August 16, 2022

***Via Email***

Facebook Oversight Board
Rachel Wolbers rwolbers@osbstaff.com
Simona Sikimic ssikimic@osbstaff.com

**RE: Request for Comment on UK drill music (2022-007-IG-MR)**

The American Civil Liberties Union and Daphne Keller of Stanford University's Cyber Policy Center welcome the opportunity to provide comments on case 2022-007-IG-MR regarding the takedown, at the behest of a UK law enforcement agency, of an Instagram clip from a new drill music video by rapper Chinx (OS).

The law enforcement agency claimed that the post, which referenced a past shooting, could provoke further violence. An internal team at Meta reviewed the post and took it down for violating the company's Violence and Incitement policy. It also removed copies of the video posted by other users. It is unclear whether Meta indiscriminately deployed technical filters, or had agents review the other posts to assess the context in which the video appeared – for example, whether an individual posted it as part of criticism of the clip's music or message. Meta notified the user who created the post when the content was removed, but did not inform the user that the removals were initiated by a request from UK law enforcement. The case summary does not describe what notifications, if any, the other affected users received.

Government-initiated removals—especially those that rely entirely on private content policies to take down lawful content—are a danger to free expression. Around the world, law enforcement bodies known as Internet Referral Units (or IRUs) are asking platforms like Instagram to delete posts, videos, photos, and comments posted by their users. Platforms are complying, citing their own discretionary terms of service as the basis for their actions. Users are not being informed of governments' involvement. Courts have little or no role.

## Removal of Protected and Political Speech

IRU flags result in the removal of lawful speech. Some IRUs have requested removal of online material ranging from Grateful Dead recordings to scholarly articles. And many of the requests target speech that platforms do not agree violate their terms or applicable law. Not surprisingly, IRUs also often flag content that is critical of government actors.

In the Israeli case of *Adalah v. Cyber Unit*, for example, the platforms thought Israel's IRU was wrong roughly 20,000 times in a single year, presumably indicating they believed the posts were not only legal speech, but also permissible under their privately-determined community standards. At the same time, platforms also honored tens of thousands of Israeli

government requests, each identifying tens or hundreds of individual posts for removal. It is impossible to know for sure what speech was affected, but Palestinian human rights advocates assert that the likely victims were racial, religious, and cultural minorities engaged in protected speech and criticism of government policy.

In the context of U.S. law, political speech "has always rested on the highest rung of the hierarchy of First Amendment values, and is entitled to special protection" *Carey v. Brown*, 447 U.S. 455, 467 (1980), "to assure unfettered interchange of ideas for the bringing about of political and social changes desired by the people." *Roth v. United States*, 354 U.S. 476, 484 (1957). Even in countries with different legal protections, the practical importance of dissent is just as high. Yet, in practical terms, the IRU system gives government actors more tools to mis-label and take down speech that criticizes government policy or actors as "hate speech," "threats," "incitement," and the like.

**Disparate Impact on Marginalized Groups' Free Expression and Access to Information**

In addition, the cooperation between IRUs and platforms is likely to have a disproportionate effect on communities of color and their freedom of expression. This case highlights that problem, with its focus on drill music, which is often created and performed by Black artists. UK police report monitoring over 2000 drill music videos, and requesting removal of many.

Bias in the enforcement of rules makes this disparate impact even more likely. For example, prosecutors in the U.S. more frequently rely on rap and drill music as criminal evidence than rock and country music. Descriptions of violence, the importance of maintaining territorial boundaries, the tragedy of lives lost or taken, and even first-person narratives of engaging in criminal activity may be viewed as entertaining or artful storytelling in the latter genres—Johnny Cash's murderous vagabond persona who "shot a man in Reno just to watch him die," or Dolly Parton's jilted lover who stabs a victim in the heart "by the banks of the Ohio." But when they appear in drill music, Chinx (OS)'s genre, officials more often interpret them as incriminating confessions or threats.

This bias has played out repeatedly in the U.S. court system. Rather than recognize the art of drill, courts have permitted states to present lyrics and music videos in criminal prosecutions, arguing that they constitute evidence of criminal behavior. In November of 2020 in Tennessee, the state showed the jury a rap video featuring the defendant, an aspiring Knoxville rapper, claiming that that the artist's sometimes violent and graphic imagery was a confession to a murder, despite the fact that the videos were recorded months before the murder and make no mention of the victim. There are a number of other examples.

If courts are getting this wrong, so too must law enforcement agencies that request takedowns, as well as platforms that consider the requests. Moderators making discretionary decisions about drill musicians are likely to mistake lawful but violent lyrics for incitement to violence. Research consistently demonstrates patterns of bias in which white subjects misperceive Black people as angry and hostile, including in situations where the Black person is actually experiencing fear. Broader studies on implicit association also show that individuals

making rapid assessments of new information are likely to rely on stereotypes based on race, gender, and other traits, disproportionately associating members of minority groups with negative attributes. The working conditions of Meta's content moderators in many ways replicate the test conditions of these studies. Both test subjects and Meta's moderators are tasked with viewing content on a screen and making rapid, potentially racialized, decisions about whether the content is, for example, threatening or violent. While we are aware of no robust research about such unintended bias in Meta's moderation – and indeed, such research may be impossible given the current lack of transparency – there is every reason to expect that even well-intentioned moderators, particularly those without cultural context, may reach biased outcomes in assessing content such as drill music videos.

This problem is exacerbated when law enforcement is trusted to flag controversial content. Law enforcement has little incentive to take the public value inherent in artistic expression into account when making decisions about what content to target.

**Applicable Law**

Advocates in Europe and elsewhere have consistently raised concerns about IRUs targeting lawful speech, and in particular speech by members of marginalized groups. But judicial review of IRUs, or of platforms' compliance with their requests, has been rare. The leading court case addressing IRUs to date is, as mentioned above, the Israeli Supreme Court's *Adalah* ruling. It appears to have been in part impeded by the lack of records retained by the IRU regarding the case, Meta's failure to notify affected users about the IRU's role, and by an overly-narrow consideration of whether the IRU's requests were coercive. Discussion of that ruling can be found in a Lawfare post by Israeli lawyers Tomer Shadmy and Yuval Shany, and in a later post by Daphne Keller.

There can be no doubt that government pressure over the past decade drove platforms' increasingly restrictive content policies. State pressure is perhaps most obvious in examples such as the EU's Hate Speech Code of Conduct, under which Meta and other platforms agreed to enforce EU law-based rules against "the promotion of incitement to violence and hateful conduct" in their Community Guidelines – despite the fact that speech prohibited in the EU may be legal in other parts of the world. But Meta and other platforms' current speech rules are also the product of other iterative, informal, or unpublicized concessions made over the years in response to pressure from powerful state actors, including in the UK.

This history matters for the protection of users' rights, regardless of whether the more compliant Meta of today is happy to honor state demands it might previously have resisted. It is particularly relevant in cases, such as this one, in which governments are "only" asking platforms to remove content under nominally voluntary Community Standards, rather than asserting authority to seek removal under national law. Deploying state resources like the London Met's IRU to suppress lawful speech would be highly questionable in any situation. But it is doubly so when speech rules applied by the platform are themselves the product of state action.

The question of coercion is further complicated by platforms' complex interactions with state actors, including in realms that are not obviously related to speech. Saying no to

governments may not be worth the potential costs: upticks in critical attention from police or prosecutors, public tongue-lashings in legislative hearings, regulatory backlash (new content laws, or laws that just so happen to hurt platform interests in areas like tax, competition or privacy), and even arrests (as have happened in Brazil and India), service blockages (Turkey, Russia), or seizures of assets or moneys owed (as recently authorized in Austria).

In the context of such potential coercion, law enforcement flagging content for private entities to censor raises serious concerns. Our comments below largely focus on U.S. First Amendment law, as our core area of expertise. But various forms of protections against prior restraint and the suppression of speech without adequate legal process can be found in legal systems around the world, including under the European Convention on Human Rights. These protections are particularly robust in the Inter American Human rights system. A thorough report by the respected UK human rights lawyer Lord Ken Macdonald suggests that similar constraints exist there. If police were to bring notices of allegedly *unlawful* content directly to platforms, he points out, the "inevitable corollary" would be that "judicial process in advance of the police issuing a takedown notice would need to take place."

U.S. law addresses analogous questions in the seminal 1963 *Bantam Books* case. There, the U.S. Supreme Court reviewed the practices of a government commission that, though it did not (indeed, could not) "apply formal legal sanctions," "deliberately set about to achieve the suppression of publications deemed 'objectionable' [by the government] and succeeded in its aim." 372 U. S. 66-67. The Commission regularly sent intimidating notices to book distributors identifying potentially unlawful literature, which led the distributors to withdraw those books from circulation. The court held this behavior violated the First Amendment.

Coercion can be accomplished through something less than an explicit threat of penalty. As mentioned above, there is a history of repercussions, and bad relationships with a national government is not something a platform can lightly ignore. As Facebook's global head of policy has described, companies are "eager to predict regulation . . . so they can adjust their policies to keep up with the times and thereby avoid risk to their business." Monika Bickert, "Defining the Boundaries of Free Speech on Social Media," in *The Free Speech Century*, ed. Geoffrey R. Stone and Lee C. Bollinger (Oxford, UK: Oxford University Press, 2018), 267.

**Recommendations**

Meta's policies for takedowns when flagged by IRUs should, to the extent possible, safeguard lawful expression, avoid racial and other bias, and resist censorship of speech critical of government.

To that end, Meta should:

1. ensure that IRU complaints as well as any removals of duplicate content are escalated to culturally-competent people, trained, vetted, and constrained by rules of procedure;

2. give accused users, including users who post duplicate content, notice not only that their content was removed, but also that it was flagged by an IRU or other government entity, which entity, and the justification given;

3. give users the ability to appeal any determination;

4. publish transparency reports that give the public information about how many IRU flags the company received, what the TOS violation claim was, whether there was a claim that the content was illegal, whether the company moderators agree, and what actions the company took. To the extent possible under applicable law and consistent with users' privacy rights, Meta should retain and disclose actual copies of flagged as well as removed content, to enable independent assessment of IRU and Meta's actions;

5. indicate whether automated content detection tools were updated to search for and take down the same content on the platform or when reposted;

6. indicate whether any takedowns in response to IRU complaints are global or regional, and use region-specific takedowns where possible. For example, content which may incite violence in a particular state will not pose similar dangers when viewed in a different country;

7. weigh factors such as whether the public has an interest in knowing what a public-official or other prominent speaker is saying; whether the flagged speech is political, including criticism of the government or government actors or policy; and whether the flagged speech was inappropriately targeted or misunderstood as a result of the genre, content, or identity of the poster as part of the decision whether or not to delete or otherwise moderate those publications;

8. refuse special "fast track," "trusted flagger," or other privileges to IRUs unless they take appropriate measures to protect free speech and other rights, including:

    a. Maintaining a high degree of accuracy in their notices;

    b. Avoiding bad faith or improper notices, including notices designed to protect the reputations of individuals in government;

    c. Maintaining adequate internal records, so that any incidents or patterns of bias and error are auditable and available as evidence in any future *Adalah*-like cases;

    d. Publishing IRU transparency reports listing the number of requests made to platforms, broken down by the basis for the request, the platform, the result of the request and, whether any effort was made to directly investigate or initiate criminal proceedings relating to the allegedly harmful content at issue. Data about enforcement efforts will be important to future policy discussions about the

relative roles and capacities of platforms or law enforcement in deterring real-world violence and other harms.

**Conclusion**

Thank you for the opportunity to submit these comments on such an important issue.