White Paper:

A Southern and Southeast Asian Lens on Online Harmful Content and Platform Accountability during Elections

White Paper: A Southern and Southeast Asian Lens on Online Harmful Content and Platform Accountability during Elections

About

Digital Rights Foundation is a registered research-based advocacy non-governmental organization in Pakistan. Founded in 2012, DRF focuses on ICTs to support human rights, inclusiveness, democratic processes and digital governance. DRF works on issues of online free speech, privacy, data protection and online violence against women. DRF opposes any and all forms of online censorship and violations of human rights, both on-ground and online.

Contact information:

info@digitalrightsfoundation.pk
www.digitalrightsfoundation.pk

The White Paper has been researched and authored by **Zainab Durrani, Maryam Saeed and Seerat Khan.**

Executive Summary

This White Paper is envisioned as a compact overview of the harms that can and do emanate from the use of online spaces, particularly those that are facilitated by large public tech platforms.

Our intent was to look at three vital harms: hate speech, online gender based violence and the disinformation stratosphere in the context of the democratic process of elections.

We have examined existing literature, previous electoral cycles and platform policies to inform this document. Our geographical focus for this Paper was Asia and we found compelling instances from multiple South Asian and Southeast Asian countries, such as Myanmar and its highly contentious 2018 elections, that have also been included in this Paper.

Our deepdive into Meta and X policies shed light on the work these platforms claim to be doing to prop up the machinery of democracy, however we also saw some instances, such as those in India and the Philippines, where community standards did not meet the mark and had a chilling effect on the democratic process.

Our recommendations are primarily focused towards social media platforms and governments in the region. We contend that any interventions emanating from tech companies should be approached with a local context lens, so as to be able to address the most pressing concerns that we are seeing in real time. We urge states to ensure that they do not engage in over regulation of platforms and develop human rights compliant regulatory mechanisms which avoid the misuse of the law to silence dissent online.

Objectives

- This White Paper sets out to identify the current measures taken by social media platforms, including Facebook, Instagram and X, to tackle content that harms public figures (human rights defenders, journalists and politicians) or marginalized communities (women, gender and sexual minorities, and religious minorities) in the Southern and Southeast Asian region, particularly prior to the elections;

- The White Paper, based on these gaps, sets out recommendations for platform accountability to tackle online harmful content.

Introduction

The problem

The internet is undeniably an essential part of everyday life for people across the globe. Although a huge gap persists in internet access and connectivity in the Southern and Southeast Asian region[1] but the gap is narrowing with internet penetration rate increasing at an exponential rate. As a result of the increased availability of the internet, online harmful content, including online hate speech and disinformation, have emerged as new sources of domestic strife in South and Southeast Asia. In this region, online hate speech and disinformation have frequently capitalized on deeply ingrained social tensions, which many political elites have exacerbated rather than diffused for political motives[2].

In Pakistan, the use of the internet has increased exponentially in the last decade. There were 87.35 million internet users in Pakistan in January 2023[3]. The internet landscape is primarily dominated by 38.4 million Tik Tok users as of July 2023[4], 43.55 million on Facebook and 13.75 million users on Instagram. Similarly, with 692.0 million internet users, India has the highest user base for Facebook of around 314.6 million followed by 229.6 million users of Instagram[5]. The internet landscape in other countries in the region shows similar trends and points to the fact that the internet has undoubtedly had a positive impact on the citizens in these countries like it has across the globe in getting access to information, enhancing freedom of expression and opening doors to cross-country communication and opportunities. However, over the years, there is increasing evidence of the scale to which the social media user base in the demographic under question is exposed to harmful content online.

Online harmful content includes but is not limited to tech-facilitated gender-based violence (TFGBV), disinformation and hate speech. Among those who get the brunt of these attacks include public figures (human rights defenders, journalists and politicians) and marginalized communities (women, and gender, sexual and religious minorities). Research indicates that those with intersecting identities (e.g., youth, socioeconomic status, gender, ethnic and religious minority, occupation, and disability status) are at a higher risk of experiencing harmful content

---

[1] The Southern and Southeast Asian region, include South Asian countries Nepal, India, and Pakistan, as well as Southeast Asian countries, Myanmar, Vietnam, Thailand, Indonesia, the Philippines, and Singapore.

[2] Liebowitz, J., Macdonald, G., Shivaram, V., and Vignaraja, S. *The Digitalisation of Hate Speech in South and Southeast Asia: Conflict-Mitigation Approaches*. Georgetown Journal of International Affaits Conflict and Security. May 5, 2021. Available at: https://gjia.georgetown.edu/2021/05/05/the-digitalization-of-hate-speech-in-south-and-southeast-asia-conflict-mitigation-approaches/

[3] Data Reportal. Digital 2023: Pakistan. 13 Feb 2023. Available at: https://datareportal.com/reports/digital-2023-pakistan

[4] Statista. TikTok users by country. Available at: https://www.statista.com/statistics/1299807/number-of-monthly-unique-tiktok-users/#:~:text=As%20of%20July%202023%2C%20the,around%2099.8%20million%20TikTok%20users.

[5] Data Reportal. Digital 2023: India. 13 Feb 2023. Available at: https://datareportal.com/reports/digital-2023-india

—thus, amplifying their marginalized status[6]. Evidence from South Asia and Southeast Asia also overwhelmingly indicates that women are disproportionately affected by online gender-based violence[7]. Among this are women in the public sphere, in particular, India and Nepal have seen a rise in the number of female journalists who experience widespread gender-based discrimination and online harassment that poses an additional threat to their participation in a male-dominated profession[8]. Similarly, there is evidence from Pakistan based on a study which analyzed 216,849 Facebook comments directed at women politicians and 843,943 at three male politicians which shows how women politicians receive harmful content including sexualised and sexist comments whereas abuse faced by men is more on their political integrity[9].

Additionally, gender and sexual minority communities are also exposed to high rates of online hate and abuse in several countries in this region, specifically Bangladesh, Pakistan, and India[10]. In these countries, gender and sexual minority rights are condemned by the government and the abuse against them tends to be more persistent, and homophobic and misogynist in nature[11]. In Pakistan, in 2023, violent attacks, hate speech and threats against transgender community persisted and murder rates were the highest in the region. This campaign was so potent that it led to sections of the Transgender Persons (Protection of Rights) Act 2018 being struck down.[12]

Targeting of religious minority communities in this region is also quite high. Countries in the South Asia have show similar trends regarding online attacks against religious minority communities such as Pakistan, where online hate and disinformation showed a huge spike in August 2023 after two Christians were accused of blasphemy which led to uncontainable violence against the Christian community and burning down of their towns and Churches[13]. In

---

[6] Digital Rights Foundation, Covid-19 and Cyber Harassment: Policy Brief 2020. Available at: https://digitalrightsfoundation.pk/wp-content/uploads/2020/06/Covid-19.pdf

[7] Bansal, V., Rezwan, M., Iyer, M., Leasure, E., Roth, C., Pal, P., & Hinson, L. (2023). A Scoping Review of Technology-Facilitated Gender-Based Violence in Low- and Middle-Income Countries Across Asia. *Trauma, Violence, & Abuse*, *0*(0). https://doi-org.abc.cardiff.ac.uk/10.1177/15248380231154614

[8] Koirala S. (2020). Female journalists' experience of online harassment: A case study of Nepal. *Media and Communication*, 8(1), 47–56. https://doi-org.abc.cardiff.ac.uk/10.17645/mac.v8i1.2541

[9] Digital Rights Foundation. 2018. Online participation of female politicians in Pakistan's General Elections 2018. Available at: https://digitalrightsfoundation.pk/wp-content/uploads/2019/01/Booklet-Elections-Web-low.pdf

[10] Dunn S. (2020). *Technology-facilitated gender-based violence: An Overview*. Centre for International Governance Innovation. https://www.cigionline.org/publications/technology-facilitated-gender-based-violence-overview/

[11] Posetti J., Shabbir N., Maynard D., Bontcheva K., Aboulez N. (2021). *The chilling: Global trends in online violence against women journalists*. United Nations International Children's Emergency Fund (UNICEF). https://www.comminit.com/unicef/content/chilling-global-trends-online-violence-against-women-journalists

[12] Pakistan Revocation of Rights of Transgender and Gender Diverse People Must be Stopped. Amnesty International. 2023. Available at: https://www.amnesty.org/en/latest/news/2023/05/pakistan-revocation-of-rights-of-transgender-and-gender-diverse-people-must-be-stopped/

[13] Bukhari, M and Shahzad, A. Pakistan crowd vandalises churches, torches homes after blasphemy accusation. 16 August 2023. Available at: https://www.reuters.com/world/asia-pacific/pakistani-christian-community-attacked-after-blasphemy-accusation-police-2023-08-16/

Bangladesh, too, perpetrators have, over the last decade, utilized social media platforms to spread rumors and to mobilize mobs to launch violent attacks on minority groups[14].

Similarly, in Sri Lanka, extremist Sinhalese Buddhist groups have used social media to encourage violence and propagate Islamophobia. Anti-Muslim riots occurred in Kandy and Negombo in 2018 and 2019, as a result of the propagation of hate speech and rumours on Facebook in the wake of the Easter bombings. False accusations that Muslims were sterilizing Sinhalese women were part of this web campaign. During the COVID-19 pandemic, the same groups incited hate speech about the topic of cremating Muslim bodies[15].

While the groups mentioned above continue to face the brunt of online harms throughout the year, election period has been reported to see rising levels of hate, TFGBV and disinformation. A recent report by a Washington-based group highlights that anti-Muslim hate speech incidents in India averaged more than one a day in the first half of 2023 and were seen most in states with upcoming elections[16].

Considering the extent of online harms in this region and limited empirical evidence, big tech companies are under pressure from governments and civil society to take immediate and decisive steps to tackle the issue. However, their efforts are piecemeal, not enough and lack contextualisation of the region to tackle the scale of online harmful content experienced by their users in the Southern and Southeast Asian region.

[14] Roy, S., & Singh, A. K. (2023). Sociological perspectives of social media, rumors, and attacks on minorities: Evidence from Bangladesh. *Frontiers in Sociology*, *8*, 1067726. https://doi.org/10.3389/fsoc.2023.1067726

[15] Liebowitz, J., Macdonald, G., Shivaram, V., and Vignaraja, S. *The Digitalisation of Hate Speech in South and Southeast Asia: Conflict-Mitigation Approaches*. Georgetown Journal of International Affaits Conflict and Security. May 5, 2021. Available at: https://gjia.georgetown.edu/2021/05/05/the-digitalization-of-hate-speech-in-south-and-southeast-asia-conflict-mitigation-approaches/

[16] Anti-muslim hate speech in india spikes around elections. Aljazeera. 26 Sept 2023. Available at: https://www.aljazeera.com/news/2023/9/26/anti-muslim-hate-speech-in-india-spikes-around-elections-report-says

Online harms in Southern and Southeast Asian region

Harm: Technology-facilitated Gender-based violence (TFGBV)

Definition: Gender-Based Violence (GBV) consists of harmful acts directed at an individual, based on their gender[17]. Certain individuals are at a higher degree of risk of facing violence, simply due to their gender. Tech-facilitated gender-based violence (TFGBV) includes the use of internet and digital platforms to inflict, assist in inflicting or aggravating violence on women and gender and sexual minority community members (including transgender, non-binary, queer individuals)[18].

Threat: The Gender Social Norms Index (GSNI) focusing on gender norms in online spaces published by the UNDP states '*The index, covering 85 percent of the global population, reveals that close to nine out of 10 men and women hold fundamental biases against women.*'[19]. As per a global study[20] by the Intelligence Unit of the Economist, 38% of women 'reported personal experience with online violence' which does not account for the element of underreporting that can potentially significantly downplay the real numbers. Framing it closer to home, the prevalence of violence in South Asia was seen at 36%, higher than the global average, as per a study published in the Lancet which undertook the review of 366 eligible studies, comprising 2 million women in total[21].

Impact: GBV in the online realm comes as an extension of the same principles that apply offline, including those steeped in sexism, racism, religiously-motivated hate. The resulting impact flows along the same lines. A Sage Journal study[22] looking into online violence on Twitter (now known as X) against women of influence (journalists, MPs, activists) in India states that '*In addition to explicit swearing behavior, many offensive tweets directed to this group attempt to dismiss the legitimacy of women politicians based on intellectual ability and*

---

[17] Stand with her: 6 women-led organizations tackling gender-based violence (2022) unfoundation.org. Available at: https://unfoundation.org/blog/post/stand-with-her-6-women-led-organizations-tackling-gender-based-violence/?gclid=Cj0KCQjwhL6pBhDjARIsAGx8D58OFcdgOIVjSN8itqCLXDM1deiSTPlZip3om0MhRM1o9-CyNjh55IcaArGsEALw_wcB (Accessed: 24 October 2023).

[18] Hinson L., Mueller J., O'Brien-Milne L., Wandera N. (2018). *Technology-facilitated gender-based violence: What is it, and how do we measure it?* International Center for Research on Women (ICRW). https://www.icrw.org/wp-content/uploads/2018/07/ICRW_TFGBVMarketing_Brief_v8-Web.pdf

[19] Nations, U. (no date) 2023 gender social norms index (GSNI), Human Development Reports. Available at: https://hdr.undp.org/content/2023-gender-social-norms-index-gsni?gclid=Cj0KCQjwhL6pBhDjARIsAGx8D58zhsNbfHbeejc-Ep2d_X3meq5CI3SY-AFIKjDdsYwwNOniyYSG6CYaArn8EALw_wcB#/indicies/GSNI (Accessed: 24 October 2023).

[20] *Measuring the prevalence of online violence against women* (no date) *Jigsaw Infographic*. Available at: https://onlineviolencewomen.eiu.com/ (Accessed: 24 October 2023).

[21] Global, regional, and national prevalence estimates of physical or ... Available at: https://www.thelancet.com/article/S0140-6736(21)02664-7/fulltext (Accessed: 22 October 2023).

[22] Kumar, P., Gruzd, A., & Mai, P. (2021). Mapping out Violence Against Women of Influence on Twitter Using the Cyber–Lifestyle Routine Activity Theory. American Behavioral Scientist, 65(5), 689-711. https://doi.org/10.1177/0002764221989777

patriotic commitments to India.' highlighting a vital result of OGBV which is the deliberate discrediting of women who occupy some significance in the online spaces, which in this instance was politicians. A review of another group (journalists, activists etc.) showed '*more direct forms of gendered and ethnoreligious online harassment, including death threats in rare cases*', shedding light on the shift in language and intent, based on the occupation of those being attacked.

A scoping review[23] capturing the available data on GBV in low and middle income countries (LMICs) in Asia shared that cyberbullying, sexual harassment, image-based abuse, threatening and trolling or gender trolling to be the most frequently mentioned behaviours. The review also noted that real numbers for the region were likely to remain unknown. This can be attributed largely due to a combination of poor legal infrastructure and lack of faith in redressal mechanisms available in South and Southeast Asia.

Pakistan's transgender community saw a strong and targeted hate campaign[24] built on disinformation and made successful by playing on the general lack of understanding and connection with the community since September 2022. DRF's own Cyber Harassment Helpline has recorded the reactions from social media platforms including Meta and X and come up against a disappointing number of actions taken against objectionable and dangerous content that was escalated to them, with context and in detail, by the Helpline. The standard response to this has been for companies to say that this behaviour does not violate their community standards.

[23] Bansal, V., Rezwan, M., Iyer, M., Leasure, E., Roth, C., Pal, P., & Hinson, L. (2023). A Scoping Review of Technology-Facilitated Gender-Based Violence in Low- and Middle-Income Countries Across Asia.
[24] Digital Waves of Hate: The struggle continues for Pakistan's transgender community (no date) Digital Waves of Hate: The Struggle Continues for Pakistan's Transgender Community | GenderIT.org. Available at: https://genderit.org/feminist-talk/digital-waves-hate-struggle-continues-pakistans-transgender-community (Accessed: 23 October 2023).

Harm: Disinformation

Definition: Disinformation is the deliberate spread of information that is false, done with the intent of manipulation, fabrication or to build and support a particular narrative. As per a UNHCR factsheet it '*...includes malicious content such as hoaxes, spear phishing and propaganda. It spreads fear and suspicion among the population*'. A key component of this harm is gendered disinformation. This is false information that has been maliciously spread as part of a strategy to impact an individual or a community, based on their gender. In her report[25] to the UNGA's 78th Session, Special Rapporteur Irene Khan calls gendered disinformation not only a way to silence women and gender non-conforming individuals but also a method of proliferating online GBV.

Threat: Silencing, intimidating and misrepresenting appear to be the primary modes of inflicting this form of electronic violence. Disinformation can weaken trust in media and information outlets, trigger the creation and spread of bias, especially against marginalized and minority communities.

Impact: A potent example would be the 'fake news race' between India and Pakistan during the 2019 escalations, cited in UNESCO's South Asia Press Freedom report 2018-19[26]. In February of 2019, a suicide bombing in Pulwama[27],India resulted in the death of 40 police officials, which was attributed to a Kashmiri bomber. Given that Kashmir is a 'red line'[28] for both India and Pakistan, this resulted in a series of air strikes which became the most serious confrontation the two neighboring countries and historic rivals had experienced in two decades. In the days that followed, everyone tuning in to the news or consuming online content saw a travesty of dis-informative content coming from the media on both sides of the border, using dated images and videos to prop up their individual narratives. This naturally did not aid in deescalating heightened tensions and the impact went straight to the safety of citizens in both the countries.

[25] UN official documents (no date) United Nations. Available at: https://www.un.org/en/delegate/page/un-official-documents (Accessed: 24 October 2023).
[26] International Federation of Journalists. (2019). Truth vs misinformation: the collective push back: South Asia press freedom report 2018-19. Available at: https://unesdoc.unesco.org/ark:/48223/pf0000368232 (Accessed: 24 October 2023).
[27] Kashmir attack: Tracing the path that led to Pulwama (2019) BBC News. Available at: https://www.bbc.com/news/world-asia-india-47302467 (Accessed: 27 October 2023).
[28] Line of control: If the Red Line is crossed (2019) Asia Dialogue. Available at: https://theasiadialogue.com/2019/10/16/line-of-control-if-the-red-line-is-crossed/ (Accessed: 27 October 2023).

Harm: Hate speech

Definition: Discriminatory speech targeted at a specific community or people where the identity factors can be '*religion, ethnicity, nationality, race, colour, descent, gender*'[29]. The Committee on Elimination of Racial Discrimination General Recommendation 35 (GR35) on Combating Racist Hate Speech identifies five contextual factors to be considered in terms of speech that should be penalized by law. These categories include (i) content and form of speech (ii) economic, social and political climate (iii) status of speaker (iv) reach of the speech (v) objective.

Threat: Aggravation of existing elements that serve as basis for hate speech can significantly increase the impact of this online harm. The existing elements and underlying triggers can include sexist, racist, homophobic and/or religiously motivated hateful verbal acts. Mythos Labs conducted a study for UN Women Asia Pacific in 2020[30] that recorded a 168% increase in misogynistic speech since 2019, due to the COVID-19 pandemic. The data for this statistic was derived from three countries: India, Sri Lanka and Malaysia.

Impact: In terms of impacting democratic principles, a 2021 study on the status of minority rights in South Asia[31] looks at presidential elections in Afghanistan in 2019. It flags the use of 'ethnic and language-based' hate speech in the aftermath of the 2019 elections as tools of combat used by the two leading parties to discredit each others' claims about electoral fraud.

DRF's own study on the experience of religious minorities in Pakistani online spaces[32] shows the impact of hate speech, in terms of curtailing free expression. Below we have pulled two participant quotes from our report to demonstrate this imposition:

1. "*When we give our opinions on something and people don't like it, people turn it against it. We don't have laws protecting us for that.*"
2. *"Even your closest friend will say that you are different, that they are privileged and closer to God. So, even in those discussions about social issues, our views are invalidated because of our religion and you can't say anything because you know the consequences."*

[29]What is hate speech? (no date) United Nations. Available at: https://www.un.org/en/hate-speech/understanding-hate-speech/what-is-hate-speech#:~:text=Hate%20speech%20is%20%E2%80%9Cdiscriminatory%E2%80%9D%20 (Accessed: 24 October 2023).
[30]Social Media Monitoring on COVID-19 and Misogyny in Asia and the Pacific (2020) UN Women. Available at: https://asiapacific.unwomen.org/sites/default/files/Field%20Office%20ESEAsia/Docs/Publications/2020/10/ap-wps-BRIEF-COVID-19-AND-ONLINE-MISOGYNY-HATE-SPEECH_FINAL.pdf (Accessed: 24 October 2023).
[31](2021) The South Asia Collective. Available at: https://www.thesouthasiacollective.org/wp-content/uploads/2022/01/SASM2021.pdf (Accessed: 24 October 2023).
[32] Religious Minorities in Online Spaces (2021) Digital Rights Foundation. Available at: https://digitalrightsfoundation.pk/wp-content/uploads/2021/05/Religious-Minorities.pdf (Accessed: 24 October 2023).

Data analysis from a Facebook study[33] conducted by researchers from the University of Sydney and University of Queensland states that gender and sexual minorities in the Asia Pacific region were experiencing forms of hate speech that was '…*culturally specific to intersectional experiences, gender communities or ethnolinguistic groups, and some which are focussed on depriving them of powers; that is, denying targets their right to make everyday decisions*…' and essentially robbing them of autonomy and this type of hate speech was not being removed by Facebook, they observed.

An important instance to record here is the ire faced by journalists, especially women like Rana Ayyub, a vocal critic of second-time Indian Prime Minister Narendra Modi's policies. Ayyub uses X as a medium to communicate her opinions at large and faces immense backlash, to the point that her 'case' was taken up as part of an international research[34] as the reaction she had to put up with was seen as emblematic of the gender-fuelled hate that is prevalent on all online platforms now. The research reviewed 8.5 million tweets directed at her and places her intersectional (Muslim, woman, journalist) identity at the center of the report[35] that aims to build an early warning system for gender-based violence.

[33] (2021) Facebook: Regulating hate speech in the Asia Pacific. Available at: https://r2pasiapacific.org/files/7099/2021_Facebook_hate_speech_Asia_report.pdf (Accessed: 24 October 2023).
[34]What 8.5 million tweets targeting Rana Ayyub tell us about online violence & the failure to stop it (no date) Article 14. Available at: https://article-14.com/post/what-8-5-million-tweets-targeting-rana-ayyub-tell-us-about-online-violence-the-failure-to-stop-it-62d104dd20f4b (Accessed: 24 October 2023).
[35] (2023) Home | International Center for Journalists. Available at: https://www.icfj.org/sites/default/files/2023-02/Rana%20Ayyub_Case%20Study_ICFJ.pdf (Accessed: 24 October 2023).

Online Harms during Elections

The use of social media platforms has become extremely popular in elections and campaigning in the Southern and Southeast Asian region. Political parties and candidates alike use different social media platforms to access their voter banks and share their political ambitions on platforms. Big Tech platforms play a vital role on the flow of information pre, during and post elections time. The flow of information during this time has many times been harmful and has targeted marginalized groups and communities across the region. In a study titled, '*Social Media, Democracy and Fake News in Pakistan: An Analysis',* social media platforms were analyzed three months after the general elections in 2018 and it was reported that the majority of the fake news on platforms was related to international relations, politicians, judiciary and the military.[36] These topics are more susceptible to sensationalism and mis/disinformation on social media platforms. The 2018 Pakistani general elections made political parties realize the importance of social media's role during the election with parties forming social media cells to campaign on platforms.[37] DRF conducted a research titled, '*Online Participation of Female Politicians in Pakistan's General Elections 2018'* which captured the online harassment female politicians faced during the elections. In the research based on the 216,849 Facebook comments directed at the women in the dataset and 843,943 comments directed at three prominent male politicians - Imran Khan, Shehbaz Sharif and Bilawal Bhutto Zardari- women were more likely to face sexual and sexist comments on platforms as compared to their male counterparts who were targeted online because of their political integrity and not their appearance.[38]

Online harmful content has also many times resulted in offline repercussions for marginalized and vulnerable communities in the region. In India, in 2018, wide spread sharing of hoax messages on Whatsapp and Facebook resulted in a dozen lynching incidents across the country. While Facebook at the time tried to manage the problem and bring in new features, many individuals including India's Technology Minister at that time stated that the measures taken were not "*not adequate to meet the challenges of the situation.*".[39] According to the Hindutva Watch, in the first half of 2023 there were 255 documented incidents of hate speech gatherings targeting Muslims in the country and about 205 (80%) of these hate speech events took place in BJP-ruled states and union territories. Hindutva tracked these hate speech events through Hindu-far right organizations pages and individual profiles on social media with scraping data

---

[36] (2020) Rehman, H.U., Hussain,S.,& Durreshehwar. Social Media, democracy and fake news in Pakistan: An analysis. Available at: https://www.researchgate.net/publication/343286800_Social_Media_Democracy_and_Fake_News_in_Pakistan_An_Analysis (Accessed: 24 October 2023).

[37]says:, F.B. (2018) Pakistan: Upcoming General Elections and the electronic and Social Media, Asia Dialogue. Available at: https://theasiadialogue.com/2018/06/25/pakistan-upcoming-general-elections-and-the-electronic-and-social-media/ (Accessed: 24 October 2023).

[38] (2018) Digital Rights Foundation. Available at: https://digitalrightsfoundation.pk/wp-content/uploads/2019/01/Booklet-Elections-Web-low.pdf (Accessed: 24 October 2023).

[39] Iyengar, R. (2018) WhatsApp has been linked to lynchings in India. facebook is trying to contain the crisis | CNN business, CNN. Available at: https://edition.cnn.com/2018/09/30/tech/facebook-whatsapp-india-misinformation/index.html (Accessed: 24 October 2023).

from X, Instragm, Youtube, Facebook and Telegram by finding verifiable videos of these hateful events on the platforms. Furthermore 70 percent of these incidents took place in states which are scheduled to hold elections in 2023 and 2024 according to the report.[40]

According to an Amnesty International Report titled '*Myanmar: The social atrocity: Meta and the right to remedy for the Rohingya*' in 2017 it was found that security forces took forward a massive campaign on Facebook (parent company Meta) on the ethnic cleansing of Rohingya muslims in the country. It was found that Meta's algorithms amplified and promoted inciteful and hateful content further exacerbating hatred, discrimination and violence against the Rohingya community in Myanmar.[41] Similarly in Myanmar in the 2018 by-elections online hate speech directed at the Rohingya spiked during campaigning according to Athan, a Myanmar based organization promoting freedom of expression and monitoring social media.[42] Meta's algorithm system is designed to be engagement-based which powers newsfeed, ranking and recommendations of a user profile in the platform. Meta profits when users stay on the platform and by curating ads that target specific audiences. Content that is hateful in nature stays on the platform longer because it gets the most engagement and in turn has repercussions for Rohingya muslims offline.[43] A case has now been filed in Ireland's High Court against Meta for it's role in the genocide against Rohingya muslims by refugees displaced due to the genocide.[44]

In Sri Lanka, in 2019, Gotabaya Rajapaksa, presidential candidate of the Sri Lanka People's Front (SLPP) shared a post on Facebook in Sinhala language which shows a series of photographs of Buddhist statues laying on the ground and suggesting that this was due to muslims razing the Sri Lankan heritage temple.[45] This Facebook post had already been fact checked by Agence France Presse (AFP) in Sri Lanka however despite that the post was shared by political parties and candidates.[46] The Centre for Policy Alternatives (CPA) in Sri Lanka prior to the elections in 2019 already called for greater transparency and monitoring election violence

---

[40]Admin (2023) 2023 half-yearly report: Anti-Muslim hate speech events in India, Hindutva Watch. Available at: https://hindutvawatch.org/hate-speech-events-india/ (Accessed: 24 October 2023).

[41] *Myanmar: The Social Atrocity: Meta and the right to remedy for the Rohingya* (2022) *Amnesty International*. Available at: https://www.amnesty.org/en/documents/ASA16/5933/2022/en/ (Accessed: 24 October 2023).

[42]Marston, H. (2020) The hate speech threat to the 2020 election, Frontier Myanmar. Available at: https://www.frontiermyanmar.net/en/the-hate-speech-threat-to-the-2020-election/ (Accessed: 24 October 2023).

[43] Myanmar: Facebook's systems promoted violence against Rohingya; meta owes reparations – new report (2023) Amnesty International. Available at: https://www.amnesty.org/en/latest/news/2022/09/myanmar-facebooks-systems-promoted-violence-against-rohingya-meta-owes-reparations-new-report/#:~:text=Rohingya%20refugee%20youth%20groups%20have,Rohingya%20that%20they%20contributed%20to. (Accessed: 24 October 2023).

[44] Geschwindt, S. (2023) Myanmar Genocide Refugees Take Meta to Irish court over disinformation claims, TNW | Data-Security. Available at: https://thenextweb.com/news/myanmar-rohingya-meta-court-disinformation (Accessed: 24 October 2023).

[45]Sri Lankans fear violence over Facebook fake news ahead of election (2019) The Guardian. Available at: https://www.theguardian.com/world/2019/nov/11/facebook-sri-lanka-election-fake-news (Accessed: 24 October 2023).

[46] Lanka, A.S. (2019) These statues in Sri Lanka were laid on their side due to heritage restrictions at a temple -- they were not attacked, Fact Check. Available at: https://factcheck.afp.com/these-statues-sri-lanka-were-laid-their-side-due-heritage-restrictions-temple-they-were-not-attacked (Accessed: 24 October 2023).

tools for ads introduced by the platform to curb the spread of misinformation and disinformation in the country.[47]

The government of Bangladesh prior to the elections in January 2019 deployed intrusive tools for social media for surveillance and monitoring of speech on platforms. Human rights defenders and journalists have both faced the brunt of the government before the elections with enforced disappearances, censorship and limiting online speech. In 2018, acclaimed photographer Shahidul Alam was arrested under the Information and Communication Technology Act (ICT Act) for provoking unrest in Facebook comments and criticizing the acts of the government during student protests in the country. The Attorney General at the time stated that with the next parliamentary election approaching, Shahidul statement can cause further political instability which is why he shouldn't be getting bail.[48] The Prime Minister at the time Sheikh Hasina stated that Shahidul was spreading fake news and was termed as 'mentally sick' in doing so. [49]

In the Southeast Asian region, Malaysia has been using Big Tech platforms to take narratives of political parties forward particularly through platforms like TikTok and Facebook.[50] In 2022 the conservative party Parti Islam se-Malaysia (PAS) won a tremendous amount of seats through promoting racial hate speech on platforms. A report by the Centre for Independent Journalism (CIJ) in partnership with Universiti Sains Malaysia, Universiti Malaysia Sabah and University of Nottingham Malaysia monitored X, Facebook, YouTube and TikTok accounts of more than 90 key political and government actors. The study found that hate speech subjects had increased to 99,563 from October 20 to November 26, compared with about 55,000 in a pilot study carried out over a longer period from August 16 to September 30. The time period from October to November is pertinently important in Malaysia's case since this parliament was dissolved in October and unofficial campaigning started for the November 19 poll.[51]

In recent times general elections were held in Cambodia in July 2023 where social media platform's role has been quite controversial. The incumbent Prime Minister ruling since 1985 has used platforms extensively for his campaigning. Platforms which he has been using are Twitter

---

[47]Letter to facebook: Urgent need for rollout of platform affordances for greater oversight of campaign spending (2019) Centre for Policy Alternatives. Available at: https://www.cpalanka.org/letter-to-facebook-urgent-need-for-rollout-of-platform-affordances-for-greater-oversight-of-campaign-spending/ (Accessed: 24 October 2023).

[48] Adams, B. and Journalist (2022) Bangladesh: Crackdown on social media, Human Rights Watch. Available at: https://www.hrw.org/news/2018/10/19/bangladesh-crackdown-social-media (Accessed: 24 October 2023).

[49] Spicer, J. and Quadir, S. (2018) Exclusive: Bangladesh PM takes aim at photographer, critics say it is part of wider crackdown, Reuters. Available at: https://www.reuters.com/article/us-bangladesh-protests-photographer-hasi/exclusive-bangladesh-pm-takes-aim-at-photographer-critics-say-it-is-part-of-wider-crackdown-idUSKCN1MM1UD (Accessed: 24 October 2023).

[50] Fitriani and Habib, M. (2023) Social Media and the fight for political influence in Southeast Asia, – The Diplomat. Available at: https://thediplomat.com/2023/08/social-media-and-the-fight-for-political-influence-in-southeast-asia/#:~:text=Social%20media%20has%20become%20an,their%20message%2C%20and%20mobilize%20support. (Accessed: 24 October 2023).

[51] Cue (2023) Malaysian polls in November saw surge in hate speech on social media: Study, The Straits Times. Available at: https://www.straitstimes.com/asia/se-asia/malaysian-polls-in-november-saw-surge-of-hate-speech-on-social-media-study (Accessed: 24 October 2023).

and Youtube.[52] Prime Minister Hun Sen asked his followers to turn to Telegram and Tiktok after deleting his Facebook account after the platform suspended his account for six-months over recommendations by Meta's Oversight Board.[53] The Prime Minister's account had been suspended when he posted a video in January that breached the policy of the platform which contained threats of violence against opposition politicians who the Prime Minister accused were insulting his family and the ruling party on the platform. The decision by Meta resulted in the government of Cambodia stating that 22 members of the Oversight board are unwelcome in Cambodia and their decision to suspend the account has been political in nature.[54]

It is now being witnessed that the role of platforms during the elections is too big to be taken lightly and platform community policies do not always adhere to taking harmful content down right away in real time which results in repercussions and consequences for marginalized groups residing in South Asia and Southeast Asia. In times like these platforms need to hold themselves accountable and be transparent regarding their content moderation and take down policies in this part of the region.

[52] Fitriani and Habib, M. (2023a) Social Media and the fight for political influence in Southeast Asia, – The Diplomat. Available at: https://thediplomat.com/2023/08/social-media-and-the-fight-for-political-influence-in-southeast-asia/#:~:text=Social%20media%20has%20become%20an,their%20message%2C%20and%20mobilize%20support. (Accessed: 24 October 2023).

[53] Post, T.P.P. (2023) PM Hun Sen asks people to turn to telegram, Tiktok after deleting his Facebook account, Asia News Network. Available at: https://asianews.network/pm-hun-sen-asks-people-to-turn-to-telegram-tiktok-after-deleting-his-facebook-account/ (Accessed: 24 October 2023).

[54] Cambodia Bars Meta Oversight Board over PM's facebook account suspension (2023) Reuters. Available at: https://www.reuters.com/world/asia-pacific/cambodia-bars-meta-oversight-board-over-pms-facebook-account-suspension-2023-07-04/ (Accessed: 24 October 2023).

How is Big Tech Responding?

Introduction

Our focus is centered around the idea of platform accountability and inquiring how prepared social media giants are for the 2023/24 election cycle. We strongly believe that the real impact we need to see is only possible if the Big Tech platforms that command the world's social media market step up to meet the challenges head on.

For this section, we have done a deep-dive into the policies, actions and frameworks of three large scale platforms: X, Instagram and Facebook (where the latter two are owned by Meta). The decision to research these companies comes purely from a determination of which platforms see the highest virtual foot traffic. Twitter has 162* million users[55] in Asia. Facebook has six Southeast Asian countries in the top ten profiles[56] in terms of number of users per country (including India, Indonesia and Bangladesh). Instagram has a solid 353.6 million users[57] in India, Philippines, Indonesia and Thailand alone.

The methodology employed was to overlook all platforms' policies relating to key themes such that are highlighted below:

Methodology for the assessment

X: This White Paper has undertaken a combination of reviews of X's policies on misinformation, enforcement options, violent entities, abusive behaviour, crisis misinformation, synthetic and manipulated media as well as X's assortment of safety tools and its transparency reports.

Meta: Our focus on Meta will detail its policies, reports and actions around both Facebook and Instagram in terms of community policies, hate speech and misinformation policies.

Our assessment has been based on identifying the current measures and policies being implemented by X and Meta in light of real-world examples from South Asia and Southeast Asia to understand the gaps.

---

[55] Twitter users, stats, data, trends, and more - datareportal – global digital insights (no date) DataReportal. Available at: https://datareportal.com/essential-twitter-stats (Accessed: 24 October 2023).
*Figure rounded from 162.9
[56] Dixon, S.J. (2023) Facebook users by country 2023, Statista. Available at: https://www.statista.com/statistics/268136/top-15-countries-based-on-number-of-facebook-users/ (Accessed: 24 October 2023).
[57] Dixon, S.J. (2023a) Countries with most Instagram users 2023, Statista. Available at: https://www.statista.com/statistics/578364/countries-with-most-instagram-users/ (Accessed: 24 October 2023).

Overview of the Policies

In 2023, the platform formerly known as Twitter was acquired by SpaceX owner Elon Musk and the platform went over a complete transition in terms of its interface and branding. The platform is now known as X and has had significant changes to its policies to counter harmful content.

X's shift in ownership has also left many feeling less safe given its higher compliance[58] with calls for information from authoritarian regimes like India and Turkey. The clamp down on dissenting voices during Turkey's election[59] earlier this year is a vital example of the kind of leaning the platform is now displaying and the sizable impact of this element on democratic procedures is not only evident but highly predictable for the upcoming round of elections in various countries in 2023-24. The Lumen Database, a project of the Berkman Klein Centre for Internet and Society at Harvard University was previously a recipient of X's legal requests, which were uploaded directly to the database as a transparency measure. The relevant tab on the Lumen website now states '*As of April 15th, 2023, Twitter has not submitted copies of any of the takedown notices it receives to Lumen.*' This is attributed to X's third-party data sharing policies being under review, as per Lumen.

Another worrying development since the platform has come under new management is the dissolution of the Twitter Trust and Safety council[60] in December of 2022, which was formed in 2016 and comprised of independent organizations working on civil and humanitarian issues. While the council was quite far removed from the format of Meta's Oversight Board, in that it had no binding power to hold Twitter accountable for its decisions, it was an important stopgap measure the reversal of which has led to further aspersions being cast[61] on the revamp the platform has gone through since it was bought out by Elon Musk in 2022. X has taken down the link that led to the work the council had done in its tenure, limiting our direct reading of the situation.

Current policies

X is heavy-handed in the matter of policy devisement, with over 30 policy documents and 39 guidelines listed in its Rules and policies tab. X also sets out six broad categories[62] (general,

[58] Chitkara, H. (2022a) *Elon Musk would've been Twitter's corporate governance nightmare*, *Protocol*. Available at: https://www.protocol.com/elon-musk-twitter-corporate-governance (Accessed: 22 October 2023).
[59] Stein, P. (2023) *Twitter says it will restrict access to some tweets before Turkey's election*, *The Washington Post*. Available at: https://www.washingtonpost.com/technology/2023/05/13/turkey-twitter-musk-erdogan/ (Accessed: 22 October 2023).
[60] MATT O'BRIEN and BARBARA ORTUTAY AP Technology Writers (2022) *Musk's Twitter disbands its trust and Safety Advisory Group*, *https://www.wtoc.com*. Available at: https://www.wtoc.com/2022/12/13/musks-twitter-dissolves-trust-safety-council/?outputType=apps (Accessed: 22 October 2023).
[61] Ray, S. (2022) Twitter shuts down its trust and Safety Council-here's what you need to know, Forbes. Available at: https://www.forbes.com/sites/siladityaray/2022/12/13/twitter-shuts-down-its-trust-and-safety-council-heres-what-you-need-to-know/?sh=76a7ab501460 (Accessed: 24 October 2023).
[62] Rules and policies (no date) Twitter. Available at: https://help.twitter.com/en/rules-and-policies (Accessed: 24 October 2023).

platform integrity, safety and intellectual property, user guidelines and account settings) of policies through which it governs the platform, with many additional subcategory policy documents.

In terms of protection against online harms, X recognizes, amongst others, violent speech, hateful conduct and violent and hateful entities as separate categories under which specific offences fall.

Platform manipulation and spam, civic integrity and manipulated media are also risk areas X identifies under its Authenticity banner[63]. In terms of content that has been tampered with, X considers the degree to which such data has been altered, if the content is shared in a deceptive manner and whether it's likely to cause or add to confusion on public issues or impact public safety.

X boasts of a 'new'* set of Safety tools which can help make the platform safer by reducing risk factors. One of these tools is 'Reply Prompts'. This is the technology that detects the use of harsh language ('insults, strong language or hateful remarks' as per X's video explainer) and sends the X user a notification to reconsider sending out a harshly worded tweet. The tool is viewed as a measure through which 'everyone can feel safe' on the platform.

The 'control who can reply' to your tweets feature is yet another Safety tool in X's arsenal that can allow for reduction or mitigation of potential harm. This could be particularly effective for larger accounts with substantial following or accounts belonging to vulnerable communities or those talking about crucial social justice issues.

'When we see a potentially harmful tweet picking up speed, we'll add labels to slow its roll' says X, when talking about its policy on addressing misleading information. It also shares that tweets will only be removed if found to be posing 'immediate and severe harm'.

Some additional curbs to misinformation[64] are (i) allowing users of certain countries (U.S, Australia, Brazil, Spain, Philippines and South Korea) to report a post as misinformative and (ii) users (currently only from the U.S.) can write 'community notes' to give additional context as to why a post may be considered misleading.

As per the 20th transparency report published by X which is available at the X Transparency Center[65], there was an 84% decrease in the number of accounts actioned for violation of civic

[63] Our synthetic and manipulated media policy | X help (no date) Twitter. Available at: https://help.twitter.com/en/rules-and-policies/manipulated-media (Accessed: 24 October 2023).
[64] How we address misinformation on X Twitter. Available at: https://help.twitter.com/en/resources/addressing-misleading-info (Accessed: 22 October 2023).
*difficult to ascertain a timeline for this as no indicator is available for the date)
[65] *Twitter transparency center*. *Twitter*. Available at: https://transparency.twitter.com/en.html (Accessed: 22 October 2023).

integrity policy in the reporting period (July to December 2021). Below this, the takeaway shared in the report offers context by attributing these decreased numbers to correspondingly low numbers of major national elections in the U.S.

The company's Platform Manipulation report recorded a 6% increase in global spam reports, where spam can include actions like coordinated activity and artificial amplification which can serve to muddy the waters with respect to any content or stream of conversation that is the target of this disruptive behaviour.

X's website also shares an initiative, the Twitter Moderation Research Consortium ("TMRC" or the "Consortium") which was launched in late 2022, through which X shares '*large-scale datasets concerning platform moderation issues*' with a global group of members, comprising of public interest researchers from various field, who are invested in studying platform governance issues. This initiative means X no longer releases these datasets to the public but only specifically to the Consortium. The Consortium is reserved for members only and if you fall in an applicable category, you can put in an application to join it. This limits public access to the data that they themselves have contributed to producing and thus diminishes the accountability quotient of the platform.

Similarly, the curtailment of free access to X's API (application programming interface), particularly from a public research perspective puts social data behind a paywall that is too high for most to access and thus, dampens attempts to create independent and reliable literature which is the cornerstone of policy advocacy efforts.

Election involvement:

In 2022, in the lead-up to the Philippines' general elections, X partnered with the local Commission on Elections '*to amplify voter education initiatives on the policy, product and partnership front to protect the integrity of election-centric conversations on the platform and encourage healthy civic debate*.[66]

The campaign included the creation of symbolic emojis curated specifically for the Filipino voter base and supporters of democracy. Additionally, pop-ups encouraging and connecting users to access factual knowledge from authentic sources was also deployed in this timeframe.

Another element to support honest and transparent elections were prompts set to deploy when misleading information was being shared through the X platform. This applied to tweets that were misleading regarding the voting process, intended to intimidate or dissuade voters or were

[66] *Safeguarding public conversation during the 2022 Philippine election* (no date) *Twitter*. Available at: https://blog.twitter.com/en_sea/topics/events/2022/safeguarding-public-conversation-during-the-2022-philippine-election (Accessed: 22 October 2023).

being used to spread information intended to undermine voter confidence in the electoral process.

X's civic integrity policy today focuses on 4 categories: misleading information on how to participate, suppression, intimidation and false or misleading affiliation. X views civic processes as elections, censuses and referendums. The penalty for violations include restricting the visibility of the post and any interactions with it from other users. While the platform remains trigger-heavy on creation of policies such as the one above, many civil society organizations have highlighted the lack of response for X for reporting cyber crime and dealing with content removal requests.

A good example of how X may not be up to the mark in terms of follow-through of its own Information Quality initiative (launched after the scandal surrounding the 2016 U.S Presidential elections)[67] is the 2018 Malaysian elections which were hailed to be a heavily contested and intense affair between the country's two main political parties. These elections saw significant influence of the use of bots to flood the Malaysian X space with pro-government narratives[68]. X's response even then was to highlight that they were working on 'improving policies' as opposed to directly addressing the issue at hand and providing clarity around it. However, X's policies around election integrity has differed for countries in the Global North with many users being informed by the platform regarding 'malicious activity' which is not the case for countries in the Global Majority.

Meta: Facebook & Instagram

Meta, which began as the Facebook company in 2004 is the parent company for social media platforms Facebook and Instagram, it also boasts of Facebook Messenger, Whatsapp and Metaverse as part of its network. The wide range of platforms and their reach makes Meta a vital entity in our everyday lives.

Overview of the policies:

Meta employs a set of tools, such as its Facebook Community Standards and Instagram Community Guidelines as well as a set of six policy areas with multiple documents to outline acceptable behaviour on its platforms.

[67] Update on Twitter's review of the 2016 US election (no date) Twitter. Available at: https://blog.twitter.com/en_us/topics/company/2018/2016-election-update (Accessed: 24 October 2023).
[68] Seiff, A. (2018) Twitter has a big bot problem in Southeast Asia, Time. Available at: https://time.com/5260832/malaysia-election-twitter-bots-social-media/ (Accessed: 24 October 2023).

The Facebook Community standards encompass the '*Violence and Criminal Behavior, Safety, Integrity and Authenticity and Objectionable Content.*'[69] verticals spread over 20 different areas. In its 2023 Q2 (second quarter) standards enforcement report on hate speech, Facebook says it actioned 17.5 million content in connection with hate speech. Below we have listed a few country-wise instances of hate speech and other online harms where the platform ran short of curtailing the harms:

In India, Facebook failed to curb the spread of hate against caste and religious minorities by not shutting down hundreds of posts targeting the aforementioned vulnerable communities, shared a 2018 report by Equality Labs[70]. '*Over 40% of all the posts that were removed – after they reported them – were restored after a period of 90 days on average. An overwhelming majority of the posts that were restored were Islamophobic in nature.'*

A 2022 Amnesty International report[71] accused Meta of proactively amplifying anti-Rohingya content in Myanmar as discussed in the section above. Since August 2017, the Myanmar security forces have undertaken a brutal campaign of ethnic cleansing against Rohingya Muslims in the country. The report by the humanitarian watchdog called Facebook 'an echochamber for virulent anti-Rohingya content' as Myanmar too experienced the phenomenon where Facebook was the internet in its entirety for the citizens of the country and observers saw a meteoric rise in its market penetration as well as real world implications of the hate that was being churned online. The report analyzes the role Meta played and the employment of a 'move fast and break things' approach that the company may have publicly distanced itself from but still employs in practice.

A report by the Newton Tech4Dev Network[72] looked at the sweeping victory[73] gained by Rodrigo Duterte in the 2016 Philippine elections with the help of trolling armies controlled by social media influencers. The report discusses the phenomenon of 'click armies' which had a sizable impact on the democratic process in the Philippines. The report also highlights the procedure through which community-level fake account operators get paid for filling a daily quota of post engagement. This is yet another concrete example of social media platforms not wielding their very impactful setups for positive change.

Instagram's Community Guidelines contain an 8-point list, some of which cover not glorifying self-injury, being respectful, and insisting on only sharing content that belonged to the user or the

[69] Facebook Community Standards (no date) Facebook. Available at: https://www.facebook.com/business/good-questions/community-standards (Accessed: 24 October 2023).
[70] Facebook's uneven enforcement of hate speech rules in India highlighted in new study (no date) The Wire. Available at: https://thewire.in/media/facebook-hate-speech-guidelines-india-study (Accessed: 24 October 2023).
[71] Myanmar: The Social Atrocity: Meta and the right to remedy for the Rohingya (2022a) Amnesty International. Available at: https://www.amnesty.org/en/documents/asa16/5933/2022/en/ (Accessed: 24 October 2023).
[72](2018) Behind the scenes of troll accounts and fake news production in the ... Available at: https://newtontechfordev.com/wp-content/uploads/2018/02/Architects-of-Networked-Disinformation-Executive-Summary-Final.pdf?trk=public_post_comment-text (Accessed: 24 October 2023).
[73] Philippines election: Maverick Rodrigo Duterte wins presidency (2016) BBC News. Available at: https://www.bbc.com/news/world-asia-36253612 (Accessed: 24 October 2023).

user had the authority to display or share it i.e. protective of intellectual property. Other than that, as the Facebook policy page says[74], the two Meta platforms share content policies.

The Meta Transparency center lists the company's policy on Misinformation[75], where a section of the policy states: '*We also remove content that is likely to directly contribute to interference with the functioning of political processes and certain highly deceptive manipulated media.*'

The hate speech policy for Meta[76], which is one of the areas covered under the Instagram Community Guidelines as well, sets out three tiers of content which cannot be posted on the platforms, that targets a person or a group of people based on protected characteristics (race, gender, age, ethnicity etc.)

In the course of conducting research, we have determined that limited information and few inquiries are available about Meta's platform Instagram. If Meta and public interest researchers do not turn their focus towards Instagram, the opacity will continue to cloud over and draw out unnecessarily the process of contouring policies to make them more human rights compliant.

Elections involvement:

In a 2022 post on 'our approach to elections'[77] Meta states that it has tripled the number of employees it has working on safety and security bringing it up to more than 40,000 globally and has 'significantly increased' its focus on elections.

The post also lists three key areas of action:

- Preventing interference: the platform accomplishes this by taking down accounts and Pages trying to manipulate public debate and coordinating with state and non-state actors to keep emerging threats in check.

- Fighting misinformation: Meta works with more than 80 partners across 60 languages to fact-check what people post and provide context on misleading content. It also engages in content removal in a case where there are attempts to interfere with the process of voting, such as incorrect voting information, and removes calls for electoral violence.

---

[74] Community standards enforcement (no date) Transparency Center. Available at: https://transparency.fb.com/reports/community-standards-enforcement/ (Accessed: 24 October 2023).
[75] Misinformation (no date) Transparency Center. Available at: https://transparency.fb.com/en-gb/policies/community-standards/misinformation (Accessed: 24 October 2023).
[76] Hate speech (no date) Transparency Center. Available at: https://transparency.fb.com/en-gb/policies/community-standards/hate-speech/?source=https%3A%2F%2Fweb.facebook.com%2Fcommunitystandards%2Fhate_speech (Accessed: 24 October 2023).
[77] Our approach to elections (no date) Transparency Center. Available at: https://transparency.fb.com/en-gb/features/approach-to-elections/ (Accessed: 24 October 2023).

- Increasing transparency: Meta focuses on ad-buying procedures such as identity verification of political advertisers and publishing its Ads Library which houses all political ads ever posted through Meta so as to provide posterity and a clear paper trail in terms of who funds these advertisements.

While the written posts and policies sound credible in theory, it is the practice that determines the real world impact.

In the lead up to the 2019 Indian Parliamentary elections, Facebook was reported[78] to have removed 687 pages and accounts, however only one of those was that of the Bharatiya Janata Party (BJP) along with 14 of its accounts that were also taken down. This prompted the perception that the Meta platform might be giving unfair advantage to certain political parties.

For the upcoming election cycle in Pakistan[79], Meta has set up an elections operations team and is also investing in voter education. Additionally: *'Meta supported a digital civic education campaign in collaboration with the Election Commission of Pakistan (ECP), PakVoter and Shehri Pakistan, to promote information about voter rights and address key election-related digital literacy topics such as tackling misinformation.'*[80] However, as discussed above, Facebook has embroiled itself in multiple national elections and not many have gone off without a hitch. This is amplified further by the company's history of negatively impacting national-level election like it did with the Cambridge Analytica scandal[81] in 2016 during the U.S Presidential elections.

However, actions such as these grand gestures in terms of deploying resources have to be taken with a grain of salt, as evidenced by a Global Witness report[82] which tested Meta and Youtube ad policies by sending in disinformative election posters, where Meta failed to robustly safeguard democratic principles (which was the premise of its elections work) by approving half the false posters it received in 2022.

---

[78] https://scroll.in/latest/1019747/facebook-ad-policy-gave-bjp-unfair-advantage-in-indian-elections-shows-series-of-reports

[79] Shahzad, A. (2023) Pakistan sets election for January, likely minus Imran Khan, Reuters. Available at: https://www.reuters.com/world/asia-pacific/pakistan-hold-national-election-jan-not-nov-vote-commission-2023-09-21/ (Accessed: 24 October 2023).

[80] Amin, T. (2023) Meta unveils strategy aimed at protecting election integrity, Brecorder. Available at: https://www.brecorder.com/news/40265527 (Accessed: 24 October 2023).

[81] Facebook scandal affected more users than thought: Up to 87m (2021) AP News. Available at: https://apnews.com/article/e0e0df2083fe40c0b0ad10ff1946f041 (Accessed: 24 October 2023).

[82] *Facebook fails to tackle election disinformation ads ahead of tense ...* Available at: https://www.globalwitness.org/documents/20391/Facebook_fails_to_tackle_election_disinformation_ads_ahead_of_tense_Brazilian_election_EN_-_August_2022.pdf (Accessed: 22 October 2023).

Avenues for Improvement & Recommendations

Based on our evaluation and reading of the current landscape surrounding information integrity and platform policy, we conclude that there is significant work yet to be done by big tech in terms of recognizing lacunas, reorganizing internally and then effectively implementing solutions, if they wish to truly be seen as free and safe spaces for citizens of the internet. We have set down below a set of actions that can guide the way forward:

Big Tech Platforms

- Tailor interventions to local context that can address the most pressing concerns with regards to online disinformation and its harsh/harmful impact. Social media companies need to conduct a transparent reform of their existing SOPs.
- Increase their understanding of local customs in all the regions they find their presence in, to lower the risk of misdiagnosing and wrongly addressing cultural differences and the threats that might be ignored due to employing a myopic lens.
- Invest in building avenues to allow for escalated engagements with users so they do not have to be reliant only on select trusted organizations to have their concerns addressed
- Given the context of possible grievous bodily harm resulting from spread of gender based false information in patriarchally-minded societies, a company sanctioned hotline should be set up for urgent escalation of such cases.
- Look beyond basic CSR duties and see this as a serious responsibility. Tech-facilitated online violence and online GBV is growing steadily and needs immediate action to be curtailed.
- Revise content moderation policies to make them consistent with the obligations of corporations to respect and promote human rights, as set out in the UN Guiding Principles on Business and Human Rights and other established international human rights standards.
- Implement a transparent responsive appeals mechanism for content decisions which needs to be adequately resourced and accessible to ensure context-appropriate and timely redressal. Given the non-transparent and arbitrary nature of content moderation decisions, individuals who are impacted rarely have redressal mechanisms to appeal decisions made against them and little control over how their content is regulated.
- Make it a practice to release bi-annual transparency reports regarding content removal across the globe. These reports are essential in developing policy frameworks for platforms and should be available in regional languages so that more people are aware of content removal requests and policies adopted by tech platforms and there's more accountability of these platforms.

- Be mindful when deploying the use of artificial intelligence (AI) and the implications it can have for marginalized groups in countries residing in South Asia and SouthEast Asia. Human rights impact assessment mechanisms should be adopted by big tech platforms to ensure that emerging technologies are not amplifying hateful content against marginalized groups.
- Changes in privacy terms and content regulation policies by big tech platforms must be transparent and available in regional languages in the region so that more audiences are able to understand data privacy mechanisms adopted by tech platforms.
- State driven legislation to hold tech accountable can be regressive, particularly in the Global South, however it should not be used as a pretext to bypass accountability altogether. A strong and binding commitment must be made by tech platforms to lead their companies in a reasonably rights-compliant manner.
- Implement an affirmative action-based approach for dealing with cases of minority groups in the region, particularly in countries where they are highly persecuted.

Governments

- Develop regulatory models which are focused specifically on content that is expressly illegal and harmful which is clearly defined with priority categories. Use of vague terms must be avoided for regulation. Any restrictions to the right to freedom of expression must be clearly prescribed by law, pursue a legitimate aim, be necessary in a democratic society, and be proportionate to the aim pursued so that the misuse of the law to silence dissent can be avoided.
- Enact human rights-compliant legislation on digital privacy and protection after meaningful consultations with civil society and the general public. Fundamental human right to dignity and privacy must be protected for every citizen in the region.
- Include Internet education and safety courses in school curriculums. Topics such as consent, social media ethics, safety practices, and what is illegal online must be covered. This will empower the younger generation to be more confident and aware while exploring the internet.
- Work with civil society and media literacy organizations to create awareness regarding harassment, online harassment and rights around free speech online. Frequent sensitization training of law enforcement led by civil society on the importance of journalist welfare and safety need to be facilitated and supported by the government.
- Must collaborate with other countries in the region and international organizations to address the similar nature of online harms that they face and how cross-border cooperation can be useful.
- Overregulation is caustic for a democracy, especially when it intends to curb free speech. States should introspect and reconsider before approaching a problem solely through the channel of a legal framework.