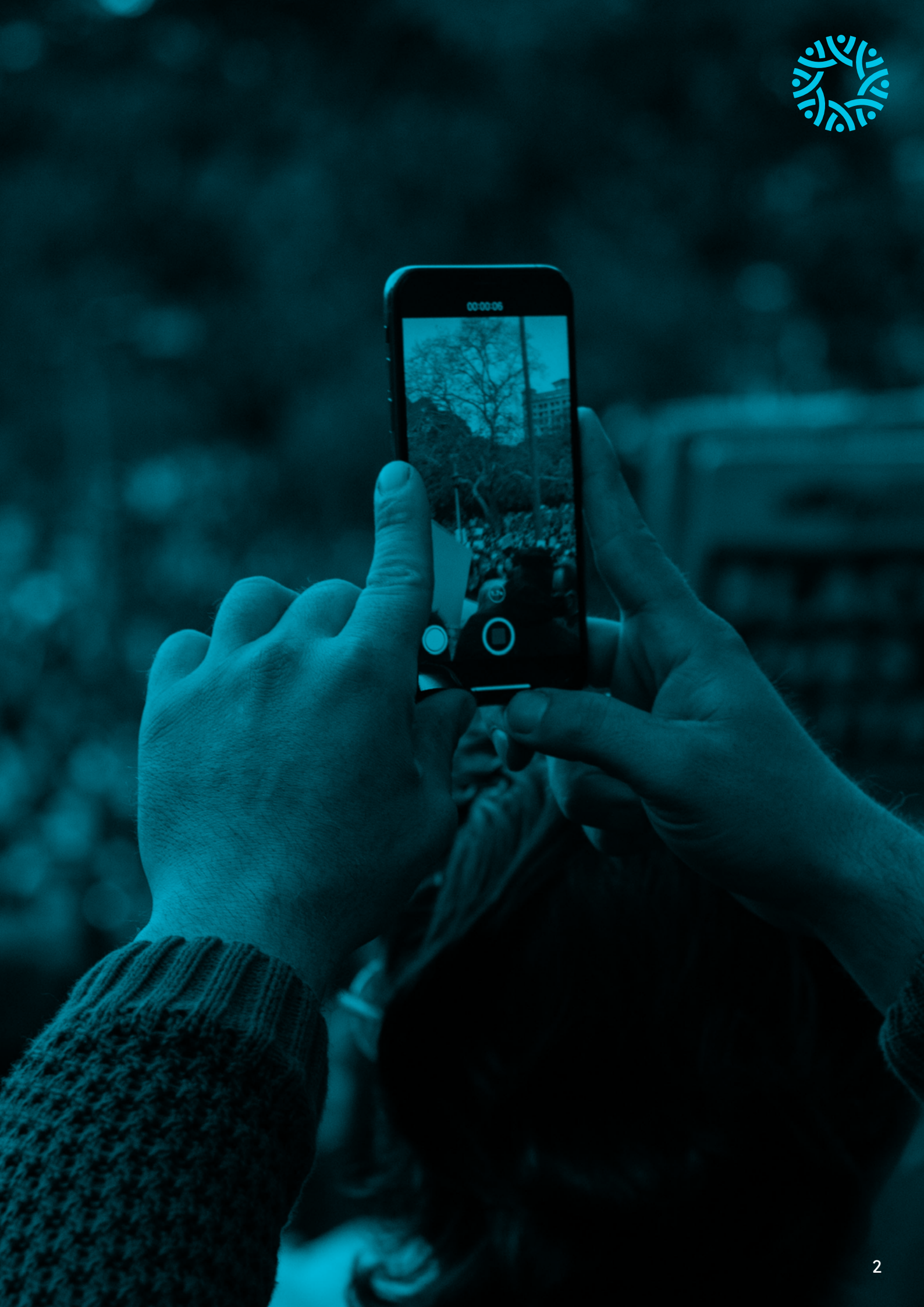
LEGAL RESOURCES CENTRE

A Critical Analysis of Content Moderation Policies and the Impact of Spreading Violence, Hatred & Disinformation in the Global South

Compiled by: Kristen Abrahams, Rachel Moore and Keja Wynn

Edited by: Sherylle Dass

Table of Contents


00:00:06

I Background and Introduction

Social media is an important – if not indispensable – medium of communication globally, with an estimated 3.2 billion social media users around the world, constituting 42% of the world's population.[1] It has become commonplace that social media platforms are not only used as centres for entertainment and the sharing of information, but also for robust debate and deliberation.

While the exchange of ideas from individuals with differing backgrounds is encouraged and required in a global context, social media platforms have also been used for the spread of disinformation, hate speech and incitement to violence. Recent examples of the ways in which social media was used to incite violence include anti-Muslim violence in Sri Lanka;[2] and racist, right-wing violence in 2021 that emerged following the national election in the United States of America (USA).[3]

As South Africa heads towards its general elections in 2024, the incitement of violence following the 2020 presidential elections in the USA is particularly worrying. It is within this context that this report is situated. This report aims to contribute to the Legal Resources Centre's grasp of the scope and breadth of social media content moderation policies, in order to inform the LRC's interventions in relation to safeguarding the integrity of South Africa's general elections in 2024. The report focuses specifically on content moderation policies in relation to disinformation, hate speech and incitement to violence, aiming to provide a basis for crafting interventions which are tailored to prevent all three, leading up to and following the national election; and thus advance democracy.

The report will examine the content moderation policies of four social media companies – Meta (Facebook), TikTok, Google (YouTube) and X (formallyTwitter) – in order to ascertain: (a) the scope of content moderation covered by these policies; (b) any gaps in these policies, particularly in relation to the South African context; (c) the extent to which these policies are being consistently implemented globally; and (d) to make recommendations based on the findings.

[1] University Canada West 'How Has Social Media Emerged as a Powerful Communication Medium?' available at www.ucanwest.ca/blog/media-communication/how-has-social-media-emerged-as-a-powerful-communication-medium/, accessed on 19 May 202.3

[2] Megha Rajagopalan and Aisha Nazim. 'We Had to Stop Facebook: When Anti-Muslim Violence Goes Viral.' BuzzFeed News 7 Apr. 2018 available at www.buzzfeed.com/meghara/we-had-to-stop-facebook-when-anti-muslim-violence-goes-viral, accessed on 19 May 2023.

[3] Craig Silverman, Craig Timberg, Jeff Kao and Jeremy Merrill 'Facebook Groups Topped 10,000 Daily Attacks on Election before Jan. 6, Analysis Shows.' The Washington Post 4 Jan 2022 available at www.washingtonpost.com/technology/2022/01/04/facebook-election-misinformation-capitol-riot/, accessed on 19 May 2023.

II Policies Regulating Hate Speech, Incitement To Violence, Misinformation And Elections

The social media companies examined in this report operate under a set of community guidelines. In relation to hate speech, these major social media companies appear to follow the same outline, using similar definitions of hate speech and including largely the same protected characteristics and traits under their policies. Notably, the policies operate at a universal level – that is, irrespective of geographical location, the policies apply as written.

META (Facebook)

Facebook's Policy on Hate Speech

Facebook has 2.98 billion monthly active users.[4] Facebook [5] prohibits hate speech on the platform because 'it creates an environment of intimidation and exclusion, and in some cases may promote offline violence.' [6] The company defines 'hate speech' as 'a direct attack against people – rather than concepts or institutions – on the basis of what (we) call protected characteristics: race, ethnicity, national origin, disability, religious affiliation, caste, sexual orientation, sex, gender identity and serious disease.'[7] 'Attacks' in this sense are defined 'as violent or dehumanizing speech, harmful stereotypes, statements of inferiority, expressions of contempt, disgust or dismissal, cursing and calls for exclusion or segregation.' [8]

---

[4] S Dixon 'Facebook: quarterly number of MAU (monthly active users) worldwide 2008-2023' available at https://www.statista.com/statistics/264810/number-of-monthly-active-facebook-users-worldwide/#:~:text=How%20many%20users%20does%20Facebook,used%20online%20social%20network%20worldwide, accessed on 29 May 2023.

[5] The Community Standards governing Facebook apply to Meta as a whole – which includes Instagram, WhatsApp, Messenger and Quest 2.

[6] Facebook Community Standards on Hate Speech available at transparency.fb.com/en-gb/policies/community-standards/hate-speech/, accessed on 19 May 2023.

[7] Ibid.

[8] Facebook Community Standards on Hate Speech: Publisher and Creator Guidelines available at www.facebook.com/business/help/170857687153963?id=208060977200861, accessed on 19 May 2023.

Facebook claims that while they seek to remove hate speech, they do allow content that may contain an example of hate speech but is shared with the intent and purpose to educate and raise awareness provided that the intent must be made clear and explicit. [9]

Facebook outlines three tiers of prohibited content, providing examples of content that goes against Facebook Community Standards, as outlined under the company's hate speech policy. [10] The protected characteristics in terms of the policy include 'race, ethnicity, national origin, religious affiliation, sexual orientation, sex, gender, or gender identity, serious disabilities or diseases.' Facebook does not indicate that this is an open list and it is thus unclear whether other characteristics, not listed, are protected under Facebook's policy. Each tier outlines prohibited content that targets a person or group of people based on protected identities and characteristics through the use of: violent and/or dehumanizing speech, mockery; generalizations and/or statements of inferiority, expressions of dismissal, contempt, and disgust, cursing, and finally, segregation and/or exclusion. [11]

Facebook's Policy on Incitement to Violence

In addition to prohibiting hate speech, the company prohibits 'harmful stereotypes' and 'dehumanizing comparisons that have historically been used to attack, intimidate or exclude specific groups, and that are often linked with offline violence.' [12]

Facebook will only remove 'language that incites or facilitates serious violence' as opposed to 'calling for violence in non-serious ways,' but the platform does not define either 'serious' or 'non-serious violence. It can be argued that this lack of clarity creates a gap which could lead to content facilitating violence being approved , on the basis that it is 'non-serious.'

[9] Op cit note 5.

[10] Ibid.

[11] Op cit note 6.

[12] Ibid.

A report by The Promise Institute for Human Rights (PIHR) illustrates how content that does not constitute the facilitation or incitement to violence can nonetheless 'feed into negative generalisations' [13] and thus further division and animosity. Thus, it can be argued that such content would be viewed by Facebook as 'non-serious' (this cannot be verified, as Facebook does not provide a definition of the term) and would not be removed as a result thereof. However, the division it engenders could lead to or exacerbate already-existing tensions between people groups and thus constitute harm.[14]

As per Facebook's Violence and Incitement guidelines, content will only be removed where it is believed that there is 'a genuine risk of physical harm or direct threats to public safety.' It can be argued that the focus here on 'physical harm,' to the exclusion of other forms of harm, creates a gap that would leave targets of other forms of violence (harm) – such as psychological or emotional violence – vulnerable. [15]

Lastly, Facebook states that it 'tries to consider language and context.' [16] The platform is operative in 157 countries across the globe – its failure to do more than 'try' to consider language and context will arguably result in millions of native language speakers being vulnerable to violence if the platform fails to identify and thus block violative content in other non-English languages and within other cultural contexts.

[13] Aya Dardari, Nicholas Levsen, Ani Setian and Jessica Peake 'Social Media, Content Moderation and International Human Rights Law' available at law.ucla.edu/sites/default/files/PDFs/Promise/Social%20Media%2C%20Content%20Moderation%20and%20International%20Human%20Rights%20Law.pdf, accessed on19 May 2023.

[14] Ibid.

[15] Vinney, Cynthia. 'What Is the Impact of Violent Media on Mental Health?' Verywell Mind 23 June 2022 available at www.verywellmind.com/what-is-the-impact-of-violent-media-on-mental-health-5270512, accessed on 19 May 2023.

[16] Facebook Community Standards on Misinformation available at transparency.fb.com/en-gb/policies/community-standards/misinformation/, accessed on 19 May 2023.

Facebook's Policy on Misinformation & Elections

Facebook removes misinformation where 'it is likely to directly contribute to the risk of imminent physical harm.' [17] This can be criticised on the basis that the requirement for violating content here is that misinformation must directly contribute to the risk of imminent physical harm. This is superfluous, as the definition of misinformation is 'false or inaccurate information... which is deliberately intended to deceive' [18] – by definition, misinformation is aimed at informing and shaping ideas through deception, which is inherently indirect. Thus, Facebook's requirement that the misinformation directly contribute to harm, is frivolous.

The ad-tech company also removes content 'that is likely to directly contribute to interference with the functioning of political processes and certain highly deceptive manipulated media. In determining what constitutes misinformation in these categories, (Facebook) partner(s) with independent experts who possess knowledge and expertise to assess the truth of the content and whether it is likely to directly contribute to the risk of imminent harm.' [19] Further, it can be argued that there is a need for Facebook to employ localised third parties, as 'an examination... reveals that in some of the world's most volatile regions, terrorist content and hate speech proliferate because the company remains short on moderators who speak local languages and understand cultural contexts.' [20] Such employment (or a commitment to this kind of employment) would guarantee some level of understanding of the context and thus would provide insight into what could possibly contribute to 'imminent physical harm'. This notwithstanding, the focus here on physical harm to the exclusion of other types of harm is questionable and will be addressed in the Conclusion of this report.

[17] Facebook Community Standards on Misinformation available at transparency.fb.com/en-gb/policies/community-standards/misinformation/, accessed on 19 May 2023.

[18] Oxford Languages, Google.

[19] Facebook Community Standards on Misinformation available at transparency.fb.com/en-gb/policies/community-standards/misinformation/, accessed on 19 May 2023.

[20] Fares Akram 'Facebook's language gaps allow terrorist content and hate speech to thrive' available at https://www.pbs.org/newshour/world/facebooks-language-gaps-allow-terrorist-content-and-hate-speech-to-thrive, accessed on 29 May 2023.

Facebook further prohibits 'content and behaviour in other areas that often overlap with the spread of misinformation.' [21] By way of example, Facebook's Community Standards 'prohibit fake accounts, fraud and coordinated inauthentic behaviour.' [22] While this acknowledgment of the overlap between misinformation and other forms of violable content is laudable, the overlap is not addressed explicitly in its Community Standards. For example, there should be an explicit recognition that there is the possibility for heightened tensions in countries leading up to and proceeding a national election. This recognition would then lead to Facebook enacting Standards which address any overlaps – by providing for the removal of content which constitutes, for example, misinformation that does not contribute directly to the risk of imminent harm, but does have the effect of perpetuating negative stereotypes and inciting violence against certain groups. Facebook's delineation of overlapping categories into 'neat' boxes is thus not reflective of the aims of certain categories – like misinformation.

Categories of misinformation that Facebook removes include: 'content that is likely to directly contribute to a risk of imminent violence or physical harm to people;' 'misinformation likely to directly contribute to imminent harm to public health and safety;' and most importantly for our purposes, Facebook claims to 'remove misinformation that is likely to directly contribute to a risk of interference with people's ability to participate in [the census and elections.]' [23] The final subsection of prohibited misinformation is dubbed manipulated media, which is media that meets the following criteria: '(1) the video has been edited or synthesised, beyond adjustments for clarity or quality, in ways that are not apparent to an average person and would likely mislead an average person to believe a subject of the video said words that they did not say; and (2) the video is the product of artificial intelligence or machine learning, including deep learning techniques (e.g. a technical deepfake), that merges, combines, replaces and/or superimposes content onto a video, creating a video that appears authentic.' [24]

[21] Op cit note 15.

[22] Ibid.

[23] Ibid.

[24] Ibid.

The 'voter or census interference' section of prohibited misinformation includes the following examples:

- 'Misinformation about the dates, locations, times and methods for voting, voter registration or census participation;
- Misinformation about who can vote, qualifications for voting, whether a vote will be counted, and what information or materials must be provided in order to vote;
- Misinformation about whether a candidate is running or not;
- Misinformation about who can participate in the census and what information or materials must be provided in order to participate;
- Misinformation about government involvement in the census, including, where applicable, that an individual's census information will be shared with another (non-census) government agency;
- Content falsely claiming that the US Immigration and Customs Enforcement (ICE) is at a voting location; and
- Explicit false claims that people will be infected by COVID-19 (or another communicable disease) if they participate in the voting process.' [25]

While the above-mentioned prohibitions are laudable, Facebook notably adds that they 'have additional policies intended to cover calls for violence, the promotion of illegal participation and calls for coordinated interference in elections, which are represented in other sections of (our) Community Standards.' [26] This is noteworthy, as it depicts the ways in which Facebook maintains isolated sections relating to violence, hate speech and elections; but does not provide for the overlap between them or address that overlap in any of its guidelines.

[25] Ibid.

[26] Ibid.

This can be depicted through the PIHR report which illustrates the 'information war' over social media and how this exacerbated tensions between Azerbaijani and Armenian forces during a physical war between the two groups in 2021. [27] The report's description of a Facebook post (which was found to be inauthentic and tailored) depicting Azerbaijani forces cutting off the ear of an Armenian soldier is illuminating:

'Facebook's rationale for its "Violent and Graphic Content" policy stipulates that Facebook "remove[s] content that glorifies violence or celebrates the suffering or humiliation of others because it may create an environment that discourages participation [on Facebook]." The policy rationale further provides that Facebook "allow[s] graphic content (with some limitations) to help people raise awareness about these issues." For content that falls within the policy, Facebook indicates that it will "include a warning screen so that people are aware that the content may be disturbing." In this case, the AAAC's post does not glorify violence or celebrate the plight of the Armenian soldiers; instead, it does the opposite by calling out the alleged Azerbaijani military violations of the treatment of prisoners of war. Consequently, the post did not violate Facebook's "Violent and Graphic Content" policy and, therefore, was properly not the subject of moderation under this policy.' [28]

It is arguable that even though the video did not constitute prohibited 'violent and graphic content' under Facebook's policy, it could constitute content in violation of Facebook's 'Violence and Incitement' policy, which prohibits content which contributes 'to the risk of imminent violence or physical harm.' [29] As the PIHR report illustrates that, however, because the video did not definitively contribute in this way, it was not in violation of Facebook's policy on violence and incitement, either.

[27] Op cit note 12.

[28] Ibid.

[29] Ibid.

However, as the report notes, 'the post casts Azerbaijani soldiers as war crime perpetrators, which could feed into negative generalisations about Azerbaijanis.' [30] It thus can be argued that the content – in perpetuating harmful stereotypes – constitutes hate speech under Facebook's policy guidelines. This notwithstanding, it is posited that because there is no provision for the overlap between categories of content – such as this example, which could constitute glorifying violence, incitement or hate speech – content not found in violation of one category would not be removed on that basis, without more. Particularly in the context of conflict – during which content can be used to exacerbate tensions – there is a need for ad-tech company policies to address content which constitutes an overlap between, inter alia, hate speech, incitement to violence, misinformation and elections policies.

[30] Ibid.

Google (YouTube)

YouTube's Policy on Hate Speech

YouTube has over 2.68 billion users. [31] YouTube's community guidelines are similar to Facebook's in relation to hate speech. Notably, the protected traits under YouTube's policy are more expansive than those found in Facebook's policy. YouTube's list of protected traits contains the following grounds that Facebook omits: age, caste, immigration status, victims of a major violent event and their kin, [and] veteran status. [32]

The above notwithstanding, the list of protected traits on YouTube, though non-exhaustive, excludes certain protected grounds in Section 9(3) of the Constitution of the Republic of South Africa. [33] These grounds are: pregnancy, marital status, social origin, colour, conscience, culture, language and birth. These grounds are included in Section 9(3), in order to provide for the diversity of people living in South Africa, many of whom would be devoid of protections on YouTube if the 'protected traits' list does not encompass the intersecting components of their nature.

[31] Daniel Ruby 'YouTube Statistics 2023: Data for Brands & Creators' available at https://www.demandsage.com/youtube-stats/, accessed on 29 May 2023.

[32] YouTube Policies on Hate Speech available at support.google.com/youtube/answer/2801939?hl=en#:~:text=Hate%20speech%20is%20not%20allowed,Caste, accessed on 19 May 2023.

[33] The Constitution of the Republic of South Africa, 1996.

These grounds are: pregnancy, marital status, social origin, colour, conscience, culture, language and birth. These grounds are included in Section 9(3), in order to provide for the diversity of people living in South Africa, many of whom would be devoid of protections on YouTube if the 'protected traits' list does not encompass the intersecting components of their nature.

If a user is found to have posted content that violates YouTube's hate speech policy, the company states that they will remove the content and notify the user via email. If it is the user's first time violating the community guidelines, then they will 'likely get a warning with no penalty to [their] channel.' [34] If the instance is a repeat violation, YouTube 'may issue a strike against [their] channel.' [35] Finally, if a user receives three strikes within a period of 90 days, the channel will be terminated. [36]

However, YouTube's language in relation to violations of their policies is not altogether certain. It provides that 'in some rare cases, they may remove content or issue other penalties when a creator:

- **Repeatedly encourages abusive audience behaviour.**
- **Repeatedly targets, insults and abuses a group based on the attributes noted above across multiple uploads.**
- **Exposes a group with attributes noted above to risks of physical harm based on the local social or political context.**
- **Creates content that harms the YouTube ecosystem by persistently inciting hostility against a group with attributes noted above for personal financial gain.' [37]**

[34] Op cit note 33.

[35] Ibid.

[36] Op cit note 27.

[37] Ibid.

All of the above behaviour represents a flagrant disregard of YouTube's content moderation policies. However, the platform's qualifications of 'rare cases' and that it 'may' remove content makes disciplinary action taken for violations less threatening than it should be. This could be improved by strengthening the provision with clearer and more certain language.

Generally, YouTube allows content where it is for educational, documentary, scientific or artistic purposes. Notably, where content falls into any of these categories, there must be a clear indication of this within the content itself. [38] This is a good approach to content which would otherwise be rejected for constituting hate speech; and other ad-tech platforms can follow suit, requiring an explicit acknowledgment of the nature of the content.

YouTube's Policy on Incitement to Violence

In addition to removing hate speech, YouTube's official policy is to not allow content that encourages violence based on the protected traits listed under the hate speech policy. Additionally, threats are not allowed and the corporation treats 'implied calls for violence' as real threats.[39]

YouTube's Policy on Misinformation

YouTube's advertisement policy directly advises that all Elections and Political content follow local legal requirements first and foremost. All advertisements must also comply with the advertising policies of Google, the conglomerate that owns YouTube. [40]

---

[38] Ibid.

[39] Ibid.

[40] YouTube Policies on Misinformation available at support.google.com/youtube/answer/10834785?hl=en, accessed on 19 May 2023.

YouTube prohibits 'certain types of misleading or deceptive content with serious risk of egregious harm. This includes certain types of misinformation that can cause real-world harm, like promoting harmful remedies or treatments, certain types of technically manipulated content, or content interfering with democratic processes.' [41] To this end, it prohibits content that constitutes:

- **'Suppression of census participation: Content aiming to mislead census participants about the time, place, means, or eligibility requirements of the census, or false claims that could materially discourage census participation;**
- **Manipulated content: Content that has been technically manipulated or doctored in a way that misleads users (beyond clips taken out of context) and may pose a serious risk of egregious harm;**
- **Misattributed content: Content that may pose a serious risk of egregious harm by falsely claiming that old footage from a past event is from a current event;**
- **Promoting dangerous remedies, cures, or substances: Content that promotes harmful substances, treatments, or substances that present an inherent risk of severe bodily harm or death; and**
- **Contradicting expert consensus on certain safe medical practices: Content that contradicts local health authorities' or WHO guidance on certain safe medical practices.'** [42]

YouTube fails to define what constitutes 'real-world harm' and addresses content which poses only 'serious risk of egregious harm.' It is arguable that the risk of egregious harm alone is sufficient to warrant the removal of content which constitutes same; and thus can be argued that YouTube's additional requirement that the risk be serious is unnecessary. [43]

[41] Ibid.

[42] Op cit note 41.

[43] Ibid.

YouTube's Election Policy

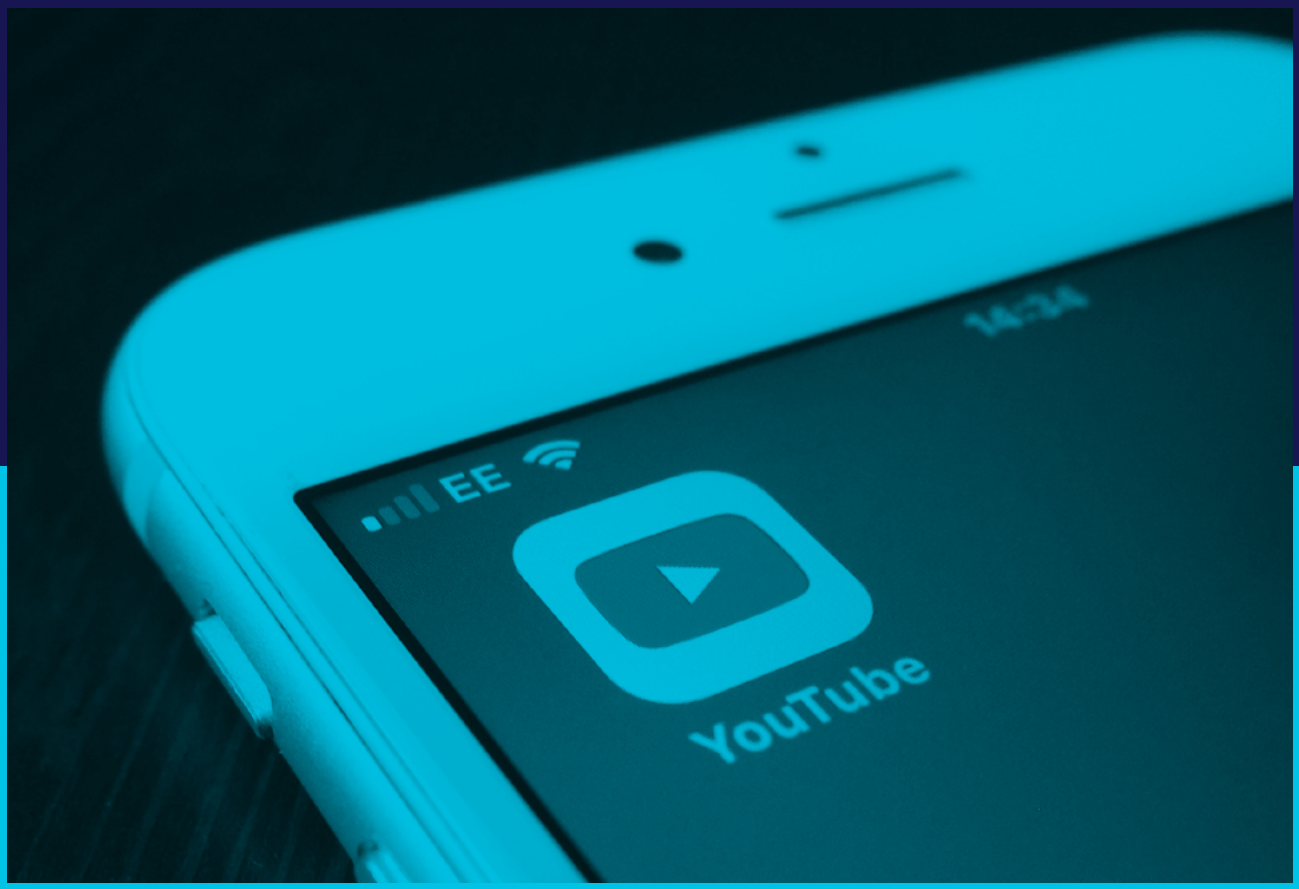
YouTube's election misinformation policy relates to content uploaded by users to the platform through video. It does not appear to also apply to advertisements submitted by users, thus it raises the question of whether the same policy applies in relation to advertisements, as they are not listed under the same page nor are ads mentioned in this particular part of Google's policy.

YouTube prohibits any misinformation that can cause real world harm or that also interferes with any democratic processes. The non-exhaustive list of election content not allowed on YouTube's platform is as follows:

- **Voter suppression – for example, telling viewers they can vote through inaccurate methods like texting their vote to a particular number;**
- **Candidate eligibility – for example, claims that a candidate or sitting government official is not eligible to hold office based on false info about the age required to hold office in that country/region;**
- **Incitement to interfere with democratic processes – for example, telling viewers to hack government websites to delay the release of elections results.**
- **Distribution of hacked materials – for example, videos that contain hacked info about a political candidate shared with the intent to interfere in an election;**
- **Election integrity – YouTube specifies that this policy – which includes false claims relating to the outcome of a past election – currently (but not only) applies to: any past U.S. Presidential election; the 2021 German federal election; and the 2014, 2018, and 2022 Brazilian Presidential elections.** [44]

[44] YouTube Policies on Elections Misinformation available at support.google.com/youtube/answer/10835034?hl=en, accessed on 12 May 2023.

YouTube's focus on 'real-world' harm is open to criticism. First, the ad-tech company provides no definition of the term 'real-world harm,' thus creating a gap through which misinformation is allowed on the platform, on the basis that the content does not constitute this sort of undefined harm; this determination is therefore left entirely to YouTube content moderators. Further – and as shown in the example outlined above – misinformation can have the effect of perpetuating stereotypes or constituting an attack against a people group, based on shared characteristics (like the Azerbaijanis), [45] but it is unclear that this would constitute 'real-world harm.' Thus, it can be argued that where misinformation has the same effect as hate speech would, misinformation would not be removed on the basis that it does not contribute to real-world harm. This cannot be what YouTube intends – as its prohibitions against hate speech illustrate – but this is a consequence which, arguably, could easily follow, based on the current drafting of its policies.



[45] Op cit note 13.

TikTok

TikTok's Policy on Hate Speech & Incitement to Violence

TikTok has over 1 billion users [46] and it describes itself as a 'diverse and inclusive community that has no tolerance for discrimination [47] TikTok's hate speech policy states: 'we do not permit content that contains hate speech or involves hateful behaviour, and we remove it from our platform. We ban accounts and/or users that engage in severe or multiple hate speech violations or that are associated with hate speech off the TikTok platform'. [48]

As mentioned above, users are banned only when they engage in severe or multiple hate speech violations on TikTok. As with Facebook, there is no clarity on what constitutes severe hate speech, as opposed to less severe. This is problematic, as content deemed by TikTok content moderators as 'less severe' could be harmful, but on the basis of TikTok's policy, would be approved of.

Hate speech includes both speech and behaviour, and is defined as 'content that attacks, threatens, incites violence against, or otherwise dehumanizes an individual or a group on the basis of the following protected attributes . . .' [49] Similar to lists of protected traits found in Facebook's and YouTube's policies, TikTok includes the following identity characteristics: race, ethnicity, national origin, religion, caste, sexual orientation, sex, gender, gender identity, serious disease, disability, [and] immigration status.' [50]

TikTok also prohibits incitement to violence, which it defines as 'advocating for, directing, or encouraging other people to commit violence.' [51] Notably, TikTok does not allow 'threats of violence or incitement to violence on our platform that may result in serious physical harm.' [52]

---

[46] Daniel Ruby: '37 + TikTok Statistics For Marketers In 2023' available at https://www.demandsage.com/tiktok-user-statistics/, accessed on 29 May 2023.

[47] TikTok Community Guidelines available at www.tiktok.com/community-guidelines/en/, accessed on 19 May 2023.

[48] TikTok Community Guidelines on Safety and Civility available at www.tiktok.com/community-guidelines/en/safety-civility/, accessed on 19 May 2023.

[49] Ibid.

[50] Ibid.

[51] Ibid.

[52] Ibid.

TikTok's focus on 'serious physical harm' to the exclusion of other forms of harm is worthy of criticism, as this provision lends itself to manipulation. Users – on the basis of this provision – would be allowed to call for or encourage other forms of violence, such as cyber-bullying; and such content would be approved as it does not constitute advocacy of violence which may result in 'serious physical harm.' [53] This provision can thus be improved either by removing the qualifications of 'serious physical' forms of harm, or by providing for other forms of harm which similarly should be guarded against.

A further prohibition on content which TikTok provides is a prohibition against hateful ideologies. Examples of such ideologies (listed by TikTok) are, inter alia, white supremacy, misogyny and antisemitism. [54] TikTok's provision of examples in this regard – and its inclusion of a prohibition against hateful ideologies – are admirable as it represents a step further in prohibiting the spread of hatred, through ideologies. However, it would benefit users if hateful ideologies were defined and/or an exhaustive list provided.

TikTok's Policy on Misinformation and Elections

TikTok defines misinformation as 'anything including misleading, false or inaccurate content.' [55] TikTok's policy on misinformation directly quotes their community guidelines which contain a paragraph on misinformation specifically as it relates to all content on the platform:

> 'We do not allow inaccurate, misleading, or false content that may cause significant harm to individuals or society, regardless of intent. Significant harm includes physical, psychological, or societal harm, and property damage.' [56]

[53] Ibid.

[54] Ibid.

[55] Ibid.

[56] Ibid.

TikTok's explicit recognition and addressing of all forms of harm can be contrasted with Facebook's specification of physical harm and YouTube's focus on real-world harm. TikTok's version is all-encompassing and thus provides greater protections against different forms of harm, as opposed to the other ad-tech companies being researched. According to TikTok, 'significant harm' includes severe forms of:

- **'Physical injury and illness, including death;**
- **Psychological trauma;**
- **Large-scale property damage; and**
- **Societal harm, including undermining fundamental social processes or institutions, such as democratic elections, and processes that maintain public health and public safety.'** [57]

[57] Ibid.

The platform collaborates with an independent fact checking system and previously fact-checked claims to monitor the content advertised on their platform. It also specifically prohibits the following kinds of misinformation:

- 'Misinformation that poses a risk to public safety or may induce panic about a crisis event or emergency, including using historical footage of a previous attack as if it were current, or incorrectly claiming a basic necessity (such as food or water) is no longer available in a particular location;
- Medical misinformation, such as misleading statements about vaccines, inaccurate medical advice that discourages people from getting appropriate medical care for a life-threatening disease, and other misinformation that poses a risk to public health;
- Climate change misinformation that undermines well-established scientific consensus, such as denying the existence of climate change or the factors that contribute to it;
- Dangerous conspiracy theories that are violent or hateful, such as making a violent call to action, having links to previous violence, denying well-documented violent events, and causing prejudice towards a group with a protected attribute;
- Specific conspiracy theories that name and attack individual people;
- Material that has been edited, spliced, or combined (such as video and audio) in a way that may mislead a person about real-world events;
- General conspiracy theories that are unfounded and claim that certain events or situations are carried out by covert or powerful groups, such as "the government" or a "secret society";
- Unverified information related to an emergency or unfolding event where the details are still emerging; and
- Potential high-harm misinformation while it is undergoing a fact-checking review.' [58]

[58] TikTok Community Guidelines on Integrity and Authenticity available at www.tiktok.com/community-guidelines/en/integrity-authenticity/, accessed on 11 May 2023.

Notably, TikTok prohibits misinformation which causes prejudice towards a group with a protected attribute. Again, this is contrastable with Facebook and YouTube's policies – as outlined above – which render misinformation as violating, only where the misinformation is linked or contributes to imminent harm. TikTok's provision for prohibitions against hate speech, through misinformation, is thus laudably a more protective and realistic provision, as it provides for the overlap between categories of violative content.

Misinformation in relation to Elections

TikTok has their own Civic and Election Integrity Policy for Misinformation. More prominently, unlike their other social media counterparts, TikTok does not allow paid political promotion, political advertising, or fundraising by politicians and political parties (for themselves or others):

> 'TikTok has long prohibited political advertising, including both paid ads and creators being paid to make branded political content. This also includes the use of promotional tools available on the platform, like Promote or TikTok Shop. In addition to our political advertising content policy, we also impose prohibitions at the account level. This means that accounts we identify as belonging to politicians and political parties have their access to advertising features turned off.' [59]

[58] TikTok Community Guidelines on Integrity and Authenticity available at www.tiktok.com/community-guidelines/en/integrity-authenticity/, accessed on 11 May 2023.

[59] Vanessa Pappas 'Combating Misinformation and Election Interference on TikTok' Newsroom 16 Aug 2019 available at newsroom.tiktok.com/en-us/combating-misinformation-and-election-interference-on-tiktok, accessed on 11 May 2023.

Prohibitions against misinformation related to elections is also detailed on TikTok and includes the following:

- **'How, when, and where to vote or register to vote;**
- **Eligibility requirements of voters to participate in an election, and the qualifications for candidates to run for office;**
- **Laws, processes, and procedures that govern the organization and implementation of elections and other civic processes, such as referendums, ballot propositions, and censuses;**
- **Final results or outcome of an election; and**
- **Unverified claims about the outcome of an election that is still unfolding and may be false or misleading.'** [60]

In combating misinformation in relation to elections on the app, TikTok has promised to update their policies on misleading content to clarify what is and is not allowed. They are expanding their fact-checking partnerships to include not only content advisory councils, but with Politifact and Lead Stories to monitor and remove misinformation related to elections. [61] TikTok has also partnered with the U.S Department of Homeland Security in establishing the Countering Foreign Influence Task Force (CFITF) as part of the National Risk Management Center within the Cyber and Infrastructure Security Agency. Their role is to counter any threat to foreign involvement with U.S. elections specifically. [62] Although these interventions are USA-specific, their inclusion provides a catalyst for advocacy calling for TikTok to implement similar, localised mechanisms in other countries in which it operates.

[60] Government, Politician, and Political Party Accounts on Tiktok Help Center available at support.tiktok.com/en/using-tiktok/growing-your-audience/government-politician-and-political-party-accounts, accessed on 12 May 2023.

[61] Op cit note 45.

[62] Ibid.

X/Twitter

X/Twitter's Policy on Hateful Conduct

X/Twitter has 450 million monthly active users as of 2023 [63] and has two relevant policies: one on 'violent speech' and the other on 'hateful conduct.'

The Hateful Conduct Policy states that 'you may not directly attack other people on the basis of race, ethnicity, national origin, caste, sexual orientation, gender, gender identity, religious affiliation, age, disability, or serious disease.' [64] The following categories of conduct are considered to be a violation of the policy: hateful references, incitement, slurs and tropes, dehumanization, hateful imagery, and hateful profiles. [65]

Notably, both hateful imagery and dehumanisation are prohibited where it is based on categories on a list. As with YouTube's policy, this list does not contain important categories which are provided for in Section 9 of the Constitution. [66] These include: colour, marital status, birth, conscience and belief.

[63] Daniel Ruby '58+ Twitter Statistics for Marketers In 2023 (Users & Trends)' available at https://www.demandsage.com/twitter-statistics/, accessed on 29 May 2023.

[64] Twitter's Policy on Hateful Conduct available at help.twitter.com/en/rules-and-policies/hateful-conduct-policy, accessed on 19 May 2023.

[65] Ibid.

[66] Section 9(3) of the Constitution.

X/Twitter's Policy on Violent Speech

The Violent Speech Policy states that users 'may not threaten, incite, glorify, or express desire for violence or harm.' [67] Content that violates this policy includes: violent threats, wishes of harm, incitement of violence, and glorification of violence. If a user violates this policy, Twitter claims that their account will be 'immediately and permanently [suspended].' [68]

Importantly, X/Twitter states that it will 'make sure to evaluate and understand the context behind the conversation before taking action.' [69] This wording is noteworthy, as it differs from Facebook's feeble statement that it 'tries to consider context and language.' [70] The certainty of the language used in X/Twitter's policy is thus a better standard against which ad-tech companies can be held accountable where content violates their policies.

Confusingly, X/Twitter makes exceptions for violent and hateful entities or affiliation with those entities where it can determine that 'they are state or governmental entities, including those that have representatives elected to public office.' [71] This provision creates a huge gap in relation to accountability for government role-players. It seems to provide immunity where the hateful entity spewing violent vitriol is a public official – the dangers of such allowance can be seen in the storming of the Capitol during and post the 2021 elections in the USA, [72] which depicts the necessity for this provision to be tightened or removed. Moreover, the fact that governmental entities are more likely to have more followers on social media and thus greater influence, bolsters the argument that this exempting provision can more easily be used to incite violence and spread misinformation and/or hate through governmental entities – with no accountability therefor, if the provision is not removed.

---

[67] Twitter's Policy on Violent Speech available at help.twitter.com/en/rules-and-policies/violent-speech, accessed on 19 May 2023.

[68] Ibid.

[69] Ibid.

[70] Op cit note 16.

[71] Ibid.

[72] Op cit note 3.

X/Twitter's Policy on Misinformation

X/Twitter has a two-tier misinformation policy that makes a distinction between synthetic & manipulated media and crisis misinformation.

1. Synthetic and Manipulated Media Policy

X/Twitter provides that: 'You may not share synthetic, manipulated, or out-of-context media that may deceive or confuse people and lead to harm ("misleading media").' [73] In addition, X/Twitter may 'label Tweets containing misleading media to help people understand their authenticity and to provide additional context.' [74]

In order for content with 'misleading media' (including images, videos, audios, gifs, and URLs hosting relevant content) to be labeled or removed under this policy, it must:

- 'Include media that is significantly and deceptively altered, manipulated, or fabricated, or
- Include media that is shared in a deceptive manner or with false context, and
- Include media likely to result in widespread confusion on public issues, impact public safety, or cause serious harm.' [75]

[73] Twitter's Policy on Synthetic and Manipulated Media available at help.twitter.com/en/rules-and-policies/manipulated-media, accessed on 12 May 2023.

[74] Ibid.

[75] 'How We Address Misinformation on Twitter' Twitter available at help.twitter.com/en/resources/addressing-misleading-info, accessed on 19 May 2023.

2. Crisis Misinformation Policy

In times of crisis, X/Twitter's policy aims to take action against any accounts that produce content that could incite harm for populations actively facing a crisis. The scope of this policy also includes international armed conflict. They define crisis as 'any situation in which there is a widespread threat to life, physical safety, health, or basic subsistence that is beyond the coping capacity of individuals and the communities in which they reside.' [76]

- **In order for conflict-related content to be considered violative under this policy, it must:**
- **'advance a claim of fact, expressed in definitive terms;**
- **be demonstrably false or misleading, based on widely available, authoritative sources; and**
- **be likely to impact public safety or cause serious harm.'** [77]

X/Twitter's specific provision for conflict-related content is necessary and provides an example of how ad-tech companies can tailor their policies to address specific situations within which the spread of hatred and misinformation, and the incitement of violence, is more likely.

[76] Twitter's Policy on Crisis Misinformation available at help.twitter.com/en/rules-and-policies/crisis-misinformation, accessed on 10 May 2023.

[77] Ibid.

X/Twitter aims to prioritise limiting the interaction of content that violates this policy especially from users that have larger audiences or are from verified accounts. Likes, Retweets, or engagement of any form with said tweet will be disabled as to prevent the spread of further misinformation that could be harmful. This can be contrasted with its policy on violent speech, that specifically exempts governmental entities (as outlined above) – which arguably have larger audiences – from having hateful and violent content removed. This contradiction is curious, and provides the basis for arguing that X/Twitter takes guarding against the spread of violence by governmental entities less seriously than it does the spread of misinformation. This gap will be addressed in the conclusion of this report. X/Twitter's policy also ensures that visibility of misinformed content will be reduced to avoid spreading to larger audiences. They aim to reduce visibility by:

- 'Making Tweets and Retweets from those accounts ineligible for recommendation in certain parts of the X/Twitter product (such as top Search results)
- Displaying Replies from the account in a lower position in conversations
- Excluding Tweets by the account itself in email or in-product recommendations.' [78]

Like Google, they also have a strike system for users who violate this policy. If you receive two notices within a 30-day period, your account will be suspended for 12 hours. Three (3) or more notices applied within a 30-day period, your account will be suspended for seven (7) days. Users also have the option to submit an appeal if they believe their account was locked in error. [79]

[78] Ibid.

[79] Ibid.

X/Twitter's Civic Integrity Policy

X/Twitter's misinformation policy regarding elections prohibits the spread and publishing of content that has misleading or false information about civic processes and the platform seeks to label and reduce visibility of content as it applies.

They define 'civic processes 'as 'major referenda and ballot initiatives, political elections, and censuses.' [80] The violation of this policy can be reduced to four categories. The first category is Misleading Information about how to participate. This could look like an advertisement or a tweet informing users that they can vote in an election by simply texting a code from their phone in a jurisdiction where it is not possible. The next category is 'Suppression and Intimidation'. For example, content spread that polling lines have closed or police activity is ongoing at certain voting sites would fall under this category. Next is the 'Misleading Information about Outcomes' category. Content spouting claims that undermine the integrity of the process of the elections such as spreading unverified claims of election rigging would fall under this scope. Lastly, there is the 'False or Misleading Affiliation' category. This would involve users creating content falsely representing themselves or their company's affiliation with a candidate, elected official, political party, electoral authority, or government entity. [81]

[80] Twitter's Policy on Civic Integrity available at help.twitter.com/en/rules-and-policies/election-integrity-policy, accessed on 11 May 2023.

[81] Ibid.

III Accountability For Non-Compliance With Policies

This report investigated what mechanisms each ad-tech company has in place for instances in which the platform violates its own content moderation policies. For example, where the platform approves content that is in violation of a policy. None of the ad-tech companies researched provide a mechanism through which the company itself can be held accountable for such violations. Instead, users can ask for either a review or appeal of the ad approval process.

IV Content Moderation And Safe-Guarding Tools

Based on the research performed, it appears that the four ad-tech companies' moderate content primarily through (1) using artificial intelligence (AI) to detect community guidelines violations; (2) human reviewers, and (3) relying on users to report harmful content. According to one report, Facebook, YouTube, and X/Twitter all employ 'an army' of content moderators to review content.[82] The digital content moderation industry is thus a booming and growing economic sector.

However, reports show that social media companies dedicate more budget and energy to fighting hate and misinformation in the West than the rest of the world. [83] According to an article in The Washington Post:

> 'although the United States comprises less than 10 percent of Facebook's daily users, the company's budget to fight misinformation was heavily weighted toward America, where 84 percent of its "global remit/language coverage" was allocated. Just 16 percent was earmarked for the "Rest of World," a cross-continent grouping that included India, France and Italy.' [84]

---

[82] Iulia-Cristina Uță 'Digital Content Moderation Industry Expected to Reach $13.60B by 2027' Brand Minds 17 Feb 2022 available at https://brandminds.com/digital-content-moderation-industry-expected-to-reach-13-60b-by-2027/, accessed on 19 May 2023.

[83] Cat Zakrzewski, Gerrit De Vynck, Nina Masih and Shibani Mahtani 'How Facebook Neglected the Rest of the World, Fueling Hate Speech and Violence in India' The Washington Post 25 Oct 2021 available at www.washingtonpost.com/technology/2021/10/24/india-facebook-misinformation-hate-speech/, accessed on 19 May 2023.

[84] Ibid.

The depiction above of Facebook's prioritisation of the US over other geographical locations is worrying and begs questions about the platform's integrity in relation to the implementation of its policies – particularly surrounding misinformation – consistently. This will be further explored below.

Facebook

Digital Content Moderators

In 2022, Facebook employed over 15,000 moderators. According to the cited article, a moderator at Facebook 'can review between 700 and 2000 posts a day.' [85] However, Facebook does not always employ a sufficient number of non-English speaking moderators; and when it does, these moderators are not necessarily people with progressive agendas and/or views. [86] In 2019, the Washington Post published an article outlining examples of failed moderation on Facebook that fuelled hate speech and violence in India. The reporters got hold of internal company documents that 'reveal that Facebook has meticulously studied its approach abroad – and was aware that weaker moderation in non-English-speaking countries leaves the platform vulnerable to abuse by bad actors and authoritarian regimes.' [87]

Internal reports at Facebook show that the company has failed to develop AI software to moderate posts in Hindi and Bengali and detect hate speech, even though Hindi is the fourth most spoken language in the world. [88] Internal reports also suggest that Facebook is only successful at detecting around six percent of hate speech content posted in Arabic. [89]

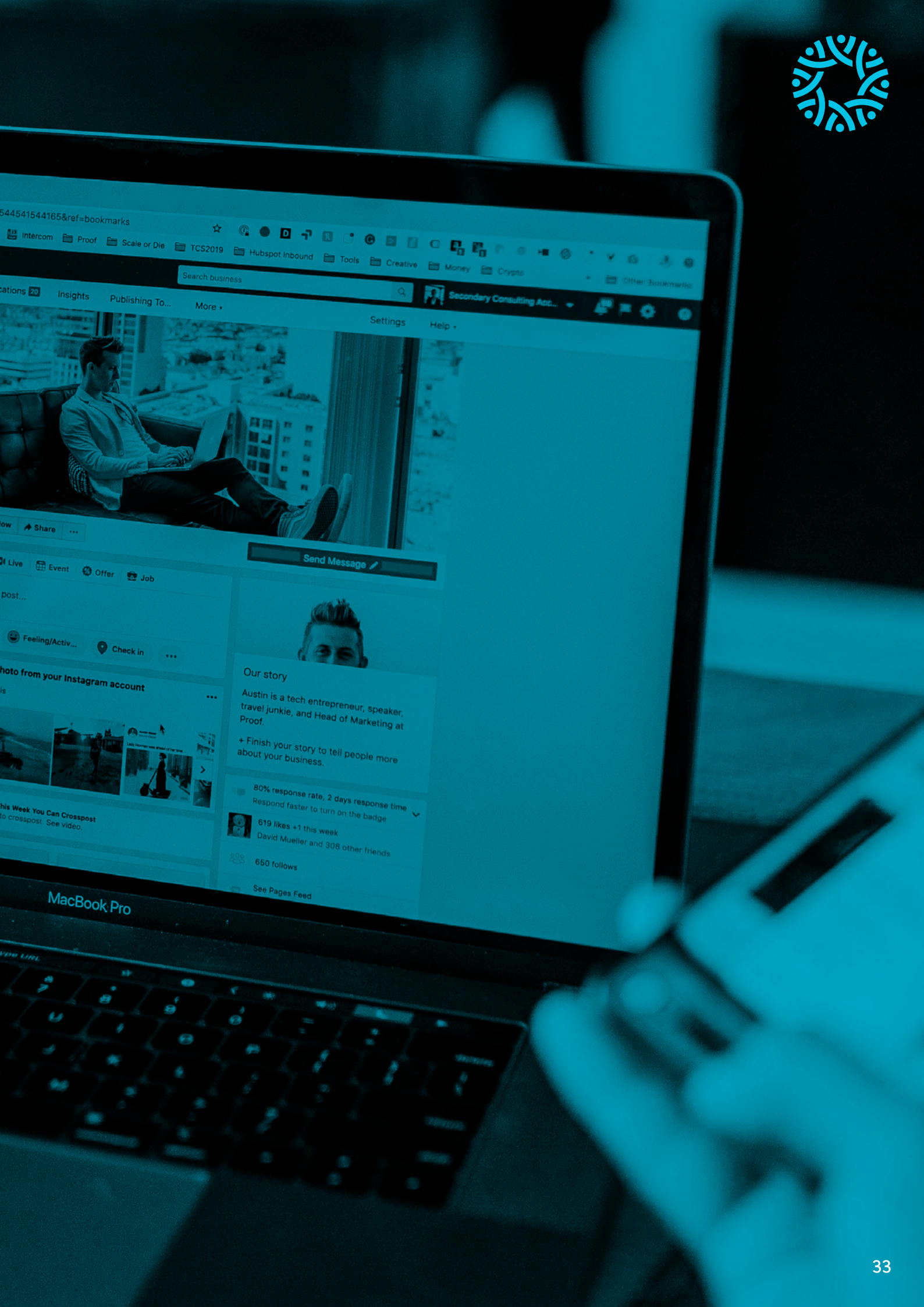
[85] Ibid.

[86] Op cit note 2.

[87] Ibid.

[88] Ibid.

[89] Mark Scott 'Facebook Did Little to Moderate Posts in the World's Most Violent Countries' Politico 25 Oct 2021 available at www.politico.eu/article/facebook-content-moderation-posts-wars-afghanistan-middle-east-arabic/, accessed on 19 May 2023.


Intercom
Proof
Scale or Die
TCS2019
Hubspot Inbound
Tools
Creative
Money
Crypto
Search business
Insights
Publishing To...
More
Settings
Help
Share
Send Message
Live
Event
Offer
Job
Feeling/Activ...
Check in
Our story
Austin is a tech entrepreneur, speaker, travel junkie, and Head of Marketing at Proof.
+ Finish your story to tell people more about your business.
80% response rate, 2 days response time
Respond faster to turn on the badge
619 likes +1 this week
David Mueller and 308 other friends
650 follows
See Pages Feed
MacBook Pro

Artificial Intelligence

Over three million items are typically reported a day by AI and users.[90] In addition to content moderators, Facebook uses AI for automatic detection of harmful content, and also relies on (and encourages) users to manually report content that violates the community guidelines. [91] The benefits of relying on AI for content moderation are the scope of content that AI is able to review, and the fact that it reports content proactively, before many users are exposed to the harmful speech. The potential of using AI in content moderation will be explored further in the Recommendations section of this report.

Even though Facebook boasts the number of posts AI detects that violate community standards, claiming the software removes over 90% of posts containing hate speech, reporters have uncovered private internal communication within the company that suggest the actual figure of hate speech removed from the platform is around three to five percent. [92]

The AI software that Facebook uses is a 'new Transformer architecture called Informer.' [93] Facebook boasted in their quarterly Community Standards Enforcement Report that in 2020, Informer successfully detected 94.7% of all of the hate speech removed from the platform. [94] However, according to a report published in Wired – a monthly magazine focusing on the effects of technology on society [95] – this number is misleading. [96]

---

[90] Ibid.

[91] Noah Giansiracusa 'How Facebook Hides How Terrible It Is with Hate Speech' Wired 15 Oct. 2021 available at www.wired.com/story/facebooks-deceptive-math-when-it-comes-to-hate-speech/#:~:text=There%20are%20two%20ways%20that,try%20to%20detect%20it%20automatically, accessed on 19 May 2023.

[92] Op cit note 68.

[93] 'How Facebook Uses Super-Efficient AI Models to Detect Hate Speech' available at ai.facebook.com/blog/how-facebook-uses-super-efficient-ai-models-to-detect-hate-speech/, accessed on 20 May 2023.

[94] Ibid.

[95] Jason Zhu 'Breaking Down Wired Magazine' available at https://medium.com/@jasonlzhu/a-6ef71ab7ef04, accessed on 29 May 2023.

[96] Op cit note 68.

This number reports the 'proactive rate' which is 'the number of hate speech items taken down that Facebook's AI detected proactively, divided by the total number of hate speech items taken down.' [97] The article explains that 'all it really measures is how big a role algorithms play in hate-speech detection on the platform.' [98] Facebook employees researching user experience abroad in 2019 found that often, content shown to users in lower-income countries differs significantly from what the algorithm shows users in the United States. [99]

The employees set up a fake account to conduct their research, which was conducted during the months when the violence in Kashmir was escalating due to the conflict between India and Pakistan regarding territory. They found that soon after registering the account, 'without any direction from the user, the Facebook account was flooded with pro-Modi propaganda and anti-Muslim hate speech.' [100]

During the same time period, Facebook users throughout India reported seeing floods of anti-Muslim posts on the platform. One said Kashmiris were 'traitors who deserved to be shot.' [101] A college student in India reported that his classmates 'used these posts as their profile pictures on Facebook-owned WhatsApp.' [102] Seeing such hateful and targeted messages across social media rightfully made this student fear for his safety and life, as they added fuel to the very real acts of violence targeting Kashmiris throughout the country.

[97] Ibid.

[98] Ibid.

[99] Op cit note 80.

[100] Ibid.

[101] Ibid.

[102] Ibid.

In 2018, the New York Times reported that the Myanmar military used Facebook as a tool to spread Islamophobia and gain support for the genocide against the Rohingya.[103] Members of the military posing as pop culture fans flocked to Facebook and filled feeds with anti-Muslim propaganda. In one example, they spread a false story of a Muslim man raping a Buddhist woman. Human rights organizations believe the posts are responsible for inciting violence and contributing to mass migration. Many of the accounts responsible for spreading this misinformation, hatred, and violence, went on undetected by Facebook.[104]

YouTube

YouTube relies on a combination of AI and human reviewers. Similarly, to Facebook, viewers are able, and encouraged, to flag harmful content.[105]

In 2022, YouTube employed 10,000 content moderators globally.[106] YouTube explains that '(our) reviewer teams remove content that violates our policies and age-restrict content that may not be appropriate for all audiences. Reviewers' inputs are then used to train and improve the accuracy of our systems at a much larger scale.'[107]

Additionally, YouTube tries to proactively moderate content as soon as it emerges by anticipating it. The company states that '(our) Intelligence Desk monitors the news, social media, and user reports to detect new trends surrounding inappropriate content and works to make sure our teams are prepared to address them before they can become a larger issue.'[108] It would be worthwhile to determine which geographical locations YouTube's Intelligence Desk monitors, and which it neglects.

---

[103] Paul Mozur 'A Genocide Incited on Facebook, with Posts from Myanmar's Military' The New York Times 15 Oct 2018 available at www.nytimes.com/2018/10/15/technology/myanmar-facebook-genocide.html, accessed on 19 May 2023

[104] Ibid.

[105] YouTube Community Guidelines & Policies on How YouTube Works available at www.youtube.com/howyoutubeworks/policies/community-guidelines/#:~:text=YouTube%20takes%20action%20on%20other,%2C%20scientific%2C%20or%20artistic%20purpose, accessed on 20 May 2023.

[106] Op cit note 65.

[107] Op cit note 74.

[108] Ibid.

Organisers from Viet Check – a fact-checking network of volunteers set up to empower Vietnamese Americans with fact-checked information and combat misinformation [109] – believe that YouTube does not consider non-English violative content as problematic. [110] As a solution, Viet Check has offered to find translation services for non-English videos, but an email exchange between a Viet Check organiser and a YouTube employee illustrates YouTube's reluctance to partner with the organisation to combat disinformation.[111] This is an example of one way in which the social media platform fails to take the moderation of non-English content seriously; and is thus comparable to Facebook's prioritisation of English over other languages.

TikTok

Similarly, to Facebook and YouTube, TikTok employs a combination of human moderators, AI software, and encourages users to flag harmful content.
TikTok explains that their moderation seeks to ensure that appropriate content is served to the right viewers: 'Our algorithms are designed with trust and safety in mind. For some content, we may reduce discoverability, including by redirecting search results, or making videos ineligible for recommendation in the For You feed.' [112] TikTok's tailoring of its algorithms in this regard is something that should be emulated by other ad-tech companies, where violative content has been found.

Similarly, other ad-tech companies can follow TikTok's example in relation to language moderation, as the social media platform provides for content moderation in more than 70 languages. [113]

---

[109] 'About Viet Check' available at https://vietfactcheck.org/about/#:~:text=We%20are%20a%20network%20of,in%20our%20Vietnamese%20American%20communities, accessed on 29 May 2023.

[110] Kate Lý Johnston 'Young Vietnamese Americans Say Their Parents Are Falling Prey To Conspiracy Videos' available at https://www.buzzfeednews.com/article/katejohnston2/vietnamese-american-youtube-misinformation-covid-vaccine, accessed on 29 May 2023.

[111] Ibid.

[112] TikTok Community Guidelines available at www.tiktok.com/community-guidelines/en/, accessed on 19 May 2023.

[113] TikTok 'Our Approach to Content Moderation' available at https://www.tiktok.com/transparency/en-us/content-moderation/, accessed on 29 May 2023.

None of the other policies of the ad-tech companies researched in this report contain a provision which elucidates how many languages it moderates content in. However, although this inclusion by TikTok is laudable, it arguably still falls short of what is needed in order to combat misinformation globally. TikTok is available in over 150 countries [114] and its provision for content moderation in just 70 languages thus represents less than half of the countries it operates it, let alone providing for the diverse range of languages spoken within one country. Thus, while TikTok's language provision is valuable, it is arguable that it still does not suffice to combat misinformation in languages which are not English, and targeting users who are non-English speakers.

X/Twitter

X/Twitter stands out as relying far less than the other companies on human reviewers. While the company does encourage users to report harmful content, and uses chat moderators, after Elon Musk took over, X/Twitter got rid of certain categories of manual reviewers and now relies more on AI to moderate content. [115] X/Twitter's reliance on AI arguably renders it more open to criticism where violative content is not detected, on the basis that AI software is more efficient and reliable than human moderators.

A Western-Centric Approach to Content Moderation

The standards and norms used to develop content moderation are often based on Western ideologies and cultures, which in turn end up unfairly affecting social media users in the Global South. [116]

---

[114] TikTok Statistics – Updated Mar 2023 available at https://wallaroomedia.com/blog/social-media/tiktok-statistics/#:~:text=TikTok%20is%20available%20in%20over,in%20the%20United%20States%20alone, accessed on 29 May 2023.

[115] Katie Paul and Sheila Dang 'Exclusive: Twitter Leans on Automation to Moderate Content as Harmful Speech Surges' Reuters 5 Dec 2022 available at www.reuters.com/technology/twitter-exec-says-moving-fast-moderation-harmful-content-surges-2022-12-03/, accessed on 19 May 2023.

[116] Patricia Waldron, Ann Cornell 'One-Size-Fits-All Content Moderation Fails the Global South' Cornell Chronicle 13 Apr 2023 available at news.cornell.edu/stories/2023/04/one-size-fits-all-content-moderation-fails-global-south, accessed on 19 May 2023.

Exporting these western-centric viewpoints and norms can result in the removal and discrimination of content that is deemed acceptable on the local level where it was posted from. Here are a few examples: In Bangladesh, a grandmother affectionately referred to a child as a 'black diamond' in an online post, and the post was flagged for racism, even though the Bangladeshi and American concepts of race differ greatly. [117] In another example, 'Facebook deleted a 90,000-member group that provides support during medical emergencies because it shared personal information – phone numbers and blood types in emergency blood donation request posts by group members.' [118] Research also shows inconsistent content moderation when it comes to religious posts. For instance, a photo was taken down and reported to be Islamophobic because it depicted 'the Quran lying in the lap of a Hindu goddess with the words, 'no religion teaches to disrespect the holy book of another religion.' [119] And at the same time, a user reported posts inciting violence against Hindus, but 'was notified the content did not violate community standards.' [120]

[117] Ibid.

[118] Ibid.

[119] Ibid,

[120] Ibid.

V CONTENT MODERATION IN SOUTH AFRICA

In June of 2023, the Legal Resources Centre and Global Witness conducted an investigation into the proficiency of content moderation on Facebook, YouTube and TikTok in detecting hate speech on their respective platforms.

Each social media platform has a verification process for adverts that users want to submit on the platform. In order to submit adverts for verification, the investigators set up fake Facebook, YouTube and TikTok accounts. From these accounts, 38 adverts were submitted for verification on each social media platform, with the scheduled date for posting two weeks later. The advertisements appeared in four South African languages – English, Afrikaans, isiZulu and isiXhosa – and were scheduled for posting on each platform at a later date. Critically, the method allowed for testing whether the platform would approve of the adverts, while still allowing those adverts to be retracted before they were actually posted. Thus, none of the adverts were actually posted onto any social media platform.

The adverts contained hate speech taken from real-life examples on social media in South Africa. They included calls to attack 'foreigners' and kill them; comparisons of refugees to animals; and general hateful vitriol aimed at foreign nationals. Critically, all 38 of the adverts contained content that violates all three social media platforms' policies on hate speech. The investigation focused on this form of hate speech leading up to the commemoration of International Refugee Day, in order to highlight the spread of hatred toward foreign nationals through social media; and to draw attention to real-life harms experienced by foreign nationals in South Africa, as a result of the spread of hate speech on social media platforms.

YouTube and TikTok both approved of all 38 adverts, while Facebook rejected the English and Afrikaans versions of one advert which compared migrants to cockroaches, but accepted the isiZulu and isiXhosa versions of this advert, as well as all 36 other adverts containing hate speech directed at foreign nationals.

With the United Nations indicating that South Africa is on the precipice of an explosion of xenophobic violence and given the recent spate of attacks on foreign nationals – catalysed through social media campaigns, like #OperationDudula – the need for obliterating the spread of hate speech targeted at foreign nationals on social media is crucial. Moreover, as South Africa's general elections fast approach, hatred targeted at particular groups is likely to rise and be used as a cheap political tactic. It is thus of paramount importance that social media companies are called on to equally protect users around the world generally; and, particularly, ensure that their content moderation policies are actually implemented in South Africa.

Given the plurality of languages spoken in South Africa, it is particularly important that content moderation in the country is performed by speakers of these languages and people who understand the local cultural context. What all three of the social media platforms have failed to do is disclose how many content moderators speak native languages in South Africa and to what extent the social media platforms' artificial intelligence mechanisms are trained to detect violating content in these languages. In its response to the investigation, TikTok reiterated that its content moderators are proficient in Zulu and Xhosa, and that ad content passes through multiple levels of verification before being approved. Meta indicated that 'both people and machines make mistakes' and that they 'know there will be things that (they) miss.'

Considering the rise of users on social media platforms every day – and the growing, if not invaluable role that social media plays in our access to information, in encouraging freedom of expression and in promoting global interconnectedness – it is critical that the content approved by these platforms is free of hatred, accurate and non-harmful. In order for South Africa's upcoming elections to be free and fair, it is paramount that social media platforms take accountability for their role in the spread of disinformation and hate speech, amongst others, and commit to properly resourcing content moderation so that users in South Africa – and our elections – are adequately protected.

For more information on this investigation, click here.

Conclusions

Before exploring possible avenues for intervention, it is necessary to outline the conclusions that can be gleaned from the research outlined above. For ease of reference, these conclusions will follow the themes in the order within which they are presented above.

Content Moderation Policies

There are no standardised guidelines across ad-tech companies, although there are similarities between them, one similarity being that most, if not all, of the ad-tech companies researched use terms which are not defined and thus make for unclear provisions. Further, no ad-tech company provides for prohibitions against disinformation, in addition to misinformation or makes specific provision for the overlap between categories of violative content, and all but one ad-tech company provides for forms of harm that are not physical, when prohibiting content that could lead to harm. While some of the ad-tech companies' policies provide specific prohibitions during times of crisis, none acknowledge or address the intersectionality between forms of harm during elections, as a particular event.

Accountability for Ad-Tech Companies

Although all ad-tech companies utilise AI in their content moderation, the development of content moderation is based on Western standards and norms and there are inconsistencies in the way in which content is moderated in the West, as opposed to countries forming part of the Global South and East.

VI Recommendations

Based on the research undertaken and as outlined above, the following recommendations are made:

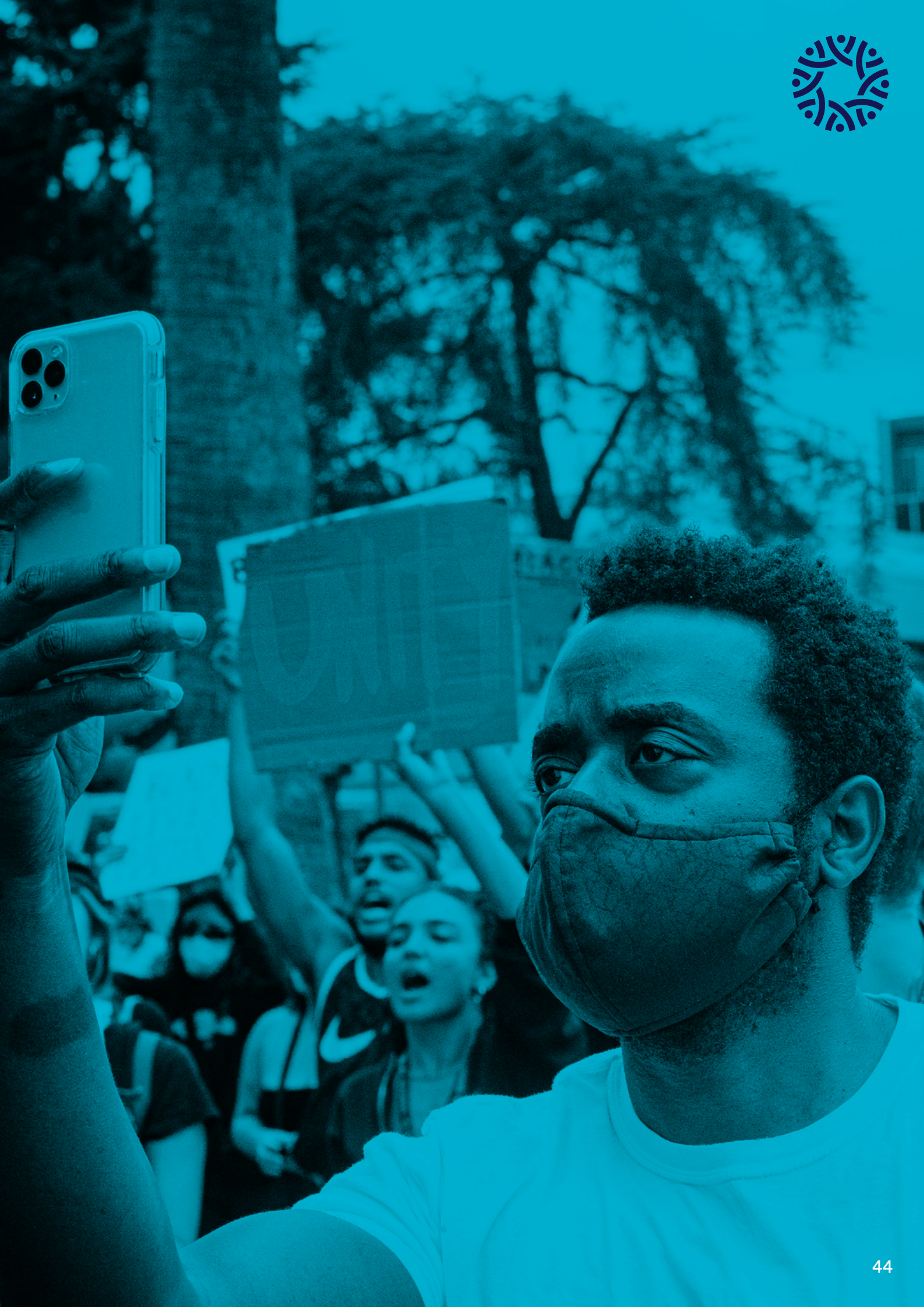
Content Moderation Policies

- **Advocacy surrounding content moderation policies can call for the following:**
  - For standardised policies and standards, based on international human rights legal standards;
  - For defined terminology accompanying standards and guidelines;
  - For examples of violative content across platforms and under each policy;
  - For protected traits to match those as outlined in South Africa's Constitution;
  - For prohibited harm to include all forms of harm;
  - For policies on disinformation, in addition to those on misinformation;
  - For specific acknowledgement of and provisions addressing the overlap between categories of violative content; and
  - For national elections to be acknowledged as an event within which there is the possibility of heightened violence, disinformation, misinformation and the spread of hatred; and for provisions that speak directly to this issue.

Content Moderation and Safe-Guarding Tools

Ad-tech companies must be called on to:
Provide statistics of implementation of their policies in the global north and south and reasons for where there are inconsistencies;

- Officially commit to implementing policies consistently;
- Disclose the number of languages in which it provides for content moderation;
- Employ a specific number of non-English speaking content moderators, based on each country in which the company operates;
- Explore ways in which AI software can be programmed with location-specific languages;
- Disclose the names of fact-checking organisations with whom it has partnered; and
- Commit to partnering with local fact-checking organisations, which must be based and have expertise in the specific country of operation.



LRC LEGAL RESOURCES CENTRE

www.lrc.org.za LRCSouthAfrica lrcsouthafrica

LRCSouthAfrica TheLRCSouthAfrica legal-resources-centre