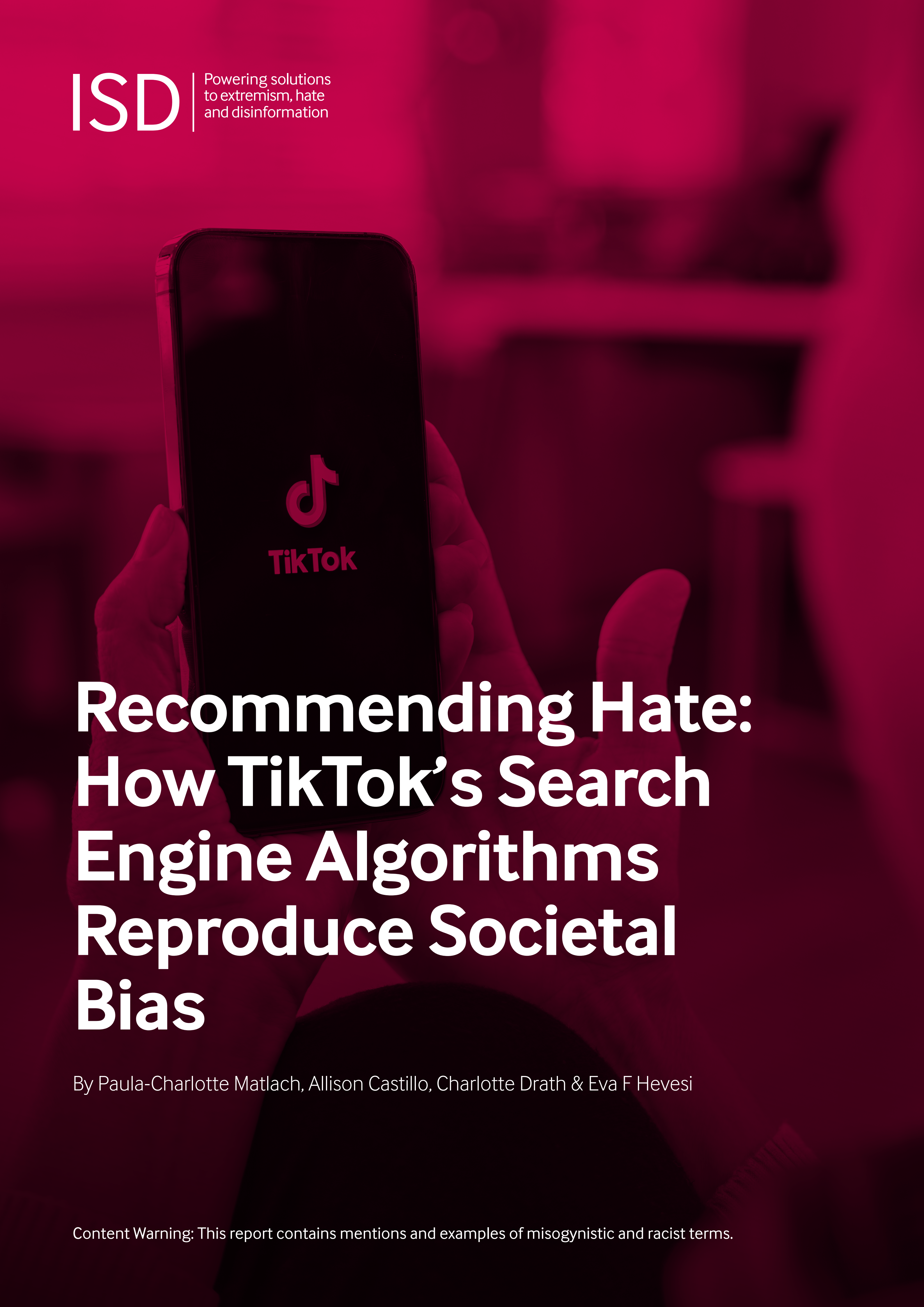
ISD
Powering solutions
to extremism, hate
and disinformation



Recommending Hate: How TikTok's Search Engine Algorithms Reproduce Societal Bias

By Paula-Charlotte Matlach, Allison Castillo, Charlotte Drath & Eva F Hevesi

Content Warning: This report contains mentions and examples of misogynistic and racist terms.

ISD | Powering solutions to extremism, hate and disinformation

Amman | Berlin | London | Paris | Washington DC

www.isdglobal.org

Contents

Overview

ISD analysed TikTok's search engine to examine its moderation processes across English, French, German and Hungarian. Our research found significant evidence of algorithmic bias: across all four languages, search results consistently demonstrated harmful associations that objectify and degrade presumed members of marginalised groups. These findings suggest that in an effort to drive user engagement and increase revenue, TikTok's search and recommender algorithms reproduce and potentially amplify societal biases. The analysis concludes with proposals for both lawmakers and the company to improve safeguards and mitigate the risk of algorithms increasing and perpetuating harm.

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 almost two thirds of the videos (197), TikTok's search engine and recommender algorithms perpetuated harmful stereotypes. This content systematically associated presumed members of marginalised groups with derogatory and violent search prompts. These algorithmic results effectively create pathways that connect users searching for hateful language with content targeting marginalised groups, exposing them to an increased risk of harassment and discrimination.
- TikTok's lack of transparency makes it difficult to understand how the algorithms match search queries and user-generated content. This makes it challenging to precisely determine what causes harmful bias to be reproduced by the platforms search engine and recommender system.
- Only 10 out of 300 videos included the search prompt in their content, descriptions, hashtags, sounds and/or top 10 comments. Of the remaining 290 videos examined, 118 videos partially matched the original prompts, 82 contained associated terms, synonyms, or translations, and 28 featured terms with similar spellings. In 62 cases, analysts found no discernible textual link to the original search terms.

Definitions

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

Algorithmic bias

Algorithmic bias refers to instances where "the outputs of an algorithm benefit or disadvantage certain individuals or groups more than others without a justified reason for such unequal impacts."

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): this 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."

Introduction

In 2015, Google's image search algorithms were found to label images of Black individuals with offensive terms such as "gorilla." This incident exposed the ways in which systemic racism can manifest as algorithmic bias, across both social media recommendation systems and search engines. Although younger users are increasingly using TikTok or Instagram to search for information, research regarding algorithmic bias on TikTok has focused primarily on the platform's personalised feeds rather than its search engine.

This analysis seeks to address this gap by examining TikTok's search engine through algorithmic probing, using targeted prompts to evaluate how search results are moderated. Our findings suggest that TikTok's search ranking algorithms both fail to adequately detect hateful terms and associate these terms with marginalised communities. This reproduces and reinforces societal biases and stereotypes. Even a single instance of online racial discrimination, like being shown a racist image online, has been found to result in negative effects on the mental health of racially minoritised groups. This highlights the urgent need for mitigation of algorithmic bias and a robust strategy to protect marginalised communities from harm online.

Methodology

To audit TikTok's search algorithms, ISD conducted a qualitative analysis using a selection of racist and misogynistic slurs, referred to here as prompts. The search results that TikTok's search engine produced based on these prompts were analysed with a focus on the links between the prompts and the outcome they produced. Algorithmic bias was identified when the search prompts were not directly or only partially present in the context of the results, suggesting that an association was made by the algorithms for the image or video to be displayed.

The analysis used twelve prompts, three for each of the four languages: English, French, German and Hungarian. All search prompts used were racist and misogynistic in nature, but their specific wording varied by language: For each language, one prompt targeting Black, one targeting Romani and one targeting Arab/Muslim women were included.

For each prompt, the first 25 search results were systematically catalogued and categorised based on whether the search prompts appeared within specific elements of the content including the creator's username, image or video material, description or hashtags, sound, and the first ten comments listed. To control for technical factors such as language preference and location, searches for each language were conducted using the same device, settings and VPN location. In total, 300 images and videos were analysed during July and August 2024.

TikTok as a Search Engine

TikTok, like other major social media platforms, allows users to discover content through a search bar using specific prompts. To prevent users from seeking out discriminatory content, TikTok restricts searches that use keywords or phrases which violate the platform's guidelines. However, previous research by ISD found this approach to be incomplete: unblocked hashtags and search terms were identified as significant contributors to the accessibility of violative content on TikTok. During this analysis, researchers again discovered a multitude of racist and misogynistic terms that TikTok's search engine did not classify as harmful, resulting in these terms producing search results.

The ranking of search results on platforms like Google, Yahoo and TikTok is neither random nor neutral. Instead, it is determined by search engine ranking algorithms; previous research have found that platforms have degraded search quality in favour of greater profits. This involves driving user engagement on social media platforms to increase revenue from paid advertisements. To achieve this, platforms have increasingly expanded searches to include results beyond the literal search prompt. These results are based on factors including user behaviour, synonyms and associated terms, location, device and language.

This algorithmic approach can be especially harmful, as it risks perpetuating societal biases by reinforcing associations between discriminatory terms and marginalised communities. It also normalises the use of such terms as descriptors for those affected. For example, in 2015, it was discovered that the search prompt 'N*gga house' directed users to the White House on Google Maps, where then-US President Barack Obama resided. Similarly, in 2011, the search term 'Black girls' predominantly produced pornographic content as "the primary representation of Black girls and women" on the first page of Google search.

Algorithmic Bias on TikTok

Algorithmic probing evidenced a significant lack of transparency in TikTok's search engine ranking system. Of the 300 videos examined, only 10 contained the complete search prompt as on-screen-texts, captions, descriptions, hashtags, sounds, and/or top 10 comments. Of these videos, five were explicitly hateful toward marginalised individuals, highlighting shortcomings in TikTok's content moderation approach, three were content-wise unrelated to the slurs and two critiqued the use of such slurs. Notably, in 197 videos the featured individuals were presumably of the same ethnicity as the slur used in the search prompt and therefore appeared to be its intended targets (figure 1).

TikTok's search algorithms appear to associate presumed members of marginalised groups with hateful and violent search prompts, objectifying and degrading them (figure 2). Further, the algorithms direct those seeking out hateful content (by searching explicitly hateful language) to individuals they may then seek to abuse.

Figure 1. Numbers of videos with no or limited textual reference to the original prompts and numbers of videos featuring individuals who were presumably affected by the search prompts. For English, 47 videos were detected, for French 51, for German 55, and for Hungarian 44.

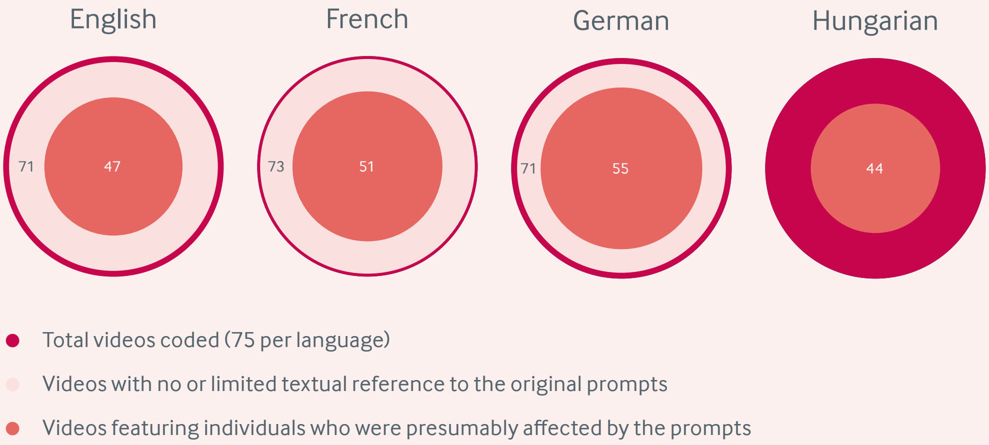
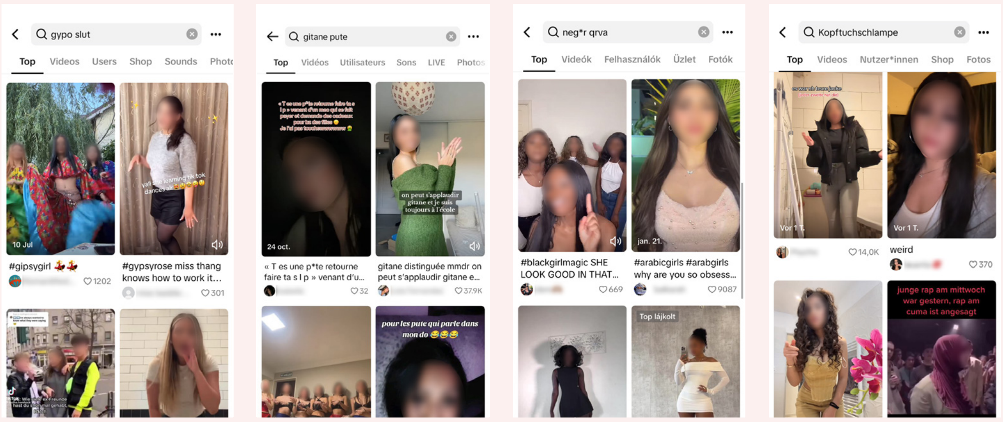


Figure 2. Screenshots from videos shown by TikTok's search engine using an anti-Roma slur in English and French, respectively; an anti-Black slur in Hungarian; and an Islamophobic slur in German (from left to right). All slurs are gendered and aimed at women of the respective communities. The search results feature unassuming users presumably affected by the slurs. The prompt's keywords were not present among all videos depicted in these screenshots.


gypo slut
Top Videos Users Shop Sounds Photo
#gipsygirl
1202
#gypsyrose miss thang knows how to work it...
301
gitane pute
Top Vidéos Utilisateurs Sons LIVE Photos
« T es une p*te retourne faire ta s l p » venant d'u...
32
gitane distinguée mmdr on peut s'applaudir gitane e...
37.9K
pour les pute qui parle dans mon do
neg*r qrva
Top Videók Felhasználók Üzlet Fotók
#blackgirlmagic SHE LOOK GOOD IN THAT...
669
#arabicgirls #arabgirls why are you so obsess...
9087
Top lájkolt
Kopftuchschlampe
Top Videos Nutzer*innen Shop Fotos
14,0K
weird
370
junge rap am mittwoch war gestern, rap am cuma ist angesagt

Underlying Dynamics

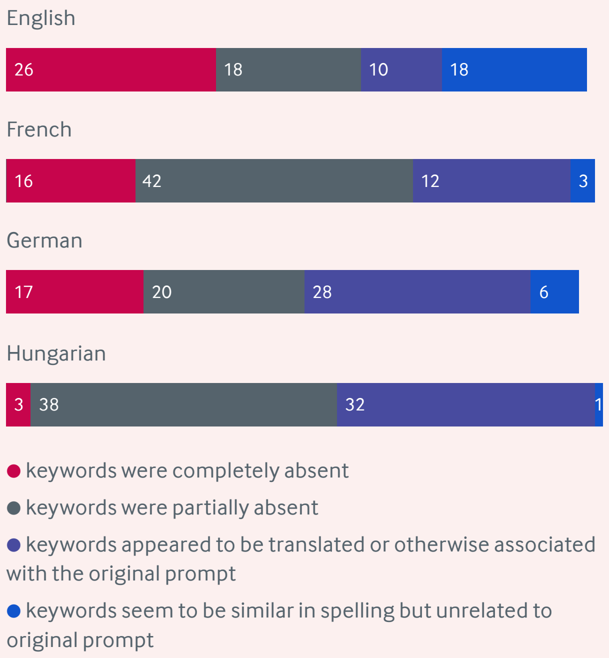
To generate search results, TikTok's recommender system considers factors such as user behaviour and the degree to which content 'matches' the search query. However, TikTok does not disclose details on what exactly the platform considers a 'match' and how these factors are weighed to produce them.

As 290 videos exhibited no or limited textual link to the original prompts, the dynamics underpinning the platform's recommender system could not be precisely determined. In 62 cases, hateful search prompts yielded search results with no discernible textual link to the original prompts in on-screen-text, captions, descriptions, hashtags, sounds and/or top 10 comments. This suggests that factors invisible to researchers, such as metadata and internal tags derived from visual content, were at play. However, among the other 228 videos, we found evidence that multiple textual factors contributed to the manifestation of bias (figure 3):

- In 118 instances, hateful search prompts produced content where the keywords were partially present. For example, one Hungarian search prompt was 'c*gány k*rva' — the first term is a non-gendered term used by some Romani people as a self-descriptor but considered derogatory by others, while the second term is a vulgar sexist slur. This prompt produced videos of Romani content creators, predominantly women. In these cases, the content creators referred to themselves as "c*gány," but the second part of the search prompt was absent. By disproportionately providing videos of women, the recommender algorithms appear to have made and reinforced a harmful association between the sexist slur and (Romani) women.

- In 82 instances, hateful search prompts produced content that did not feature the original keywords but included associated terms, synonyms or translations. For example, analysts used the French prompt 'guen*n p*te,' a hateful slur towards Black women, which consists of a term for a monkey species and a vulgar sexist slur. This led to videos where one of two keywords was present in the absence of the original prompts: "chimpanzé" (a likely association to the first part of the prompt), or "p*ta" (a possible Spanish translation of the second part of the prompt, with strong derogatory connotations). Both examples highlight the need for sensitivity and safeguarding in the development of search query expansion algorithms to avoid embedding and reproducing harmful associations through this practice.

Figure 3: Graph showing numbers of videos per language where either no textual link was identified (62) or one of three possible categories indicated a potential textual link, namely the partial presence of keywords (118), the assumed translation or association between prompt and other keywords identified in the videos (82), and the presence of keywords similar in spelling but unrelated to the original prompt (28).



- In 28 instances, hateful search prompts were absent but other terms close in spelling were found. For example, the term 'Nigeria' was found present in videos when searching for the German gendered anti-Black slur "N*gerin" while the prompt itself was absent. Although the two keywords are similar in spelling, they differ vastly in meaning. Presumably, TikTok's search algorithms processed the original prompt and 'auto-corrected' the search, which resulted in a harmful association of the anti-Black slur with content related to Nigeria, including the portrayal of Black individuals who are presumably victimised by the original hateful prompt. While spelling correction is a common practice among search engines, transparency is typically maintained by notifying users using labels such as "Showing results for:" or "Did you mean searching for:". In the cases examined here, no such label was displayed.

These surface-level observations emphasise the complexity and opacity of the various decisions that come into play during the key moment of 'matching' search queries with content. They underscore the urgent need for improvements regarding transparency on the curation and delivery of content to users and the mitigation of algorithmic bias.

Conclusions and Recommendations

These observations evidence that TikTok's search engine fails to adequately detect hateful terms. In addition, bias in TikTok's search engine ranking and recommender algorithms harms marginalised groups by associating them with explicitly derogatory search prompts. Due to limited data access, researchers were only able to derive the possible mechanisms of algorithmic bias through publicly available data like video descriptions rather than more descriptive data such as TikTok's internal classification labels.

Previous research has shown that all manifestations of gender-based violence – from offensive slurs to femicides – form a continuum of violence. As opposed to categorizing and conceptualizing gender-based violence as individual and episodic actions, the concept of the continuum of violence recognizes this violence as interconnected and systemic, rooted in normalized misogynistic beliefs and biases. Like societal bias, algorithmic bias reproduces and reinforces existing beliefs. If left unaddressed, the algorithmic bias ingrained in TikTok's search engine will continue to play a part in the continuum of gender-based violence, making the platform unsafe for marginalised communities. This also underscores the need for further research into the manifestation of algorithmic bias on social media platforms in general and on TikTok in particular.

To address these shortcomings, TikTok should:

- Complement the use of AI-based systems to detect and moderate harmful content, including search prompts, with human oversight. This requires teams with specific expertise on (illegal) hate speech against women, transgender, non-binary and genderqueer people. This specialism allows for nuanced approaches that recognise the role of subtle and veiled misogyny, and can help mitigate algorithmic bias. TikTok should be transparent about the group-specific qualifications of human moderators (e.g. expertise/intersectional training in the field of gender-based violence and discrimination).

- Incorporate gender analysis and feminist methodology when assessing the risks of algorithms and machine-learning (ML) models embedded in their services and ensure that relevant teams (such as those designing, testing, and evaluating algorithms) are diverse and trained on how to conduct gender analysis to detect and mitigate biases and discriminatory patterns.

- Improve transparency to better enable the prevention, detection and ultimately the addressing of discrimination and bias embedded into algorithms. Although TikTok currently provides information regarding the basic parameters used in its search and feed personalisation, it is unclear how exactly content is 'matched' to search queries and user interests. TikTok should provide public information about its search engine and recommender algorithms' rationale; it should also provide the assumptions regarding potentially affected groups, the main categorisation choices and for what they are designed to optimise, the specific relevance of the different parameters, and the decisions about any possible trade-offs.

- Strengthen its commitment to inclusivity and ethical practices by identifying, assessing, and mitigating the influence of its search engine and recommender algorithms on systemic risks. This is part of their obligation to create a safe and fair digital space, and to mitigate negative effects on fundamental rights and of gender-based violence. These commitments fall under article 1, 34, and 35 of the EU's Digital Services Act (DSA) and under the United Nations Human Rights Committee Resolution A/HRC/RES/53/29.

- Involve representatives of groups potentially impacted by hateful content in TikTok's DSA risk assessment methodology. This includes consultations on harms resulting from the design of its search engine and recommender algorithms as well as on the development of related risk mitigation measures. TikTok should publish up-to-date and detailed reports on the results of its risk assessment, including information on how the feedback from representatives of the different groups was considered.

Lawmakers should:

- Consider ways to enable regulators to independently monitor and respond to the outcomes of algorithmic decision-making on platforms, when enforcing or supporting the enforcement of risk assessment and mitigation requirements of the DSA or risk-based governance approaches in other jurisdictions. This could be addressed through adequate personnel and financial resources.

- Expand the existing requirements for recommender system transparency. Article 27 of the DSA mandates the providers of online platforms to make public the criteria used for procuring recommendations and the reasons for the relevant importance of these parameters. However, this information is insufficient to understand how bias continues to be embedded and reproduced by the underlying systems. Therefore, platforms should be mandated to publish detailed information on how these criteria are factored in to produce results which 'match' user interests and search queries.

- Implement mandatory access to non-public data of providers of very large online platforms (VLOPs) and very large online search engines (VLOSEs), as mentioned in article 40(4) of the DSA. This should be done promptly so that vetted researchers can contribute to detecting, identifying and understanding the negative effects on users' fundamental rights and the continued exertion of gender-based violence. Amongst others, this access should entail data on reach and information on internal classification labels, if appropriate.

ISD | Powering solutions
to extremism, hate
and disinformation

Amman | Berlin | London | Paris | Washington DC

www.isdglobal.org