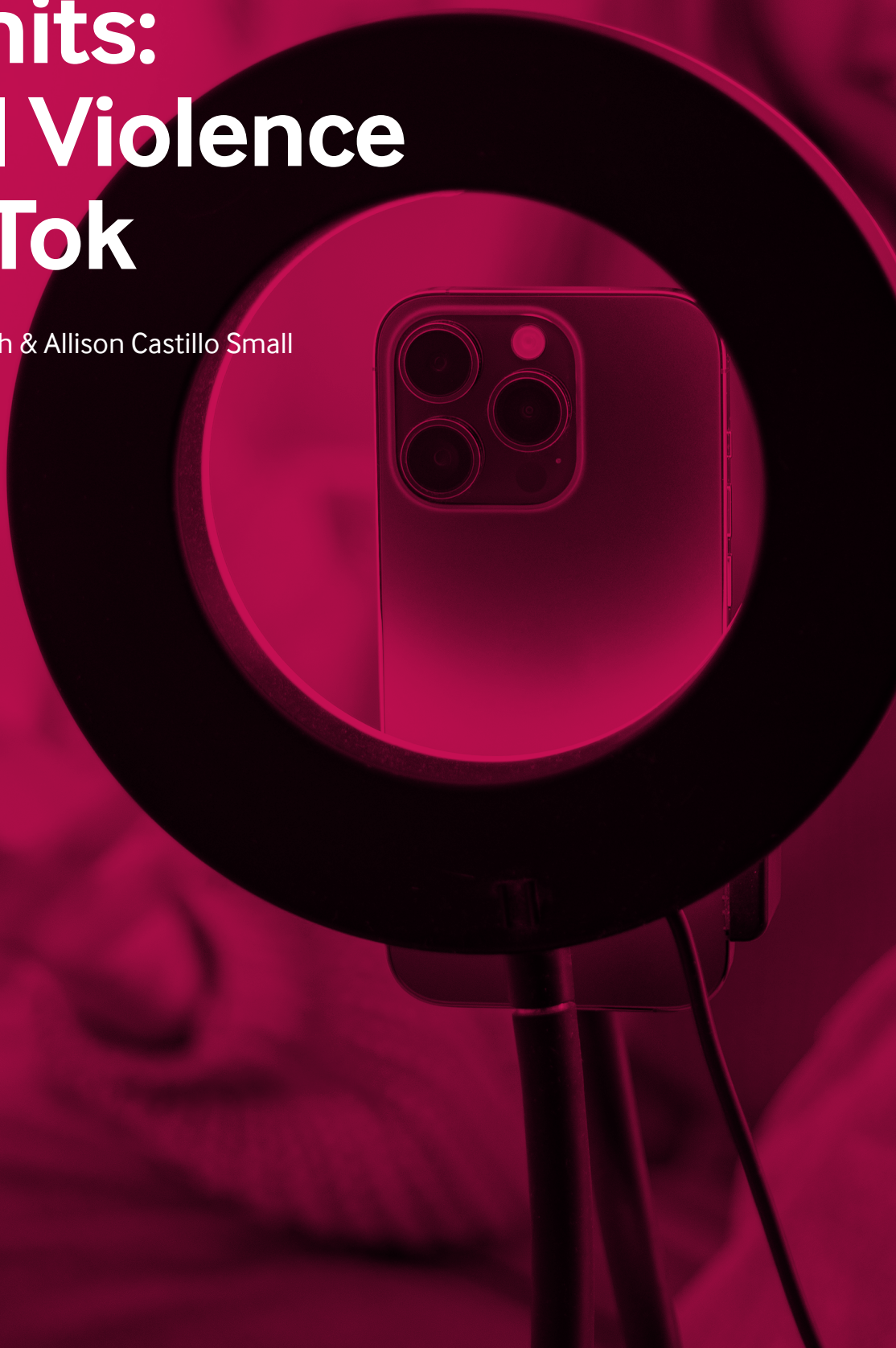
ISD | Powering solutions to extremism, hate and disinformation

# Off-limits: Sexual Violence on TikTok

Paula-Charlotte Matlach & Allison Castillo Small



ISD | Powering solutions to extremism, hate and disinformation

Amman | Berlin | London | Paris | Washington DC



**www.isdglobal.org**

# Off-limits: Sexual Violence on TikTok

TikTok's efforts to remove shocking and graphic content – including sexual and physical abuse of both children and adults – remains inconsistent and might be inadvertently marginalising victims and survivors of sexual violence. ISD has examined the platform's content moderation practices on sexual violence in English, French, German and Hungarian. On the one hand, ISD analysts found significant gaps in enforcement across languages, locations and devices. On the other hand, TikTok's efforts to uphold their own Community Guidelines and stop the dissemination of violent or sexualized content may have also affected legitimate discussions and information concerning gender-based violence – for example, by blocking terms associated with sexual violence completely. TikTok's guidelines state that "reproductive health and sex education content" are explicitly allowed.

However, it appears that the enforcement of these guidelines is inconsistent, which could result in a disproportionate negative impact on women and the LGBTQ+ community. In conclusion, TikTok's moderation strategy on sexual violence is failing to strike a balance between moderating violent content and ensuring freedom of expression as the platform is obliged to under article 34 of the DSA. In conclusion, two contradictory observations were made: TikTok's moderation policies simultaneously risk the restriction of legitimate conversations about sexual violence and insufficient consideration of sexual violence keywords.

This briefing is part of a series examining online gender-based violence (OGBV) on TikTok in English, German, French and Hungarian. It is part of the project Monitoring Online Gender Based Violence Around the European Parliament Election 2024, funded by the German Federal Foreign Office.

# Key findings

In May 2024, TikTok updated its Community Guidelines, including a detailed section on their consideration of sexual content and sexual education[1]. In the "Sensitive and Mature Themes" section, the platform asserts that while nudity, sexually suggestive content involving minors, and graphic violence are restricted, discussions about sexuality, reproductive health and sex education are permissible. They also state that the application of these guidelines may differ across regions, considering the importance of local context and culture. Although TikTok claims to maintain a baseline of internationally recognised human rights in its localisation efforts, prior research by ISD has shown that more transparency on TikTok's content moderation decisions, as well as a broader linguistic, cultural and geographic scope are generally needed.

When it comes to moderation strategy, TikTok employs a combination of automated technology and human moderators to detect and manage harmful content. Machine learning algorithms analyse various signals from the videos, including keywords, images, titles, descriptions, audio, and metadata, to identify potential violations. The platform states that when a potential violation is found, the automated moderation technology can immediately remove content or pass it on to a moderation team, which conducts further reviews to ensure context and nuance are considered appropriately.

Despite these self-proclaimed efforts**, ISD has found crucial inconsistencies in TikTok's moderation strategy on sexual content.** This briefing evidences several gaps in TikTok's keyword filter for searches on sexual violence. **The inconsistencies found lead to two contradictory outcomes: Potential restriction of legitimate conversations about sexual violence on one hand on the one hand and insufficient consideration of sexual violence keywords on the other.**

# Definitions and methodology

## Definitions

For the purposes of this briefing, ISD utilises the following definitions:

**Gender**
Gender refers to a "system of symbolic meaning that creates social hierarchies based on perceived associations with masculine and feminine characteristics." A person's gender identity refers to "an individual's internal, innate sense of their own gender."

**Gender-based violence (GBV)**
This term refers to "violence directed against a person because of that person's gender or violence that affects persons of a particular gender disproportionately." Women and the LGBTQ+ community, including transgender and gender-diverse persons, experience disproportionate rates of GBV.

**Online gender-based violence (OGBV)**
OGBV is defined here as a subset of technology-facilitated gender-based violence (TFGBV), which refers to any "act that is committed, assisted, aggravated, or amplified by the use of information communication technologies or other digital tools, that results in or is likely to result in physical, sexual, psychological, social, political, or economic harm, or other infringements of rights and freedoms." For a more detailed review and discussions of terms and definitions please refer to ISD's report "Misogynistic Pathways to Radicalisation".

## Methodology

This briefing is based on manually collected data from TikTok in English, French, German and Hungarian. The qualitative analysis was carried out by a team of analysts who conducted and compared search results for the keyword "rape".

| Keyword | Language |
|---|---|
| Rape | English |
| Vergewaltigung | German |
| Viol, le viol | French |
| Nemi erőszak, megerőszakolni | Hungarian |

Table 1: The keyword "rape" and translations to German, French and Hungarian.

To gather the data, analysts used a systematic approach by which each term was used as a search term in the TikTok browser version, switching VPN locations between France, Germany, Hungary and the UK. For each search term and location, it was recorded whether the term was blocked or produced search results. If terms were blocked, the type of notice given was also recorded.

To standardise results as much as possible, searches were conducted through an incognito browser window without a TikTok account using a laptop located within the European Union (EU). Analysts compared search results across the four languages to identify differences in moderation practices across geographies.

# Inconsistencies in moderation of sexually violent content across geographies and languages on TikTok

Analysts found that TikTok's moderation practices are inconsistent across different languages and locations, both in the application of search filters and in the support offered to users who encounter sexually violent content on the platform (table 1).

On the TikTok desktop version, searches for the term "rape" in English triggered supportive messages and links to resources for survivors in the UK and Germany but produced results without any formal notice in France and Hungary. In comparison, the TikTok Android app showed supportive messages and linked resources when searching for the word "rape", across multiple phones set to different languages (see figure 2). Notably, despite displaying a prompt in Hungarian, French and German, the text-based prompt for phones set to either language forwarded users to English-language resources and displayed UK-based phone hotlines.

For both the desktop and mobile search, the German equivalent "Vergewaltigung" yielded no search results across all four examined locations, accompanied by notices citing potential violations of Community Guidelines (see figure 3 for comparison).

Notably, resources for those affected by sexual violence are displayed on searches for the English-language term "rape" in Germany but not on searches for the German term "Vergewaltigung". Whereas the support message triggered by the search term "rape" centres survivors and those affected by sexual violence, the message triggered by the term "Vergewaltigung" simply states that the word is associated with potentially violative content (figure 3).

The French terms "viol" and "le viol" produce rape-related content without any user notices across all geographies. Hungarian terms such as "nemi erőszak" and "megerőszakol" similarly yield results, although they are often unrelated to rape.

As outlined above, the results shown in table 1 were based solely on searches conducted with a desktop computer using VPNs in order to standardise the analysis. However, as already noted, analysts found additional inconsistencies when conducting searches on the TikTok mobile app, showing further variations across languages, locations, and devices, regardless of the use of VPNs. ISD found that search results and moderation practices appeared to differ depending on the type of device used (e.g. phone or laptop), account language, origin of SIM card and account activity. For instance, while the search term "rape" is blocked and labelled when using a UK VPN on a desktop, the TikTok app does not consistently enforce the blocks and labels for UK-created accounts (figure 3).

| Keyword | Language | (VPN)-Location | Search Result |
|---|---|---|---|
| Rape | English | United Kingdom (UK) | Search blocked with support message; Link to TikTok's sexual abuse resource support page (language of the prompt appears to be based on device location regardless of VPN); The Survivors Trust phone number displayed |
| Rape | English | Germany | Search blocked with support message; Link to TikTok's German-language sexual abuse resource support page |
| Rape | English | France, Hungary | No block, produces search results |
| Vergewaltigung | German | UK, France, Germany, Hungary | Search blocked with German-language community guidelines note |
| Viol, le viol | French | UK, France, Germany, Hungary | No block, produces search results |
| Nemi erőszak & megerőszakolni | Hungarian | UK, France, Germany, Hungary | No block, produces search results |

Table 2: Table indicating search results for the term "rape" in different languages across different VPN locations using an incognito browser window (UK, France, Germany, Hungary).

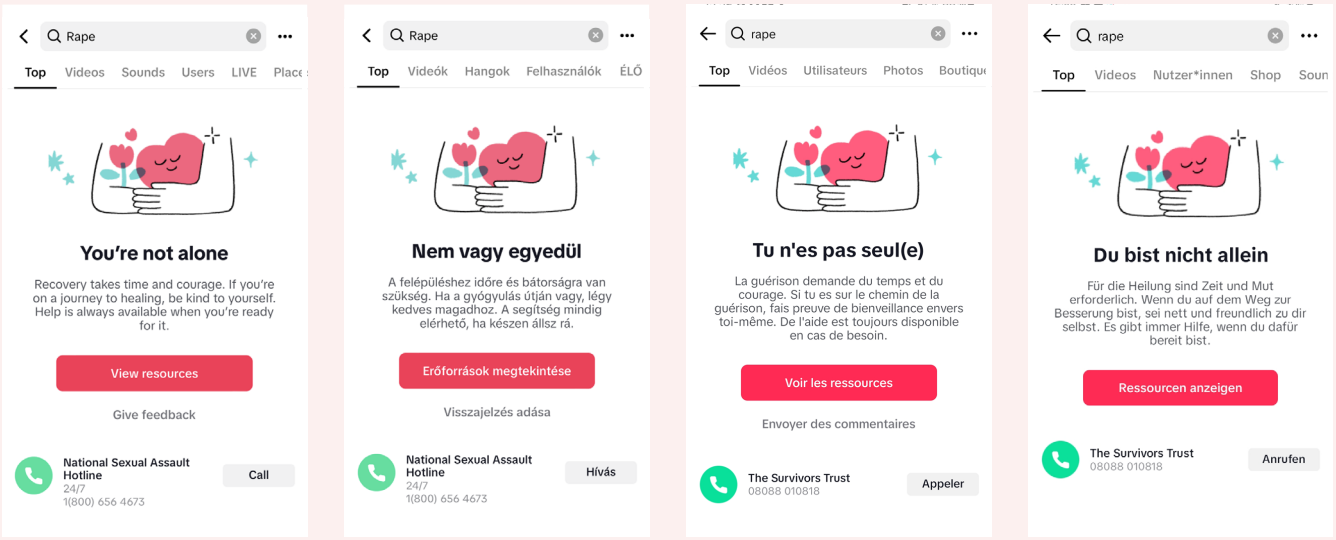



Figure 1: Screenshots showing a supportive message for survivors/victims of rape when searching for the English keyword "rape". Independent from language settings of the phone (from top right English, Hungarian, French and German), further resources are offered when searching for the English term "rape".

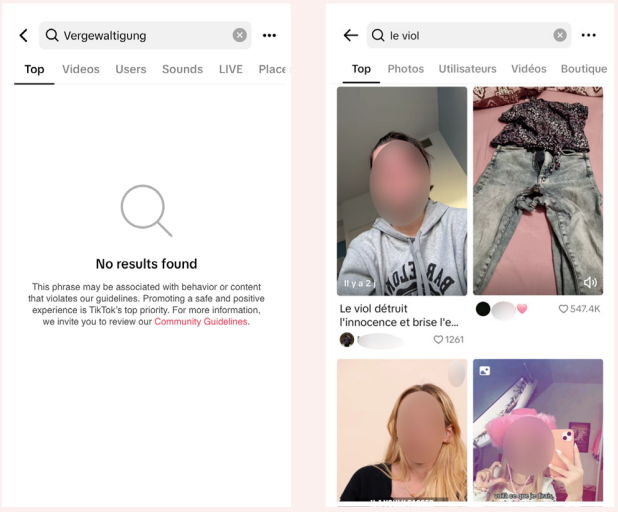



Figure 2: Images showing inconsistent blocking of translations of the term "rape" in different languages. A search for the German keyword "Vergewaltigung" (left) and for the French term "le viol" (right), both of which translate into "rape", leads to a note on community guidelines, and no search block, respectively.

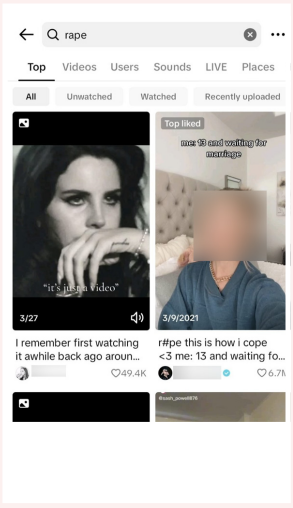



Figure 3: Inconsistent search blocking for the term "rape" using a phone with a UK-created English-language TikTok account and an Italian SIM card in Sweden.

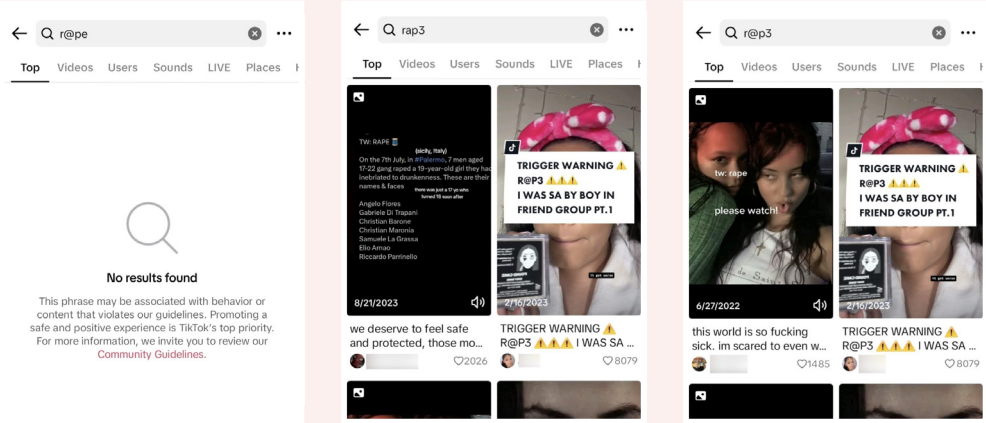



Figure 4: Inconsistent blocking for searches about "rape" in an English-language Tiktok-account using a mobile phone located outside the UK. The searches for the term "rape", "rap3" and "r@p3" were not blocked, while the search for "r@pe" was. None of the searches triggered a user support message.

Inconsistencies were also observed within the same language when using common intentional misspellings such as "rap3", "r@pe" and "r@p3". Only one of these spelling variations was blocked (figure 4).

## TikTok sexual violence moderation gaps: Algospeak & overblocking

Although TikTok's Community Guidelines on "Sensitive and Mature Themes" explicitly allow educational content about sexual health, many creators report that their material is removed or "shadowbanned", which is consistently denied by platforms. In response to TikTok's rigorous approach to topics that users felt were "secretly unwanted but were not explicitly violating TikTok's community guidelines", many users have started to rely on the usage of "netspeak" or "algospeak", to evade algorithmic content moderation that they deem "unjustified". The use of algospeak refers to the "habit of coming up with substitutes for words that [users] worry might either affect how their videos get promoted on the site or run afoul of moderation rules". Examples for algospeak include made up terms such as "unalive" instead of "kill", as well as novel spellings such as le$bian with a dollar sign. Apart from using creative misspellings for "rape" such as "r4p3" or "r@pe", content creators are also using words such as "grape" to refer to sexual violence or "seggs" when discussing sexual health. This development highlights a significant disconnect between platform policies and user experiences.

The nature of algospeak also varies by language and cultural context. For instance, while English speakers may use the netspeak "grape", French users seem to employ a purple circle emoticon or a purple heart to circumvent content blocks and discuss sexual abuse (figure 6).

As such, TikTok's decision to block terms associated with sexual violence, such as "rape" from searches completely, follows a moderation logic that constitutes overblocking as it also affects non-violent content such as reproductive health and sex education content which is explicitly allowed under TikTok's own community guidelines. This approach risks silencing vital conversations on gender, sex and sexual violence. Experts have previously found that online moderation efforts which conflate sexual content and harmful speech often inadvertently discriminate, especially against queer people and sex workers, and against women, transgender, intergender and non-binary people. During this research ISD analysts noticed that users appeared to be moderating their behaviour by using algospeak in conversations on rape. It is likely that this could be a result of TikTok's overblocking.

Blocking keywords related to sexual violence can prevent experts, educators, activists and survivors of violence from sharing and receiving support. Many creators claim and have even attempted to demonstrate in their content that YouTube's, Meta's and TikTok's algorithmic moderation targets words and visuals related to sexual

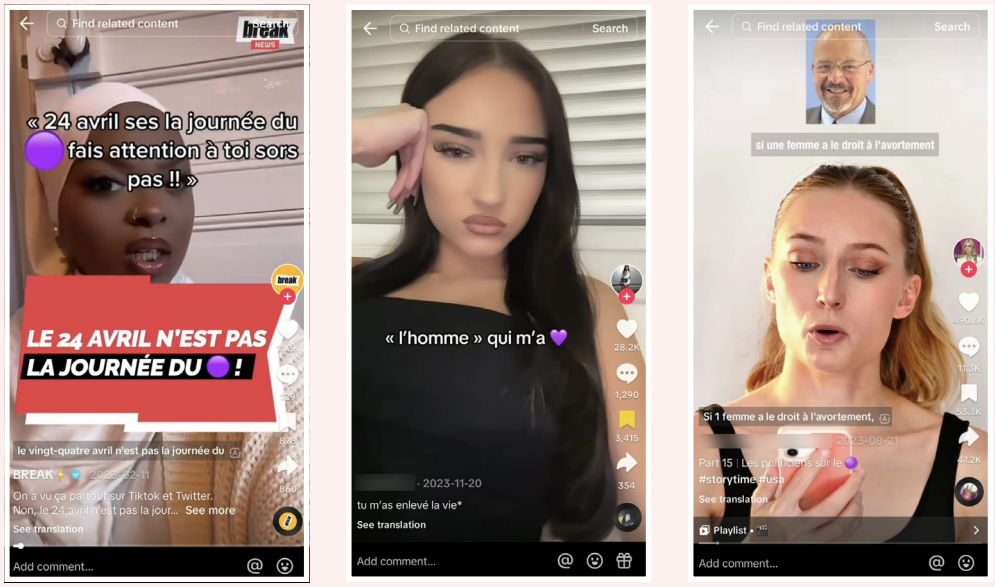



Figure 5: French content discussing sexual abuse using purple circle and purple heart emojis. From left to right: "The 24 of April is not the day of [purple emoji – rape]", "The man that [purple emoji – abused] me" and "Politicians discussing [purple emoji – sexual violence]".

and gender education, which has led to content removal or account suspension despite explicit policy permissions. Facing such restrictions, e.g. in the form of shadowbanning, creators who produce informative content on marginalised identities and social issues have expressed feeling "silenced". Many affected creators and communities have in response turned to coded language and algospeak to circumvent this approach. This becomes a barrier to effectively accessing and sharing sexual health content and information on gender diversity, which can also inhibit community support and relegate crucial discussions on sexual health, gender and gender-based violence.

The banning of language can also exacerbate and reinforce societal stigma associated with sexual education, gender diversity and surviving sexual violence. Sexual education and LGBTQ+ content creators report profound feelings of discouragement and a sense of exclusion, discrimination and stigma due to content removal and visibility restrictions.

# Conclusion & recommendations

ISD found significant inconsistencies in TikTok's keyword moderation approach to sexual violence, including the notices and explanations provided on a blocked search. These differences in moderation practices create an uneven playing field for users, as inconsistent moderation practices imply that users are not equally safeguarded across all locations and languages.

Consequently, TikTok should further enhance standardisation and transparency in its moderation decisions, especially those affecting sexual and reproductive health content. A first important step to fix these gaps could be to allocate proportional resources across languages to ensure equitable access to support and information. The platform currently has a disproportionate number of English-language moderators (2334) compared to other languages (e.g. French: 650; German: 837 and Hungarian: 47).

In addition, TikTok could pivot towards a more nuanced and comprehensive moderation approach. For example, searches using keywords that produce content on sexual assault could prompt notes of support guiding users to helpful resources and at the same time produce search results for educational content and other non-violent content.

TikTok's attempts to combat hateful content and disinformation in searches appear to rely heavily on keyword-blocking content restriction and removal. However, efforts highly focused on keyword enforcement are destined to fall short due to the vast and creative variations users employ to continue promoting this content, whilst at the same time inadvertently limiting important discussions on sexual and reproductive health and safety. Under article 34 of the DSA, TikTok is obliged to manage negative effects in relation to gender-based violence alongside negative effects on civic discourse and freedom of expression, an act of balance that the content moderation strategy observed here is failing to perform.

TikTok's decision to completely block terms associated with sexual violence, such as "rape", constitutes overblocking and affects content which is explicitly allowed under TikTok's own community guidelines. In addition, sexual education and gender activist content creators who rely on platforms for their livelihood also reported that platforms are demonetising and removing gender-related and social justice content and that the lack of transparency and clarity regarding guidelines violations on TikTok hindered their ability to appeal sanctions. Consequently, it is important for TikTok to transparently outline their reasoning for blocking keywords, so as to make the information clearer to both the users and creators affected by these decisions. In addition, more nuance and context should be applied in moderation decisions, for example by adding a human layer of moderation for specific contexts such as the discourse on rape and sexual assault in general.

Finally, implementing comprehensive and context-sensitive keyword detection mechanisms could lead to a more balanced and supportive strategy for monitoring and removing harmful content while providing necessary support to survivors, activists and experts and preserving crucial conversations on sexual education and violence.

# Endnotes

1 Previous versions of TikTok's Community Guidelines were not available for review on the platform's website or in the DSA Digital Services Terms and Conditions Database. Consequently, ISD is unable to assess whether the "Sensitive and Mature Themes" section was recently added to the policy, or if it has changed over time.

ISD | Powering solutions to extremism, hate and disinformation

Amman | Berlin | London | Paris | Washington DC



**www.isdglobal.org**