs and hate speech, pornographic content and visual violence – but CSAM (child sexual abuse material) is among the worst of it and underscores the importance of moderation. "Sadly, we have seen an increase of this kind of content over the years," he explains. The high stakes of CSAM moderation means that not every company is willing to dedicate teams to the work – but for WebPurify, it was baked into their mission. "We know that even the small dent we are making is impactful and very well worth it."

To be in this business, you have to be mission-driven and fully invested. "You have to have a passion for what you're doing, otherwise it won't work," continues Josh. "It's not about financial gain, it's about your passion for having an impact. If you didn't, you'd quit on day five or the first time you saw a CSAM image."



Over the past 17 years, WebPurify has come to understand the unique challenges of CSAM moderation. In order to identify this content accurately and remove it quickly, an organizational and technological approach has been developed to support CSAM moderators – and protecting their mental wellbeing is a key part of this.

Here, we shine a spotlight on the WebPurify CSAM team: who they are, how they work, and what they've achieved.

"This is why we started this company, and to find out we're responsible for over 500 arrests in the last year alone is so gratifying."

JOSH BUXBAUM

WEBPURIFY'S TEAM HAS IDENTIFIED OVER A MILLION PIECES OF CSAM CONTENT THAT LED TO OVER 500 ARRESTS IN 2022.

How WebPurify CSAM moderation *works*

1. POTENTIAL CSAM IS INITIALLY FLAGGED BY AI

AI has its limitations, but plays an important role in the WebPurify CSAM workflow – providing the first line of defense in identifying potential CSAM content. But at the moment, AI alone isn't sufficient for this work. First, the technology isn't accurate enough as of yet – it might miss some CSAM, and incorrectly identify some innocuous content. Second, the effects of missing that content or of wrongly referring someone's content to the National Center for Missing and Exploited Children (NCMEC) are severe. "Bottom line: the stakes are much higher with CSAM content," explains Josh, "and the margins for error next to none, so AI decisions cannot be solely relied on unless a submission is an exact match, right down to every single pixel, to a previously verified CSAM submission."

2. IT'S SENT FOR HUMAN REVIEW

The CSAM content moderators provide the critical second layer of defense. "We have a great internal tool called Montage that helps us label any content," explains CSAM moderator, Ahmed. "If it qualifies as CSAM then we label it accordingly. If not, we approve it. Any time we're unsure, we escalate the content to our subject matter expert for review." Furthermore, WebPurify's moderation tool has built-in functionality to limit the mental health risks of viewing this sensitive content: for example, Montage presents all imagery in black and white, which has been scientifically shown to reduce the psychological impact of disturbing imagery. It also converts videos into storyboards (a series of still frames from the video in a grid pattern) so moderators can quickly understand what's happening without having to hear audio or view the scene in motion. And, there's a blurring functionality that can be used to conceal aspects of the image that are particularly graphic. All of this is centered on the core idea that prevention tools are necessary when it comes to content moderation's impact on mental health. As Ahmed puts it, "we don't wait until the content affects us to do something about it."

3. SUBJECT MATTER EXPERTS (SMES) HANDLE EDGE CASES

"One of the priorities for our projects is ensuring content is moderated quickly. We don't want a moderator to get hung up on one image for too long and become a bottleneck, so when they encounter one that requires extra attention or is unclear in some way, they will escalate it to a Subject Matter Expert (SME). This person is highly trained for that specific project and has the time to research the image and ensure it is moderated properly. If a project is extremely high volume, it will have a dedicated SME, but in some cases SMEs cover a few different projects. When not occupied reviewing an escalation, they'll work with the quality control team, reviewing sample sets of the team's work for accuracy."

4. LABELED CONTENT IS REPORTED TO THE CLIENT

Moderators use various frameworks to evaluate what they're seeing. For example, the Dost test – developed by a California district court – is a key feature of child pornography law, adopted by virtually all state and lower federal courts as part of the definition of child pornography. Likewise, the Tanner score, which refers to the stages of physical development in a young person, and the industry classification scale from the National Center for Missing & Exploited Children (NCMEC) are also taken into account. Once the content is identified and labeled, it is sent via the client to NCMEC, who can then refer the content and associated account details to law enforcement.

5. REAL-WORLD IMPACTS ARE SHARED WITH MODERATORS

Sharing arrests stemming from the detection and reporting of CSAM is an essential part of WebPurify's mental health program. WebPurify's clients share data, so the impact can be measured. The successes are reported to the entire content moderation team, not only the CSAM moderators, so everyone can celebrate their accomplishments. In what can feel like isolated work, this sense of community support and recognition is a powerful tool. Regular reminders of the very real-world impacts the team's work is having make a difficult job that much more gratifying.

Who are **WebPurify's CSAM moderators?**

It's not easy work, but the CSAM team has low levels of turnover – something Josh attributes, in part, to who is selected for these positions.

There are many attributes a CSAM content moderator requires: strong communication skills, attention to detail, empathy and motivation to make a difference, emotional resilience, and an ability to compartmentalize. Not everyone is well-suited to the work or interested in pursuing it. So how does the team identify the best candidates?

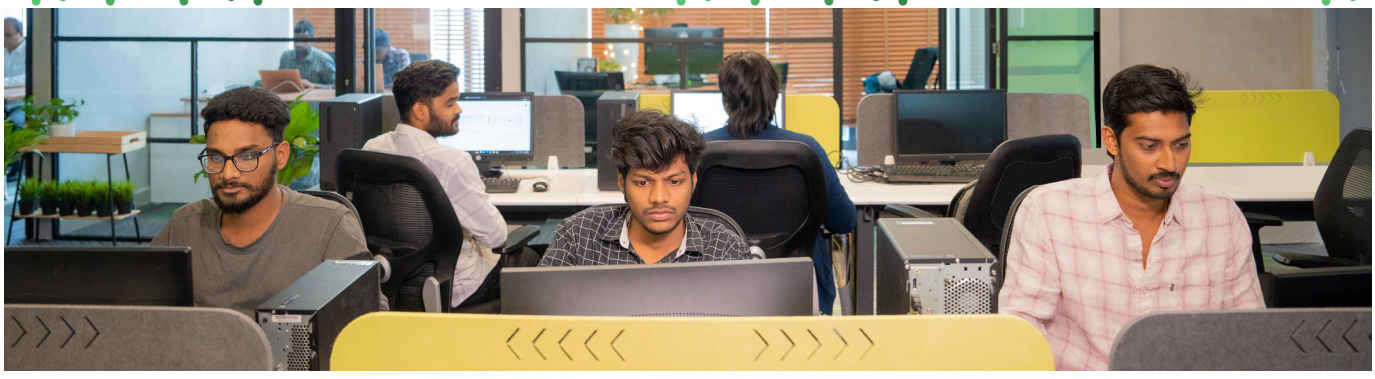
"For work of this nature, we tend to recruit from within," explains Josh. "It is sometimes difficult to determine if a new moderator will be fit for such work, even after an exhaustive interview process and close observation during training. If someone has worked with us for many years, we know their character, and we only offer these roles to those we know are ready for the complex nature of the work. They are typically interested due to the significant impact these positions have.

"That said, if an experienced moderator does not want to join this important team, we are completely understanding. We have plenty of other workflows for them to moderate. WebPurify's CSAM moderators stand out as the most resilient and experienced members of our team, willing to review some of the most disturbing content on the internet for the greater good."

Moderators must undergo extensive training on how to identify and report CSAM content, and this training typically includes both classroom instruction and hands-on practice. They also go through several mental health awareness programs.

"These moderators are a proven, experienced team. They've been in this career for many years and are likely to continue on their path – they're passionate about their work and the impact it has on the world. That's what makes it such a rewarding job, in spite of how tough it is.

"It's a really well-respected position," Josh continues. "There's this subtle understanding that the folks sitting in that team are having a really big impact."



CSAM *mental health* **risks**

That sense of purpose is key to maintaining the team's motivation, and is underpinned by the meaningful mental health support WebPurify provides. This team requires specialized support, and with a dedicated CSAM team it is delivered in a targeted manner.

"This work presents unique risks, and we realized we would need the proper tools to mitigate them," Josh explains. "We work with trauma specialist Duane Bowers to help mentally prepare new team members." Through a storied 30-year career, Duane T. Bowers, LPC, has become a trauma, dying, death and grief expert; respected international educator; and noted author. "He's worked with survivors of trauma and exploitation, with police officers and people who work at crime scenes. But what's different in this context is, you can't really train on it in the classroom. It would be illegal for us to store any kind of CSAM imagery for training purposes, so when new team members sit down on their first day to shadow current moderators, that's the first time they're actually seeing this stuff. No matter how prepared you are, your first exposure to this type of content is jarring. But over time it becomes easier to deal with as you utilize trauma reduction tools and techniques, and see the positive impact of your work."

"Your first exposure to this type of content is jarring. But over time it becomes easier to deal with as you... see the positive impact of your work."

JOSH BUXBAUM

A DEDICATED TEAM

"CSAM can come up in any project," explains Josh. "So even with companies inclined to think CSAM is a less likely threat given their industry or audience, it can pop up. Those cases can be shocking for customers, and traumatic for the company. That's part of why we decided to set up a dedicated team. That way we could have the highest impact, but also set up specific workflows and support systems to minimize that trauma."

SUPPORT SYSTEM

"There's always risk for someone who works with CSAM content for a prolonged period," says CSAM moderator Sandeep. "Traumatic stress, depression and anxiety are possible side effects, but we have a strong support system in place here, and constantly communicate with the rest of the team. We know that we're all in this together."

MANDATORY BREAKS

The structure of each team member's day is also designed to support their mental wellbeing, with mandatory breaks built in and dedicated spaces for resting and relaxing. Each day also wraps up with a quality check of the day's reviews.

"We usually work for two hours at a time, taking breaks," explains another member of the CSAM team, Srinivas. "We're actually trained to do this, and to take our minds completely off the content before we continue reviewing. We might go play pool or foosball, or perhaps board games like chess, or just watch entertaining videos." This system also helps maintain accuracy and focus.

Mental health *protection*

WebPurify provides trauma-informed training, plus access to mental health professionals and a mindfulness program with one-on-one coaching as part of their Employee Wellbeing and Assistance Program. The workplace culture encourages peer support, and the work day is structured to provide regular breaks and help the team disengage from the content. The tool is also designed to minimize psychological impact to employees. This hands-on approach to mental health support, combined with clear feedback on the team's positive impact, creates high levels of protection, commitment and retention in spite of the job's inherent risks. For more on WebPurify's approach, see our eBook on Best Practices for Mental Wellness in Content Moderation.