This document reflects my experience and understanding of internal moderation practices at the time, based on the information available to me. It does not allege intent, motive, or wrongdoing by any individual.

1. What moderation currently looks like

As of today, the platform has around 125,000 reports waiting to be reviewed, with three active mods (two, now that I'm leaving). I joined the moderation team on December 14, 2024, a little over a year ago. During that time, there were several periods where I was the only active site moderator, with ticket mods helping me when they happened to receive reports on their side.

That workload is not realistic. Even if moderation were a full-time job, no one could keep up with it. Every report requires context checks, evidence, logging, and sometimes discussion. And when you're new, you need to ask questions. The problem is that very often, there was no one to answer them…

Because of that, some reports were never handled. Not because mods didn't care, but because there was no guidance, no framework, and no clear decision to make. Devs don't give us clear instructions. Some of them don't even seem to know their own site's rules.

For months, moderation had no real support, no communication, and no roadmap. Things improved slightly later, but before that, it was a complete mess. That's also why many mods left. Rules change constantly, sometimes overnight. One day we're told something must be banned. The next day, we're asked why we banned it and told we should have only removed it.

So we operate day by day, adapting on the fly, while everything that happens behind the scenes is invisible to users.

Mods end up in the middle of everything. When users complain, devs point at us. When rules are unfair, users blame us.

But we don't make the rules. We only apply them. We have no real decision-making power, yet we take the hits from every side. And that comes with abuse. A lot of it. Death threats, rape threats, doxxing threats, etc, daily.

I asked a former dev to make moderator accounts anonymous and to disable replies to moderation comments, because the harassment was constant. I would log in just to moderate, and the first thing I'd see would be insults or threats. That was often how a workday started.

Mods don't abuse power because we don't really have any. Devs give us a few tools and expect us to deal with the consequences. And at any moment, those same devs can change their minds, reverse decisions, or rewrite rules. We're expected to follow instantly.

Sometimes the rules change, and we learn about it at the same time users do. The difference is that we're still expected to act as if we knew all along, answer questions we don't have answers to, and pretend everything is under control.

At one point, mods raised concerns about the sustainability of the workload and the need for compensation for lead-level responsibilities. Instead of addressing the workload or compensation directly, the proposed solution shifted toward finding a "lower-cost external labor or automated moderation solutions" for those roles. The implication was that existing leads would be expected to train their own replacements without compensation.

This approach raised serious ethical concerns and further reinforced the idea that moderation labor was considered disposable rather than essential.

2. Moderation is mostly performative

What most people think of as moderation on the platform is, in reality, very superficial. We don't have the tools needed to properly moderate a platform like this.

We can't actually ban users. What we do is closer to a shadowban. A user can no longer publish bots or comments publicly, but their account still works. They can still log in, chat with bots, or create bots privately and use them.

This has serious consequences.

We've banned minors who still have access to the platform. We've banned users for creating CSAM content who can continue doing it privately.

When we "remove" a bot, it isn't deleted. It just becomes unpublishable. Everything moves to private, and moderation has no access to private bots.

That means anyone can create illegal or abusive content in private, and we literally can't see it, stop it, or intervene. We don't even know it exists.

On top of that, mods don't fully control their own actions. We can't unban users. If a mistake happens or if an appeal is legitimate, we have to ask a dev to step in.

So even when moderation wants to fix an error, there's no guarantee it ever will be fixed, unless the user manages to directly DM a dev. I'll come back to that later.

This is why I'm saying moderation is mostly performative.

From the outside, it looks like action is being taken. But in practice, enforcement stops at the public layer. We're given very limited tools, no access to critical parts of the platform to enforce anything in a meaningful way.

What exists is the appearance of moderation.

3. Hidden rules and opaque guidelines

There are rules users will never see.

Moderation operates with internal rules that are not written anywhere publicly. As mods, we sometimes have rules about what's allowed and what isn't. Users, on the other hand, have no way of knowing.

The public guidelines only cover what is legally safe to say, but many moderation actions are based on rules that simply don't appear anywhere. If you read the guidelines carefully and try to follow them in good faith, you can still end up breaking rules you didn't even know existed.

That's not a user problem. It's a transparency problem.

An example: Sexual abuse themes are allowed in bots. However, sexual abuse toward the user can't appear in the first message. For it to be considered CNC, the user has to be in a position to consent and to withdraw that consent. That means the first message can't already contain the abuse.

You will not find this rule written anywhere in the guidelines. Because it doesn't look good, legally, to spell this out clearly. Saying "this is allowed but only under specific conditions" is legally riskier than staying vague. From a legal standpoint, that protects the platform. From a user standpoint, it creates confusion and punishments without warning.

And that's just one example.

Another common one is bots based on real people or celebrities. Are they allowed? Yes and no. Even mods don't have a clear rule to rely on.

What we do know is that bots based on OnlyFans models and VTubers are removed. Not because of a clear ethical line, but because those creators regularly send DMCA requests, and it becomes a legal issue. For other public figures, there is no consistent answer.

We get a huge number of reports about this. Tags were removed to make this content harder to find. But there are still no clear instructions on how to handle it.

Mods are often accused of being inconsistent. The reality is that the rules themselves are inconsistent, incomplete, or deliberately vague. Some rules we learn on the fly. Some rules we never get a clear answer for. Some rules we know, but are explicitly told not to explain to users.

The result is a system where users are punished for content they had no reasonable way of knowing was against the rules. And mods are stuck enforcing policies they don't control, don't fully understand, and aren't allowed to clarify.

4. Favoritism and unequal rules

This is probably the part most people care about, so I want to be precise. Infractions are not handled the same way for everyone.

At first, the system was relatively simple. Everyone was treated the same, with one exception: verified creators. When a verified creator broke the rules, mods were required to contact them privately and ask them to fix the issue instead of applying a sanction immediately. Non-verified creators, on the other hand, would receive the penalty right away. That alone already created problems.

Those messages had to be sent from our personal Discord accounts. There was no anonymity, mods were directly exposed. In a recent case, a moderator followed the rules exactly as instructed and ended up being heavily harassed because their personal account was exposed. I won't go into names or details, but this wasn't a mistake, it was the expected outcome of the system.

There's something important to understand here: mods are required to stay silent. We aren't allowed to publicly explain moderation actions. We can't share evidence or clarify what actually happened. Moderation actions are considered private, and we are expected to protect user privacy at all costs.

That creates a very uneven situation. Users are free to say whatever they want publicly. Moderation is not allowed to respond. When mods stay silent, it doesn't mean anything other than we are not allowed to speak.

This is why you've seen situations where creators make public statements about moderation actions, and mods never reply. Not because we don't care or we agree. It's just because we're not allowed to. Meanwhile, mods are expected to absorb harassment, threats, and abuse while staying professional.

Now, about enforcement itself.

At some point, we were told there were too many bans. That was true, especially when I first joined. New directives followed: bans should be avoided whenever possible, except for extreme content or cases involving minors.

Moderation was later restricted from acting on external evidence in cases involving minors. Enforcement was limited to situations where users explicitly stated their age on the platform itself. As a result, mods were often forced to leave accounts active despite credible concerns, unless the user publicly self-identified as a minor on the platform.

One limitation was that creators above a certain follower threshold (at 666 followers, I kid you not) could no longer be banned directly. If you posted bannable content with 500 followers, moderation could act. If you posted the exact same content with 1k followers, moderation had to ask devs for permission. (That permission was never granted.)

So the result is simple: the same violation leads to different outcomes depending on visibility.

And, some of the rules we were asked to apply were vague to the point of being disturbing. For example, we were told to ban CSAM only if it was "very bad." That

raises an obvious question: when does sexualizing a minor become "very bad"? Personally, I believe it's serious the moment it exists, but that wasn't the directive.
Rules like this are impossible to apply consistently and put mods in an impossible position.

Now let's talk about verification.

Originally, verification was handled by mods. Content was reviewed, rules were checked, and once certain criteria were met, a creator could be verified. That changed.

Verification became a discretionary decision made by devs. Content didn't need to be reviewed. Numbers didn't matter. It became a matter of preference.

So here's the contradiction: if follower counts don't matter for verification, why do they suddenly matter for enforcement?
An infraction is still an infraction, regardless of who commits it and how many followers they have.

Favoritism exists, not because mods are friends with large creators, but because the system itself limits what mods are allowed to do. Our tools enforce inequality. We are physically blocked from applying the same rules to everyone.

There were also instances where content indicating potential real-world harm (mass shooting) was removed by mods and escalated internally for awareness. In several cases, no response or follow-up occurred, leaving mods without confirmation that the risks had been assessed.

5. Devs' intervention and overturned moderation decisions

Some users may remember what happened last year with the wave of bans related to bots based on Leland Coyle from Outlast. This character is canonically a former member of the KKK and is explicitly written as racist, antisemitic, and extremist.

I was the moderator who handled that situation.

It started when a creator I knew sent me one of these bots. When I reviewed it and saw explicit references to the KKK, it was immediately clear that this content couldn't stay on the platform. I discussed it with the rest of the moderation team, and everyone agreed. Regardless of the character being fictional, the KKK is a real-world extremist organization responsible for real violence and real deaths. This isn't fictional harm.

Because that first creator was known in the community and hadn't caused issues before, I chose to remove the bot rather than issue a ban, even though a ban would have been justified.

I then checked whether other bots of the same character existed and found around fifty. I informed the moderation team and asked if I could handle them. I was explicitly given authorization to do so. This type of content was considered bannable.

I reviewed every bot manually, and every action was logged. That's how moderation works internally. Actions are reviewable by the rest of the team. If a moderator removes content without justification or outside agreed-upon rules, they are removed from the team. Internal accountability does exist.

Out of those fifty bots, some were only removed. Others, those that explicitly included KKK references, racism, antisemitism, homophobia, or hate-based content, resulted in bans. Around eight accounts were banned, some of which were already inactive at that time.

That same evening, one of the banned creators contacted the devs about it. The devs then asked the moderation team what had happened. I presented the full logs and evidence. The creator denied having written racist or KKK-related content, despite the logs clearly showing otherwise.

The creator continued private conversations with the devs. Shortly after, a decision was made to unban all the accounts involved, including those who never appealed, never contacted anyone, and never disputed the decision.

Some justifications given were things like "I copied the description from Wikipedia" or "I didn't know what the KKK was." One of the creators involved was later found engaging in Nazi roleplay on other platforms.

This is important to state clearly: it takes very little time to understand what the KKK is. This isn't a fictional faction or abstract lore. Real people were murdered. But those accounts were still unbanned.

The entire moderation team agreed with the original actions. The reversals happened solely after private complaints.

This was not an isolated case. We've had minors unbanned despite clear evidence, simply because they contacted devs directly and denied being minors. Moderation then had to re-ban them afterward. Any time someone bypassed normal appeal channels and contacted devs directly, moderation decisions were reversed.

This creates a clear inequality. Users who follow normal appeal processes are often ignored since we don't have the tools to unban them. Users who know how to directly contact devs are far more likely to get decisions reversed.

In some cases, devs communicated one version of events to users and a different one to mods. Mods enforce rules, only to later be publicly undermined. This makes moderation look incompetent or malicious, even when actions were taken correctly. Public opinion of moderation is already hostile. When decisions are reversed without policy changes or internal clarification, mods become scapegoats.

I've personally moderated some of the most disturbing content on the platform, including extreme CSAM and necrophilia. In one case, I banned a rather large creator for one of the most graphic necrophilic bots I encountered. That ban was

overturned because the creator claimed they were "going through a hard time" and wanted to create "edgy content."

Decisions like this aren't based on consistent policy. They are based on convenience, emotional responses, and who complains loudly enough.

Mods are then expected to adapt, present these reversals as professional decisions, and absorb the backlash, despite having no control over the outcome.

6. The image policy and the payment processors

Around February or March last year, mods learned at the exact same time as users that NSFW images would no longer be allowed. We were explicitly told not to say that we learned it at the same time, even though it was obvious. I'll say it clearly now: we found out through the public announcement, just like everyone else.

So overnight, mods were expected to enforce a rule we had just discovered ourselves, and to answer questions we didn't have answers to.

The reason given was that NSFW images have to go so the site can receive payments. On paper, that makes sense. But almost a year later, where are the subscriptions?

The image filter has been adjusted multiple times. Mods have spent countless hours removing images. We've taken a huge amount of harassment for it. And yet, there have been no subscriptions, no clear updates, and no transparency.

I want to add something. I've seen people complain "you can't post porn on a porn site anymore." That comparison doesn't really hold. The platform is not a pornographic website. A porn site is a specific legal category with strict regulations, and that's precisely why those sites can work with payment processors. This site is an AI generation platform. Pornographic content can exist inside bots, but the site

itself isn't classified as a porn site. Complaining about porn images being removed here is closer to complaining about not being able to post porn on Facebook.

So for months, moderation enforced these image rules because we were told they were necessary. Then, recently, devs publicly said something very different.

After months of telling mods that NSFW images had to be removed, that the image guidelines were validated, devs publicly stated that if an image passes the automod filter, it's fine.

That means mods spent months enforcing rules we were told were mandatory, only for users to later be told publicly that those same images were actually acceptable all along, as long as they passed the filter.

Once again, mods end up looking abusive, arbitrary, or incompetent when in reality, we were following the instructions we were given.

The filter itself is an AI system. It's not reliable. Some explicit images pass, others don't. There's no clear logic a human can apply consistently. And yet, we were enforcing rules based on that system, without knowing the criteria ourselves.

The worst part is that mods weren't informed of this change. I personally learned that "if it passes the filter, it's okay" through a conversation with someone else. Not internally or officially.

So once again, users are told one thing, mods are told another, and we are left to deal with the confusion and the backlash.

This situation sums up the broader problem very well: poor communication, constant contradictions, and mods enforcing rules that change without warning, sometimes without even being told they've changed.

7. Why I'm leaving and speaking only now

I'm leaving moderation because I can't keep doing this.

For a long time, I believed in this project. I genuinely did. I loved working with the moderation team. They are good people, and I don't regret working alongside them for a single second. But I reached a point where I can't justify staying.

Moderation on the platform is unpaid. It's volunteer work. We're not employees. In my case, I'm disabled and can't work a traditional job. For me, moderation was my work. I treated it seriously. I spent close to full-time hours doing it for over a year because I cared and I believed it mattered.

If the site hasn't shut down despite the amount of illegal and abusive content that exists, it's because mods remove it. When people say "there's no child exploitation here," that's because mods actively delete it. You don't see it because someone is doing the work.

But the workload is overwhelming, the team is too small, and people burn out. We are exposed to abuse daily, with no protection, no clear structure, and no support. I've seen things every day that felt wrong. I kept telling myself it would improve, things would stabilize, and communication would get better. But it didn't.

There were many moments where I should have left earlier. When I was told to only ban CSAM "if it was very bad." When bans related to extremist content were overturned. When decisions kept changing without explanation.

I stayed because I wanted to believe this project could become something better. I don't believe that anymore.

This post isn't asking you to feel sorry for mods. It's not asking you to like us. It's not even asking you to agree with me. It's about transparency.

Most users never see what happens behind the scenes, but I did. And almost no one ever talks about it.

I'm speaking now because I'm leaving, and because once I'm gone, I have nothing to protect and nothing to gain.

Mods are not decision-makers. We don't act out of spite or personal issues. We try to enforce unclear rules that change constantly, often without being properly communicated. We act as a buffer between user frustration and devs' decisions, without any real protection.

If a moderator gets harassed, threatened, or doxxed, there is no legal or institutional support. I've lost count of how many times people wished me dead for removing bots. That was treated as normal. At one point, we were even told we couldn't remove bots that contained threats against mods.

In one case, a large creator made repeated violent threats toward mods. We asked devs to do something about it. Instead, devs joined that creator's server and politely asked them to stop.

That's the level of protection mods have.

Nothing in this post is meant to justify moderation decisions or excuse mistakes. It's meant to explain how the system actually works, because from the outside, you can't see it.

I'm done pretending everything is fine. That's why I'm leaving, and I'm saying this now.