MODULATE

modulate.ai
contact@modulate.ai

The Importance of Mental Health for Content Moderators

Moderating online platforms is grueling work. Moderators are often underpaid and overworked, but even when they are not, their job description basically involves getting besieged with the worst content the internet has to offer, for hours at a time…and then getting up the next day to do it again.

It's easy to see that it's a moral necessity for any platform employing content moderators to support them through such difficult work; especially when it comes to protecting their mental health. But most platforms are at a bit of a loss here - larger platforms can at least afford to pay for mental health care, but beyond that, what exactly can they do?

Quite a lot, actually. As a company focused on reducing online toxicity, Modulate has had the opportunity to learn from best practices from many top online platforms as well as our own data labeling community. We've learned that while professional support is of course crucial, there are a number of soft techniques platforms can also employ to improve their moderators' wellbeing. So in honor of Mental Health Month, we wanted to take the time to share these lessons more widely.

1. Normalize taking a minute.

Depending on a moderator's job, all of the content they review might be extremely toxic; or perhaps only one in a hundred or a thousand pieces of content. Between this and individual differences, every moderator has different tolerance for getting exposed to toxicity, but nobody can go indefinitely engaging with this stuff without getting worn down.

It's important to remember that input to the brain affects output, and even if you can separate yourself from the content emotionally, a large amount of input can change your automatic thoughts. That's why it's important to give moderators regular breaks during a moderation shift. Our recommendation is to take a 15 minute break for every 2 hours of work by default.

At Modulate, we deal specifically with audio clips, which may include dangerous situations, admissions of current abuse, implied sexual assault, or anything else that may trigger past trauma or otherwise be harmful to hear. Research has shown that audio connects to emotions and memories more directly than text, this sort of content is especially crucial to moderate, but also especially harmful for moderators. Some moderators may start hearing clips in their head on repeat, feel queasy or sick, or have unresolved distress surrounding someone's well being. When something like this comes up, moderators must have the discretion to just take a minute to themselves to recover their equilibrium.

Remember, each moderator is a person, who has faced their own share of challenges, and each of them may be impacted by various types of toxicity in unique ways. Keep that in mind if you find yourself starting to question whether your moderator has really earned that brief break.

2. Give them space to understand and digest what they are seeing.

Most content that gets reported for moderation is itself narrow - a single audio clip, a snippet of text chat, or an image. In order to do their jobs well, moderators may often need to dig into the surrounding context…but even when it's obvious what moderation action should be taken, it's worth noting that your moderators may still deeply value the closure associated with understanding how such a toxic piece of content could have come about in the first place.

When reviewing so much toxic content, many moderators have reported challenges squaring the amount of hate they see with claims that people are "mostly good." But the reality is that most content violations come from generally well-intentioned people who are making a mistake, having a bad day, or misunderstanding the norms of the space they are in. Giving your moderators the space to confirm this for themselves can be a powerful opportunity for them to restore their faith in humanity and confidence in the work they are doing.

3. …and give them spaces to process things collectively, too.

As mentioned, it's easy for even the best moderators to be overwhelmed by the negativity and toxicity they witness and, frankly, lose faith in humanity. A simple measure that can help hugely with this is to give your moderation team a secure space for them to share particularly ridiculous, egregious, or upsetting content with each other.

That might sound counterintuitive - why would you let them signal-boost the very worst stuff? But the advantage is that this gives others an opportunity to chime in and say things like "you're right, that is horrible." And that can be overwhelmingly valuable to help your moderators retain a mentally healthy (and accurate) understanding of the world despite constantly witnessing such a skewed fraction of human behavior.

4. Remind moderators that they are the first line of defense, not the last one.

Especially in cases with players in danger or harassment, it can be easy for moderators to feel powerless - they see a situation where help is needed, but can't do anything to resolve it themselves. It's important to remind them that it takes multiple perspectives and training to help someone in any given situation. For example, let's say you go to the doctor's one day for some foot pain. As your primary care physician, the doctor would take a look at your foot and give you a treatment if it's something small; like a bad bruise, some muscle or joint pain, or an infected cut. But if it's a larger issue, they will refer you out to a specialist, and if it's extremely urgent, a hospital. For good reason, there's going to be a few different people who look at your foot and assist in the treatment; all the way from the physician's assistant to the care team on your floor at the hospital.

The same goes for moderators; if they see an instantaneous breach of conduct, such as a banned word or an insult, they can report it and take proper action. But if they suspect a case of severe bullying, child grooming, or other insidious harms, their job might center more on simply flagging the situation, rather than trying to resolve the whole situation. Indeed, there are a number of incredible organizations any platform and moderation team should be aware of who can help investigate these more complex harm types; Trust & Safety teams should make sure they are working well with third-party agencies like NCMEC (National Center for Missing and Exploited Children), NCADV (National Coalition Against Domestic Violence), AFSP (Americans For Suicide Prevention).

Hiring amazing moderators should never be the *end* of your trust & safety strategy - they are heroic contributors who deserve to be celebrated, but they can only provide that value if you give them the infrastructure and support they need to do so in a healthy and robust way.