Comment of Electronic Frontier Foundation to Oversight Board case 2022-007-IG-MR

Government involvement in content moderation raises serious human rights concerns in every context. Those concerns are very troubling when the involvement originates with law enforcement. They are even more troubling when law enforcement targets artistic expression, an area well beyond their expertise. And they are especially troubling when law enforcement advances concerns for gang activity, concerns disproportionately aimed at youth and communities of color.

Government Involvement in Content Moderation Raises Serious Human Rights Concerns

As revised in December 2021, the Santa Clara Principles on Transparency and Accountability in Content Moderation specifically scrutinize "State Involvement in Content Moderation". As set forth in the Principles:

"Companies should recognise the particular risks to users' rights that result from state involvement in content moderation processes. This includes a state's involvement in the development and enforcement of the company's rules and policies, either to comply with local law or serve other state interests. Special concerns are raised by demands and requests from state actors (including government bodies, regulatory authorities, law enforcement agencies and courts) for the removal of content or the suspension of accounts."

The Santa Clara Principles also include pertinent principles for state actors, stating that "state actors must not exploit or manipulate companies' content moderation systems to censor dissenters, political opponents, social movements, or any person."

These concerns are even more dire given that the EU's Digital Services Act will soon require that Meta employ "trusted flaggers," entities given priority status to "flag" content for platforms. Some have questioned[i] their effectiveness. Nonetheless, under the DSA, trusted flaggers are designated by governmental agencies, and may include law enforcement agencies such as Europol. Although trusted flaggers are supposed to only flag illegal content, the preamble of the DSA explains that its rules should not prevent the providers of online platforms from making use of trusted flaggers to act against content incompatible with their terms of service. This opens the door to law enforcement overreach and platforms' over-reliance on law enforcement capacities for the purpose of content moderation. These and other concerns have also been articulated[ii] by the DSA Human Rights Alliance,[iii] a group of digital and human rights advocacy organizations representing diverse communities across the globe.

The origins of the human rights concerns are twofold:

First, preferred flagger status gives government entities outsized influence to manipulate content moderation systems for their own political goals—to control public dialogue, suppress dissent, silence political opponents, or blunt social movements. As an example, consider community standards addressing misinformation: it is far too easy for a government to flag all criticism of it as "fake news."

And once such systems are set up, it is easy for government—and particularly law enforcement—to use the systems to coerce and pressure platforms to moderate speech they may not otherwise have chosen to moderate. This leaves platforms inherently biased in favor of the government's favored positions.

Second, governmental entities are often not worthy of the trust the platforms' preferred status gives them. For example, each EU member state assigns different rights and responsibilities to its police force. The abuse of this system could enable in countries like Poland and Hungary, which have been on the record for their anti-human rights views, must not be taken lightly. Hungary, for instance, has already been found guilty[iv] by the European Court of Human Rights of violating the right to respect for private life of an Iranian transgender man; similarly, Poland is ranked[v] as the lowest in the EU when it comes to respect of human rights and, in particular, those of the LGBTQ+ community. The Israeli Cyber Unit has boasted of high compliance rates, up to 90 percent across all social media platforms with its takedown requests.[vi] But human rights organizations have expressed concerns that these requests unfairly target Palestinian rights activists, news organizations, and civil society.[vii] One such incident prompted this Oversight Board to recommended that Facebook "Formalize a transparent process on how it receives and responds to all government requests for content removal, and ensure that they are included in transparency reporting."[viii] Vietnam has also boasted of its increasing effectiveness in getting Facebook posts removed, but has been accused of targeting dissidents in doing so.[ix]

Government entities may also commonly simply lack relevant expertise. Under the DSA, for example, once a trusted flagger is so recognized, they are "trusted" to flag content on a variety of platform types, regardless of the limitations of their expertise.

Drill music is an excellent example of how this system fails. Meta has recognized the need to consider both freedom of artistic expression and law enforcement concerns. But police aren't experts on music and have a history of linking it to violence; a claim unsupported by recent research.[x] As such, the flags raised by the police come to Meta completely one-sided, rather than with experts supporting both sides. Moreover, this particular issue also raises concerns far beyond law enforcement and freedom of artistic expression. An informed decision will also consider a variety of sociological factors.

<u>The Music of Inner City Black Communities is Disproportionately Overtargeted by Police</u>

Originating from the streets of Chicago, drill music is a creative output of inner-city Black youths. And whilst anti-establishment narratives are communicated and the videos often depict content of a violent nature, it is a vital mouthpiece for voices that are seldom elevated into the realms of mainstream engagement—from the streets of south Chicago to south London.

The takedown requests pertinent to this case are part of a larger from London's police force—the Metropolitan Police, or the Met—to remove drill music from online platforms, based on the mistaken, and frankly racist, belief that it is not creative expression at all, but biography. The Met started an "enhanced partnership" with streaming platform YouTube in 2018 which has since facilitated a pervasive and punitive system of content moderation for drill rappers in London. The partnership advanced previous efforts by the Met to surveil drill music, most

notably since 2015 when the force launched Operation Domain[xi] to monitor "videos that incite violence" on YouTube. In June 2019, Operation Domain was replaced by Project Alpha which involves police officers from gang units operating a database of 34 different categories,[xii] including drill music videos, and monitoring social media sites for intelligence about criminal activity. According to a 2022 report by Vice,[xiii] 1,006 rap videos have been included on the database since 2020. The BBC has reported[xiv] that since November 2016, the Met made 579 referrals for the removal of "potentially harmful content" from social media platforms and 522 of these were removed, predominantly from YouTube.

Project Alpha contravenes data protection, privacy, and freedom of expression rights. A heavily redacted document[xv] notes that the project was to carry out the "systematic monitoring or profiling on a large scale, or in a public place," with males aged between 15 to 21 the primary focus. This is in line with the Met's previous gang-focused activities, such as the 'Gangs Matrix,'[xvi] where racial profiling and breaches of data protection laws were implemented to give the force a foundation to surveil innocent people and disproportionately target Black men and children.[xvii]

The operationalization of the Met's systemic racism into the realm of trusted flaggers is furthermore stifling artistic freedom. The Met has refuted accusations that Project Alpha suppresses freedom of expression. But this is in direct conflict with the force's actions and own narratives, with former Met Police Commissioner Cressida Dick blaming drill videos[xviii] for the surge in violent crime and murders across London, and Detective Inspector Kieran McAuliffe affirming[xix] that "It's fine to use it [drill] as an outlet" but "Just think about what you're saying." Such framing of drill music is crude and fundamentally disregards the genre's intrinsic role in providing a means of expression for communities forced to the periphery of mainstream political and social vernaculars. It has also contributed to self-censorship whereby popular YouTube channels have advised artists to censor content that could be deemed offensive to avoid potential removal once the video goes live.

The Met's police officers also provide evidence in U.K. court rooms for the prosecution, despite seldom understanding the socio-political culture that the music was produced within. In one trial,[xx] the prosecution introduced drill lyrics as evidence and claimed that they either "describe the life [the defendant] already leads or they describe the life he aspires to lead." Drill lyrics and music videos are not simple or immediate confessions to engagements in criminal activity, yet the Met Police continue to affirm the opposite and in doing so exacerbate the over-policing and racial stigmatization[xxi] of Black children and youth across urban areas in London. This "street illiteracy"[xxii] has further exacerbated conceptualizations of drill music as an illustration of real-life activities that the artists have themselves seen or done, rather than an artistic expression communicated through culturally-specific language and references that police officers are seldom equipped to decode or understand. A 2021 report by JUSTICE[xxiii] called the "misuse of drill music to secure convictions" as "one of the most profound examples" of systemic racism in the UK, and the prolific requests by the Met to YouTube for video removal highlights this further.

Similar trends are also evident in countries like the United States where New York City mayor Eric Adams recently blamed drill[xxiv] for violent crime in the city and called for the removal of

drill videos from social media, echoing similar panics about hard rock[xxv] and rap music[xxvi] in the United States in past decades.

Transparency Reporting Should Provide Details About Government Moderation Requests

The Santa Clara Principles includes several recommendations for reporting on government involvement in content moderation, each of which a company with the resources of Meta should be able to comply. We urge the Oversight Board to recommend that Meta implement them.

"Users should know when a state actor has requested or participated in any actioning on their content or account. Users should also know if the company believes that the actioning was required by relevant law. . . . But companies should clearly report to users when there is any state involvement in the enforcement of the company's rules and policies.

Specifically, users should be able to access:

- Details of any rules or policies, whether applying globally or in certain jurisdictions, which seek to reflect requirements of local laws.
- Details of any formal or informal working relationships and/or agreements the company has with state actors when it comes to flagging content or accounts or any other action taken by the company.
- Details of the process by which content or accounts flagged by state actors are assessed, whether on the basis of the company's rules or policies or local laws.
- Details of state requests to action posts and accounts."

The companies should thus disclose any back-channel arrangements they have with government actors, including trusted or other preferred flagger systems, and reveal the specific government actors to whom such privileges and access are granted.

The Principles further provide the following under the Numbers principle:

"Special reporting requirements apply to decisions made with the involvement of state actors, which should be broken down by country:

- The number of demands or requests made by state actors for content or accounts to be actioned
- The identity of the state actor for each request
- Whether the content was flagged by a court order/judge or other type of state actor
- The number of demands or requests made by state actors that were actioned and the number of demands or requests that did not result in actioning.
- Whether the basis of each flag was an alleged breach of the company's rules and policies (and, if so, which rules or policies) or of local law (and, if so, which provisions of local law), or both.
- Whether the actions taken against content were on the basis of a violation of the company's rules and policies or a violation of local law."

Moreover, Meta should notify individuals when their speech is moderated at the request of a government actor.

"Specific information about the involvement of a state actor in flagging or ordering actioning. Content flagged by state actors should be identified as such, and the specific state actor identified, unless prohibited by law. Where the content is alleged to be in violation of local law, as opposed to the company's rules or policies, the users should be informed of the relevant provision of local law."

---

[i] https://scholarlycommons.law.wlu.edu/cgi/viewcontent.cgi?article=4565&context=wlulr
[ii] https://www.eff.org/deeplinks/2021/10/new-global-alliance-calls-european-parliament-make-digital-services-act-model-set
[iii] https://www.eff.org/pages/dsa-human-rights-alliance
[iv] https://www.amnesty.org/en/location/europe-and-central-asia/hungary/report-hungary/
[v] https://www.ilga-europe.org/rainboweurope/2021;
[vi] https://www.gov.il/BlobFolder/generalpage/files-general/he/files_report-2018.pdf
[vii] https://www.hrw.org/news/2021/10/08/israel/palestine-facebook-censors-discussion-rights-issues; https://7amleh.org/2021/05/21/7amleh-issues-report-documenting-the-attacks-on-palestinian-digital-rights; https://prospect.org/world/how-secretive-cyber-unit-censors-palestinians/
[viii] https://www.oversightboard.com/decision/FB-P93JPX02
[ix] https://thediplomat.com/2020/07/facebook-vietnams-fickle-partner-in-crime/
[x] https://www.independent.co.uk/news/science/violent-music-death-metal-songs-happy-pharrell-williams-a8819551.html
[xi] https://www.met.police.uk/foi-ai/metropolitan-police/d/march-2022/information-about-operation-domain/
[xii] https://bylinetimes.com/2021/10/25/tracking-without-transparency-met-police-expands-social-media-surveillance-operations/
[xiii] https://www.vice.com/en/article/bvnp8v/met-police-youtube-drill-music-removal
[xiv] https://www.bbc.com/news/uk-55617706
[xv] https://www.theguardian.com/uk-news/2022/jun/03/met-police-project-alpha-profiling-children-documents-show/. These documents were sourced through a Freedom of Information Act request by Point Source. https://pointsourceinvestigations.wordpress.com/
[xvi] https://twitter.com/ICOnews/status/1063371272162369536
[xvii] https://www.theguardian.com/uk-news/2022/jun/03/met-police-project-alpha-profiling-children-documents-show/
[xviii] https://ico.org.uk/media/action-weve-taken/decision-notices/2021/2619961/ic-58919-t8h2.pdf
[xix] https://www.bbc.com/news/uk-55617706
[xx] https://www.bbc.com/news/uk-55617706
[xxi] https://www.gov.uk/government/news/to-end-racial-disparity-we-require-your-absolute-focus
[xxii] https://academic.oup.com/bjc/advance-article-abstract/doi/10.1093/bjc/azz086/5706791?redirectedFrom=fulltext&login=false
[xxiii] https://yjlc.uk/resources/legal-updates/justice-report-report-finds-misunderstanding-drill-music-leading-unfair
[xxiv] https://www.thefader.com/2022/02/11/new-york-city-mayor-eric-adams-drill
[xxv] https://www.vice.com/en/article/r3za83/satanic-panic-interviews
[xxvi] https://www.npr.org/transcripts/11361420