This is a submission from British tech justice non-profit Foxglove.[1] Foxglove works to make technology fair for everyone. We challenge the misuse of technology by powerful corporations and states.

Summary:

For over five years, Foxglove has worked extensively with social media content moderators. As part of this work, we have launched world-first litigation against social media companies in multiple jurisdictions, including challenging Facebook's software design, and content moderation failures, which fanned the flames of hate and violence during the Tigrayan conflict in Ethiopia.

During this time, Foxglove has gained a thorough understanding of this work and attempted to raise public awareness – as well as communicate to elected leaders – about the extent to which the most powerful social media platforms are **utterly reliant** upon human content moderation to function at all.

If we were to deliver one key message for members of the Committee to take away from this submission it would be this: social media content moderation is the single most vital, under-resourced and under-regulated job in the modern digital economy. **If the Committee truly wants to understand the spread of hate and violence online, it is crucial Committee members understand the work, conditions and, as it stands, the exploitation of the skilled human workers whose job it is to police it.**

Background:

Social media platforms now dictate the public square across the world. Just look at the events of the last week: posts on X from Elon Musk have dominated the UK's domestic political discourse and, whatever one's opinion of them, dictated the parliamentary agenda, provoking a response from government and the Opposition.

Musk's posts are a microcosm for other social media platforms' business models. The algorithms that curate user experience on platforms like X, Facebook and Instagram recommend content to maximize engagement. Their machine-learning models favour controversy, misinformation, and extremism - content that provokes a response, the most sensational and extreme content is promoted into users' feeds. Just as Musk's posts efficiently seized the attention of Westminster this week, so do recommender systems bombard social media users with toxic and sensational content to keep us clicking.

Why? Social media companies generate most of their income from advertising. The sums involved are huge, in some cases substantially larger than the GDP of dozens of countries. In 2023, Meta reported an advertising income of $131.9 billion[15] (somewhere in the ballpark of Slovakia's GDP). TikTok's ad revenue in the US alone is expected to reach $12.3 billion for 2024[16] and although X's revenue from advertising is falling, advertising still accounts for approximately 90% of its income.[17]

This advertising-based revenue model means profit is directly related to user engagement. Platforms deploy the tactics described above to keep users engaged - algorithms designed to maintain users' attention at all costs, to keep them online, and to keep advertising targeted. Put simply, more eyeballs on more feeds means more clicks – and more clicks means more ad money. The non-stop prioritisation of extreme content can have a normalising effect on user response. As Facebook whistleblower Frances Haugen said: "*the algorithms take people who have very mainstream interests and push them towards extreme interests [...] You see a normalisation of hate, a normalisation of dehumanising others, and that is what leads to violent incidents*".

That is how the business model of social media companies translates into violence on our streets. How do companies like Meta square the essential truth that their platforms are designed to amplify hate speech and violence with their public duty not to cause harm? In theory: content moderation. **As The Verge's Nilay Patel once said: "The essential truth of every social network is that *the* product is content moderation."[4]**

<u>The work of a Content Moderator:</u>

It is worth setting out what the work of content moderation actually means. The job of a moderator is to review content uploaded by users, whether photo, video or text, and then assess if it is acceptable to a platform's policies. If it is acceptable, it can be left up. If not, if it breaks the rules, it needs to be taken down. The content moderators review can be truly horrific: beheadings, child pornography, torture, suicide and the dismembering of humans and animals. It is the job of human moderators to review content that is too complicated or gruesome for platforms' automated tools to be able to assess.[6] **In other words, moderating traumatising content is not one unfortunate aspect of the work: it is the core function**.

That means human content moderation will have to remain the critical component of social media's safety supply chain for at least the foreseeable future. Without moderators, this nightmarish content would be live on Facebook, YouTube, Instagram, X and the rest. The social media companies' theory is: content moderation *is supposed to* allow social media companies to thread the needle of using their recommender systems to push the most sensational, shocking and graphic content as widely as possible, while ensuring anything actively dangerous is removed. In practice, content moderators are trying to put out a wildfire with a water pistol.

The number of content moderators employed for this work is woefully inadequate, for the scale of content that must be reviewed every day. For example – according to previous comments by Meta, it employs about 15,000 content moderators globally, with most of those based in the United States. Every day, about 350 million photos are uploaded to Facebook, as are around 1 billion stories. Asking 15,000 moderators – even if they were properly resourced – to keep up with that scale is impossible.

The scaling problem is much worse in Global Majority regions where the fewest moderators are employed. For example, until the beginning of 2023, Facebook's content moderation hub for East and Southern Africa was in Nairobi, Kenya. It employed 260 content moderators to cover a region with a population that is conservatively home to around 500 million people. The Nairobi hub was shut down in early 2023, after workers there attempted to form a trade union. The closure and mass lay-off of workers is the subject of an ongoing court case at the Employment and Labour Relations Court in Nairobi supported by Foxglove.

It's not just Facebook. When Elon Musk took over Twitter, now X, in 2022 he fired 80% of the Trust and Safety (content moderation) team[7]. That meant there were simply not enough workers to keep the platform safe in August 2024. It is impossible to know the true extent of social media's under-resourcing because the companies are not required to publish the number of moderators they employ – and X certainly isn't telling. But the violence in August demonstrates the consequences of the systematic national and global under-resourcing of content moderation. **Until social media companies are forced to properly resource human content moderation social media platforms cannot, and will not, be safe to use.**

<u>Working conditions:</u>

The woefully inadequate numbers of content moderators to cover the scale of content being uploaded daily to social media is only the first critical flaw in social media's business model when it comes to public safety.

Psychiatric damage: Repeated exposure to violence and other toxic content without psychiatric care makes content moderators seriously ill. 144 **Facebook** moderators underwent psychiatric assessments as part of one of Foxglove's legal cases in Kenya. The medical evidence revealed that **every single moderator** assessed was diagnosed with severe post-traumatic stress disorder caused by exposure to graphic social media content including murders, suicides, child sexual abuse and terrorism[8]. The doctor who carried out the assessments, the Head of Mental Health Services at Kenyatta National Hospital, the largest hospital in Kenya and the East Africa region, said that in his professional opinion the primary cause of PTSD in all 144 cases was the work of Facebook content moderation. There is not a social media company we know of – anywhere, including in the UK – that provides sufficient psychiatric care to protect moderators' mental health.

Outsourcing: Social media companies routinely outsource content moderation - they do not conduct their core product of their business in house. This choice allows platforms to play down the importance of content moderation as low-paid, low-status work, in comparison to their directly employed white-collar workers. There are also pay disparities between moderators in global minority vs global majority countries. In 2022, TIME reported that Facebook content moderators in Kenya were paid around $2.20 per hour. In the US, moderators were paid around $15-18 for the same work.

Outsourcing is also the social media companies' attempt to shift responsibility for moderators' working conditions. In court cases supported by Foxglove where content moderators have challenged their poor working conditions social media companies have argued they cannot be held accountable by the courts of countries where they have caused harm, because the work is outsourced.[10] We believe many of the content moderators for UK content are employed by outsourcing companies in Ireland.

Union busting and the culture of fear: As described above, when content moderators in Kenya organised for a fairer workplace, Meta sacked them all – the entire content moderation workforce for East and Southern Africa[11]. When workers spoke to parliamentarians about the terrible conditions in their workplace in Germany, Meta, and outsourcing company Telus, sacked the spokesperson who gave evidence[12]. Elon Musk fired dozens of X employees when they dared to criticise him publicly.[13]

Alongside that, content moderators often describe how their jobs were originally advertised under false pretenses – as 'IT administrators', 'language experts' or 'admin assistants'. Many content moderators only learned the nature of the job – and that they would be working for some of the biggest and most famous companies on the planet – when they began work. Moderators are routinely forced to sign highly restrictive NDAs which ban them from discussing any aspect of their work with anyone, including their families and loved ones. Not only does this exacerbate the mental toll of the work which, as we have seen, routinely causes PTSD, but also has a chilling effect on any attempts to organise in their workplaces for better conditions.

Final thoughts:

Despite social media companies' claims of high standards for removing hate speech, their platforms are swimming in extreme content that incites violence. That is by design. Platform's recommender systems are organized for profit. They are not organized for our communities, the people they serve or with an interest in protecting public safety.

Corporate secrecy means we can't know exactly how content is curated – other than relying on first-person testimony of content moderators and other whistleblowers, like Frances Haugen. But we can connect the dots and listen to their evidence to understand that these platforms are full of toxic content by design – because that is how their business model generates the huge engagement that underscores their advertising revenue. That business model actively encourages the spread of harmful content which, in turn, contributes to broader societal harms.

Social media companies made scaling up user numbers their singular priority, then woefully underinvested in human content moderation, despite it being **critical** to platform safety. They chose not to moderate content properly, or to implement their own community standards, by design. The moderators they do have work in dangerous conditions that routinely expose them to life-threatening illnesses including PTSD. It may in fact be practically impossible to moderate Facebook safely at the scale of users it has reached. If so, Foxglove contends that is an argument that they have reached a scale that is incompatible with public safety. When companies accumulate such a monopoly they cannot be kept safe, the only responsible policy decision is to break them up.

Last summer's violence serves as a stark reminder of the power of social media platforms and the urgent need for meaningful reform to curb the spread of disinformation and protect public safety. At this point, it is painfully obvious that social media companies will not act to get their house in order on safety unless they are forced to by courageous national governments, working in the public interest to protect the safety of their citizens.