Behind the For You

Content Moderators under TikTok's Control

Lúcio, anonymous content moderator

Suggested citation:
Lúcio. 2026. "Behind the For You - Content Moderators under TikTok's Control". ["Por trás da for you: O dia a dia do moderador de conteúdo sob o controle do TikTok"]. In: M. Miceli, A. Dinika, K. Kauffman, C. Salim Wagner, and L. Sachenbacher (eds.) *Data Workers' Inquiry.* Creative Commons BY 4.0. https://data-workers.org/lucio/

Introduction

My name is Lúcio, and I have been working in content moderation for about four years. It may seem like a short time, but in this universe — where everything moves at a different speed from the rest of the world, everything is due "yesterday," and the rules change from one moment to the next — this experience is equivalent to a lifetime. Based on my personal experience and numerous conversations with colleagues over the years, this piece offers an analysis of the ways in which managers and technology companies build a system to extract ever-increasing productivity from their employees at the expense of their health and well-being. The sensitive content we deal with has a profound impact on our health, and this report focuses on the additional challenges we face because of management strategies that prioritize numbers over human well-being: a relentless metrics system, constant competition among colleagues, and a lack of career prospects.

My goal in sharing this experience is twofold. First, to reveal what goes on behind the scenes of moderation at a multinational company. Second, to show those already familiar with the field that our experiences and struggles are rarely unique. I describe the supervision and management system but this account is, above all, about how this work transforms lives and shapes minds, exploring the often-ignored reality that exists behind the crucial work of ensuring a safe and reliable digital environment for the end user of social media. Through this account, it becomes clear how large digital platforms build systems that, while promising security to users, neglect the mental health and professional development of those who uphold that promise behind the scenes. This is a warning about the true human costs behind the curation of what you see—and what is kept from your eyes—on your *For You Page*.

PART ONE

Getting into the world of moderation

In 2022, I started working as a content moderator at Byte Dance Brazil, in the TikTok short videos area. Byte Dance is the company that owns TikTok, and I was hired as a full-time employee in Vila Olímpia, São Paulo, a neighborhood known as the city's technology hub and home to the largest finance, technology, and investment companies. For more than three years, I witnessed firsthand the workings of a Trust and Safety department, which is responsible for user safety and the quality of the content moderation in the platform. During this time the operation, which started small, grew exponentially.

Content moderation ensures that what is consumed by users of a platform complies with its rules of use and community guidelines. As a content moderator on social media, our role is to analyze media, photos, videos, and images created by users and posted within the platform. The importance of the moderator in this process is to ensure that what appears on your timeline, the For You page on TikTok, conforms to the community rules created by the platform, providing a safe environment for users. Some platforms have a similar process for comments.

Moderation always involves complex decisions that require balancing users' protection with their freedom to express themselves within the platform. From this description of content moderation for social media, it's easy to imagine that content moderation involves seeing everything bad on the internet. Even before I started, I had heard accounts that moderating social media was terrible, but I thought that my experience with community moderation, which involved experiencing all kinds of possible offenses, had prepared me a little. Still, I struggled to adapt. I had already learned to deal with prejudiced statements, insults, and racism against Black or Latinx people, but that didn't prepare me for the graphically sensitive content I experienced.

The content moderation at Byte Dance was structured into two main branches, or rounds, which are composed of smaller teams. Most of my time at Byte Dance, I was a Round 2 moderator, which means I dealt with content that was growing in engagement on the platform. Our team was responsible for moderating content with the potential to go viral, and we received everything that reached a certain number of views.

Usually, this content was part of a trend. There were a lot of dances, reaction videos, fan content for K-pop groups, New United and other young bands, and content selling access to sexual content platforms like OnlyFans and Privacy. Over the years, the company expanded our responsibilities to other products beyond short videos, including live streams and messages, images, and audios from chats. The content is very varied, but for all of us, the importance of the work is clear, especially to protect children. I saw firsthand the amount of disguised sexual content that could easily reach the recommendations page, as well as the frightening speed with which challenges or viral challenges spread that could cause injury or death. There were countless, but I particularly remember the incredibly dangerous "skull break challenge", where two people kicked the legs from under a third, making them fall over on their back. Content moderation plays an essential role in notifying and tracking these trends, as well as removing the horrifying and violent videos that arise as a result of them.

The responsibility we carry, and the disturbing nature of some of the content we evaluate, weigh heavily and leave deep scars on those who do this work. It is no coincidence that the moderation team has a high turnover rate and most of my colleagues are undergoing psychological or psychiatric treatment. But the biggest blow is, and always has been, for colleagues who are Round 1 moderators. They dealt with content considered sensitive, most of which was reported by users. Here we are talking about content that was *really* sensitive, involving dead bodies, brutal violence, mutilated people, and the worst types of abuse, including child sexual abuse. The psychological consequences of working with sensitive and toxic content are not the focus of this article. Still, it is important to make it clear that all the management practices documented here are implemented in the context of extremely demanding work that is potentially harmful to the mental and physical health of workers.

The content has always had a huge impact on the mental health of moderators. For those who join, it is a startling shock. After a few months at the company, we all notice emotional desensitization, as the content becomes just another video, or yet another photo. The negative impact on mental health is enormous[1]. But this piece goes beyond the content, focusing on the moderators' routine and management's strategies to constantly increase productivity by creating an exhausting schedule to achieve ever-higher goals, fostering dangerous levels of stress. My goal is to show that the challenges moderators face come not only from the content but also from a toxic environment that sees human beings and the workforce only as numbers.

I worked at Byte Dance for approximately three years, and today I am at a business process outsourcing company (BPO), moderating content for another social network, and I see much of the same treatment. I recognize the same patterns as at TikTok, which shows these experiences are not something specific to one company, but rather to an industry that exploits workers. In addition to the content, there is intense pressure to meet metrics, creating competition among employees to produce more and more.

[1] Inquiries developed by other workers have already explored the psychological impacts of such work in depth, particularly the report by Fasica Gebrekidan https://data-workers.org/fasica and the proposal for mental health intervention by Kauna Malgwi https://data-workers.org/kauna/

PART TWO

Metrics and more metrics

From my first days, it was clear that the routine at the company was very strict and metrics — or key performance indicators (KPIs) in the language of business managers — were essential. The moderators' lives revolved around three metrics: moderation time, speed, and quality.

“Moderation time” refers to the time we are effectively logged into the system reviewing content. The moderator has to deliver six hours of moderation per day, which might seem easy, but in reality is very challenging considering all the activities that are part of the job. All teams have a series of meetings to update and clarify policies that do not count towards this time, even though they are mandatory and essential. Bathroom breaks or stretching your legs didn't count towards this time, nor did the mandatory well-being sessions. In other words, achieving this target requires a lot of daily coordination, and unexpected meetings or delays because you have to go to the company's physical office disrupt plans and make meeting this delivery metric more complex.

The second metric is the average time the moderator takes per task, in our case per video. The target to be achieved was one video every ten seconds. Which is difficult in itself, considering that most content lasts at least 30 seconds. But the greater challenge was maintaining a good speed while simultaneously keeping the quality average at 97%, meaning making the correct decision in the vast majority of cases.

“Quality” is measured by a dedicated quality team that re-moderated a small number of cases. If there is a discrepancy between the moderators' decision and the quality team's decision, either in the appropriate action to be taken or in the categorization of which policy the content violates, it is recorded as an error by the moderator and lowers their average. The system is structured very unfavorably for the moderators, as the quality team has time to evaluate each content carefully, while the moderators work under pressure.

Furthermore, the quality team only re-moderates a small number of cases, which makes this metric very unpredictable. If I did around seven thousand cases during the week, the quality team re-moderated around 20 random ones, and my quality metric was based solely on these. If you have more cases being reviewed by quality, you have a higher chance of having errors, which lowers your average. However, for a person who has fewer cases reviewed, the impact of each one on the average is much greater, generating a distorted result that can be much higher or lower than your actual accuracy rate.

The Quality Lottery

Since there is no fixed formula or percentage determining how many cases per moderator the quality team checks, the quality metric is always unpredictable.

The most extreme example is when only one of your cases is reviewed by quality. If it's correct, your score that week is 100%; if it's wrong, it's 0%. A score like that tanks your rating for the entire month, and it's virtually impossible to get back on target.

You could be very lucky, deliver fewer cases and get high quality scores. If in one week I delivered 3,000 cases while my colleagues delivered 7,000, they would likely have more cases re-reviewed, and I could come out ahead in the process.

But in reality, the score is not a number that truly expresses how much that moderator actually got right or wrong: it's a game of luck, with no margin for error.

The moderators' lives revolve around these performance metrics for a very simple reason: low scores almost always mean dismissal the very next month, and high scores guarantee bonus payments. The three metrics are the basis for the monthly performance scores, which in turn are reviewed twice a year, almost like a school report card at the end of the semester. The performance reviews occurred in August and March, and although our lives were governed by numbers, the process wasn't very clear. What we knew was that low scores almost always meant dismissal, and a series of dismissals always happened right after the reviews. High scores guaranteed generous performance bonuses, reaching up to three months' salary for the top performers, and could even lead to a promotion. It's easy to imagine the stress for the moderation teams during the review period.

PART THREE

Conflicts with quality

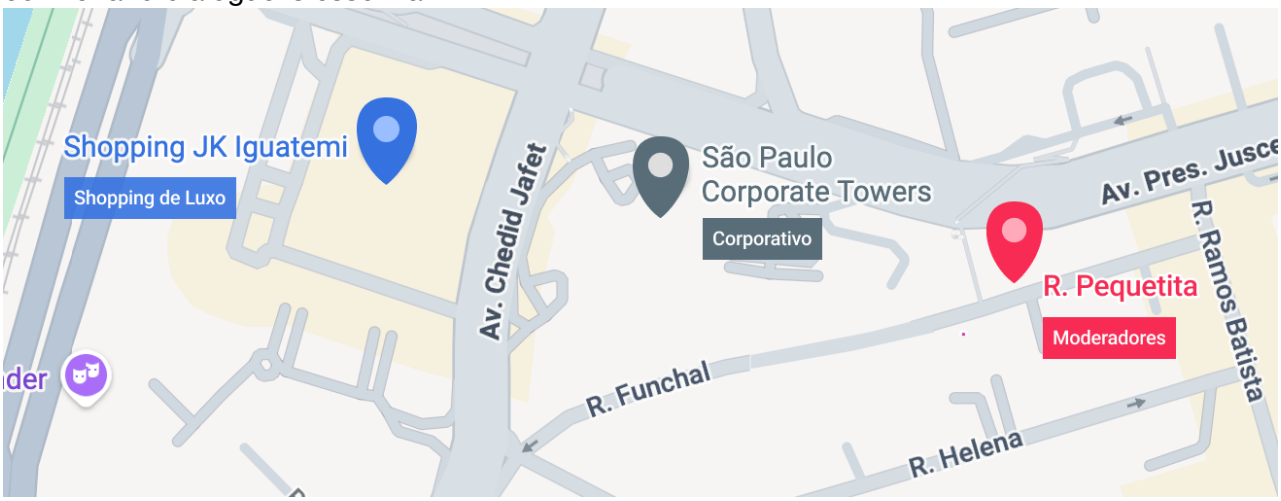
When I started, I spent the first few months moderating a lot, trying to improve, dedicating myself to studying the policies — which change constantly — and working with focus. However, it has always been extremely difficult to meet all the metrics, and I soon realized I wasn't the only one, and that the metrics system wasn't designed with our success in mind. All of us had weeks where the number of cases reviewed by the quality department was low, and this score became completely unpredictable. There was constant friction and conflict between the moderation and quality teams over the quality metric.

The coworking space where the moderation teams were located had rotating desks and was never sufficient for everyone.

While moderators are held accountable for daily case volume, time, and accuracy during the moderation process, the quality team is held accountable for ensuring the processes are being carried out and followed as stipulated. This creates friction between the two groups, who are pursuing different things. The quality team's performance is affected only by their own work, while the moderation team's results are affected by the work of this other department. It wasn't an individual conflict, but there were constant clashes that needed to be discussed but weren't.

There is a conflict between the moderation and quality areas, as each chases its own goals. But both are essential to maintaining user safety on the platform. The only reason for the constant friction was the company's decision to separate the quality and moderation processes, as if each had a different objective, forgetting that moderation's outcomes depend on quality. This is also reflected in the physical separation, as the people performing quality control were not even on the

same floor as us. This makes everyone forget we're all in the same boat; it's not a war or a conflict and dialogue is essential.



The buildings where management and moderators were located were physically separated.

During my years at the company, it was clear that in any situation where moderators had less opportunity for growth and to explain their perspective to the quality team, there was a significant difference in scores, worse performance, and more pressure on both sides. In contrast, in teams where moderation worked in conjunction with quality, that's where we saw a significant improvement: the work environment was calmer and lighter, if you had doubts you knew where to go, and there was more alignment.

My experience makes it clear that this unnecessary separation between areas greatly worsens the moderators' lives. ByteDance management chooses to keep the processes this way and, tellingly, allocate the trust and safety department to a separate building from the administrative and corporate departments.

São Paulo Corporate Towers - the fancy building where the offices of management are.

PART FOUR

Competition

As mentioned, the evaluation of our metrics directly affected our compensation, and the difficulties caused by the structuring of the quality processes were just one part of the problem. There is also an opaque system of monthly ratings, where the performance of an individual moderator is compared to their team and the entire department. In other words, that unpredictable quality score, along with the moderation time and speed metrics, are compared to the results of all moderators. During the dreaded performance reviews, each person receives a grade ranging from A to D, which is an aggregate of all the semester's metrics and averages.

To better understand the situation we were in, and how fierce competition was, it's necessary to understand that there is only a limited number of each grade that can be distributed during this evaluation. For example, in a team of 20 people, there could be a maximum of two A's, four B's, five C's, and the rest D's. So even when my quality average was 97% and I'd delivered all my production hours and met every requirement that did not guarantee I'd receive a good A or B grade.

When the entire team or the vast majority were meeting these metrics, it was no longer enough: the TARGET for achievement rose based on other people's numbers. If someone scored 100% on everything, you had to have that score too, and outperform them, to get the A grade. The official target is no longer what you need to deliver to keep your job; it becomes the absolute bare minimum to avoid a negative grade. And remember, all of this influenced the annual bonus, potential promotions, and salary increases.

There was a very clear, if unspoken, policy that you cannot know anything about your colleagues, and you also could not ask. No one knew the next person's salary, their salary grade (salary promotion), or the monthly ratings within the team. It was even more frustrating, since you were compared to the team without knowing the actual numbers of the top performers. Without knowing how to improve, who is the best, or what they do for that. It was a blind competition where, even if you were exceeding the metrics, you couldn't know how to succeed. There was a moment when everyone was delivering above the metrics, and there was no way to get a better grade because the target was no longer the baseline goal, but the basic expectation.

As moderators, we often didn't know our own metrics, let alone those of our colleagues, or how leadership aggregated each of these indicators to establish the comparison. Competition among moderators thrives in secrecy and silence, with terrible consequences for our mental health, generating additional stress and anxiety.

Clearly, the only one who benefits from this management structure is the company, which can extract more productivity from its employees at the cost of their health and well-being, while also undermining the goal of a trust and safety department that can carefully assess risks. There is a management problem with leaders who are not trained to lead, and quality processes separated from moderation instead of aligned to improve platform safety. This excessive control over productivity, which extends to requiring that bathroom or water breaks be logged, is extremely unnecessary, dehumanizing, and only serves micro-management.

The Dynamics of Competition

Let's use as an example a team of moderators who don't moderate short videos, as was my case for a time. Each of them has the following averages in a month.

Remember that the established targets are: 6 hours of production, 97% quality, and an average of 10 seconds per case.

PERSON A
Moderates chats
Daily production time: 6h
Pace: 9 sec/case
Quality: 98%
Cases per week: 9,000

PERSON B
Moderates chats
Daily production time: 6.5h
Pace: 9 sec/case
Quality: 96.5%
Cases per week: 9,200

PERSON C
Moderates live streams
Daily production time: 6h
Pace: 25 sec/case
Quality: 97%
Cases per week: 7,000

In this fictional example, moderators A and B have a clear advantage because they are handling shorter and simpler content. Person C, who handles much longer live streams, knows they are at a disadvantage: their pace is much higher than the metric and even though they deliver equally good quality, they will have a lower grade.

Still, the incentive structure likely motivates person B to work overtime and produce more. They all work on the same team and have these metrics analyzed as a whole. In any case, none of them know about the others' results, which drives constant competition.

Such a system is not humane and it does not work. The consequences are high employee turnover, terrible mental health in the moderation team, falling quality and bad corporate health. Poor management makes the entire operational system sick and for those already exhausted, more stress means psychological destruction and mental health deterioration.

PART FIVE

Dead end

In content moderation the pressure to meet and exceed metrics is huge. Evaluating each piece of content according to a large set of rules and criteria demands constant attention, and dealing with sensitive content also takes its toll: emotionally, physically, and psychologically. And it must be made absolutely clear that we are talking about an entry-level position, a role where you would—theoretically—develop into new areas. However, at ByteDance, there was no established, structured career plan for the moderation area. Not even close!

The most natural progression would be to advance to a quality analyst position. But in 2021, when there was a massive expansion of the quality team, the company preferred to hire more quality professionals from the market instead of increasing internal promotions. In other words, those who were already moderators waiting for openings in quality— which would be the direct promotion—ended up losing opportunities because the positions were either not made available internally or were offered in insignificant numbers.

And it doesn't matter how many years of moderation experience you have; in every company you join, moderation will be the entry-level position, and the lack of growth is the rule. You need to specialize in other functions to get a promotion. Your work alone is not enough. Beyond your job, you must know how to handle people management, processes, and development. More is demanded from those already inside than from external hires.

Throughout my entire time as a TIkTok moderator, there was no prospect of growth or advancement. Many excellent, capable colleagues entered as moderators and left as moderators. There was no solid, established career plan, so hitting targets was just a way to avoid being fired. There was no development plan for everyone to grow within the company, for all that moderation knowledge and interaction with the network and users to benefit the moderators and prepare them to engage with other areas of trust and safety or within the company.

Increasingly, social media responsibility has become a necessary topic in public debates and inside social media companies. Content moderation is a fundamental part of that process, however those who are moderating have little participation in this debate. We are merely numbers: the content must be analyzed and that's all that matters.

AI as saviour?

Today, companies are training AI to replace human content moderators. TikTok has already started layoffs in moderation, resulting in mass dismissals in Brazil[2], Malaysia, Germany[3], and around the world[4]. The problem is that there are always malicious users testing the platform to try to promote themselves and only moderators understand the context, the platform's dynamics, and where users are testing the limits of the policies. It is common and routine for moderators to identify that a user is trying, or even managing, to bypass the rules and they can start developing action plans.

AI, in contrast, will never be able to identify new tactics for bypassing safeties and publishing sensitive or dangerous content in time to prevent it from reaching users. AI is good for handling the volume of cases to be moderated, not for the quality of those cases. The machine learning process depends on the availability of a large amount of data. In other words, human moderation will always be indispensable to keep the AI updated. The new ways of circumventing the system must first be identified by or reported by users to the responsible trust and safety departments before an action plan can be created to collect and moderate the necessary data to train the AI again.

Without the moderator — someone who adds context, culture, history, and diversity to the process — the AI will just be another heap of 0s and 1s repeating what it was taught before. It adds no new information, no diversity, and thus fails to protect users.

[2] TikTok demite centenas, enquanto amplia uso de IA para moderação de conteúdo". October 11, 2024. Época Negocios. Fabiana Rolfini. https://epocanegocios.globo.com/empresas/noticia/2024/10/tiktok-demite-centenas-enquanto-amplia-uso-de-ia-para-moderacao-de-conteudo.ghtml
"Em ano eleitoral, TikTok faz demissões no Brasil em sua equipe de moderação e segurança". March 19, 2024. Estadão. Henrique Sampaio. https://www.estadao.com.br/link/empresas/ano-eleitoral-tiktok-demissoes-seguranca-brasil-nprei/?srsltid=AfmBOoomwgyZFpd3dreAttPk1SwHxsgUXa_ihBc1qgSKzJ2siNEX4qFT

[3] "TikTok: Gemeinsam gegen Massenentlassungen". November 18, 2025. Ver.di. https://www.verdi.de/themen/arbeit/++co++19d30bc8-622c-11f0-937c-b715127af524

[4] "ByteDance's TikTok cuts hundreds of jobs in shift towards AI content moderation. October 11, 2024. Reuters. Rozanna Latiff. https://www.reuters.com/technology/bytedance-cuts-over-700-jobs-malaysia-shift-towards-ai-moderation-sources-say-2024-10-11/
"TikTok to lay off hundreds of UK moderators as it shifts to AI". Aug 22. 2025. Financial Times. Anna Gross and Tim Bradshaw. https://www.ft.com/content/d277c456-4518-4994-8afd-c45a399db342

Conclusion

My experience as a content moderator reveals a fundamental contradiction at the heart of major digital platforms: the same industry that boasts of creating safe and connected spaces builds its foundations on the systematic exploitation of invisible workers. This account demonstrates that the challenges of moderation go far beyond graphic and disturbing content. They are rooted in a business model that treats human beings as disposable resources.

At the end of the day, everyone is just doing their job, trying to pay the month's bills, and the company takes advantage of this state of need among workers to extract more and more from every possible role. The rivalry between areas and teams fosters competitiveness and increases productivity, but at the employees' expense. This piece shows how large tech companies use metrics — in this case, moderation time, speed, and quality — not only to monitor the proper functioning of the Trust and Safety systems, but as tools to increase productivity without raising salaries or hiring more staff.

The artificial separation between moderation and quality exemplifies how companies fragment processes that should be collaborative, creating unnecessary conflicts that harm both workers and the safety system as a whole. Simultaneously, the forced competition for limited grades generates a toxic environment of rivalry and anxiety. This is not simply inefficient management, but a deliberate architecture designed to maximize productivity at any cost, even if that cost is the mental health of the workers.

Today, working at another tech company, I understand this is not unique to ByteDance. Part of the shattered expectations that workers experience regarding working conditions is tied to it being a well-known name, a *big tech* company. But the reality I found is structural; at this new company, the same problems exist.

The lack of interest in bringing moderators into the center of policy debates, the lack of paths to specialize and leverage the knowledge moderators have about content, and, in general, bleak career prospects, condemn talented moderators to professional stagnation, wasting valuable knowledge that could enrich other areas of trust and safety.

All of this is a system. Where you're being employed directly by a company or contracted through a third party makes no difference. It's not about tech companies per se, it's about a system that uses and exploits our labor and our physical and mental health, without wanting us to be anything more than a mass of workers. Beyond the control that extends even to our most basic breaks, the situation is worsened by the imposed silence: salaries, metrics, or conditions cannot be discussed, creating a culture of isolation that benefits only the employer. This testimony serves as an urgent warning: we can no longer ignore the conditions of the workers who uphold our digital experience. Content moderation is fundamental to online safety, but it must be recognized as specialized work that deserves dignified conditions, career prospects, and adequate protections against the inherent psychological harms.