Clicks, Code, and Consequences: Big Tech's Gamble with Human Lives and Election integrity in the 2024 Year of Democracy

Global Coalition for Tech Justice

Executive Summary

The beginning of 2025 was marked by a dramatic shift and the rollback of guardrails by Meta, X, Google/YouTube and TikTok after Big Tech oligarchs sided with US president Donald Trump to further enrich themselves. Meta, which owns Facebook, WhatsApp and Instagram, changed its hate speech policies and sacked fact-checkers, raising fears of fuelling widespread violence and human rights abuses just like it did in Myanmar in 2017. Meanwhile, X's Elon Musk meddled in Germany's national elections by promoting the far-right AFD party, and made Nazi salutes on Trump's inauguration day. But already before this turn for the worse, these social media companies had undermined electoral integrity and endangered human lives in dozens of jurisdictions voting in the world's most consequential elections megacycle in contemporary history.

2024 was the biggest election cycle in our lifetimes. Elections and referenda took place in 78 countries and territories on 110 unique election days. The electoral calendars of several of the world's most populous nations aligned, including India, the United States, Indonesia, Pakistan, Bangladesh, Mexico, South Africa and the European Union with its 27 member states. The total population of the countries holding elections in 2024 was around 4.16 billion– more than half of the world. This coincided with a time where there are 5.24 billion people on social media platforms globally, accounting for nearly 64% of the world's population, and 95% of the people online. For a large proportion of these people, social media is a prominent, if not primary source to access news and information, and tech companies have vouched to ensure the integrity of the content they platform.

Meta operates the world's largest social media platforms by far, with a global user base of well over 3 billion people, mostly outside its home US market, alongside Google's YouTube with 2.5 bn users, TikTok's 1.5 bn users and X's 200-250 million users. Our movement, the Global Coalition for Tech Justice, wrote to these Big Tech companies in July 2023, with a series of demands for 2024's Year of Democracy, including additional investments and transparency in election integrity plans, and for locally appropriate, rights-respecting content moderation and meaningful safety efforts. While Meta and TikTok responded in letters containing general information and met with the Coalition on a small handful of occasions; X and Google ignored our calls and refrained from any meaningful engagement over the course of the next 18 months. None of the companies stepped up to meet the demands we set for fully-resourced and transparent elections plans for all countries going to the polls.

As part of our campaign, we covered elections and corresponding unregulated illegal or harmful online speech, which went largely unaddressed by Big Tech platforms. Already in the first quarter of 2024, it became clear that there was little change or improvement in the approaches to elections adopted by the tech giants. And as usual, it was the most vulnerable people – in or from Global Majority countries – who paid the highest price for the failure of Big Tech companies to prioritise tackling tech harms over profits. In situations when the companies did act, it was often too little, too late.

The investments and resources that Big Tech companies spend in the Global Majority towards election integrity represent an inequitable fraction of their expenditures in the Global North and the US in particular. In addition, aside from the US and EU, and with the exception of large social media markets like India and Brazil, most other countries have very limited negotiating power to ensure election integrity efforts from platforms in their regions. Even then, Big Tech platforms generate risks they cannot effectively mitigate anywhere.

From Pakistan, through South Africa, Tunisia and Brazil the companies' toxic algorithm promoted posts inciting violence against minorities, activists, LGBTQ+ people and women. In India, Meta's largest market, Muslim men were beaten to death by radicalised Hindus, who believed in online conspiracies about Muslims seducing Hindu women to increase the country's Musilm population – a "great replacement" theory passionately peddled by the far-right across the world.

Unbridled and un-checked political speech didn't only affect people in the US, where presidential candidate Donald Trump spewed racist rhetoric and falsely accused migrant workers of driving up crime. In Tunisia, the authoritarian president Kais Saied used Facebook as his personal megaphone to dehumanise rights defenders and activists, before arresting and imprisoning them.

Big Tech also miserably failed to be transparent about and respond to the growing threat of AI-generated content, which predominantly targeted women in the 2024 election season. In Brazil – where platforms like X pushed back against regulatory efforts – a social media campaign using deepfake pornographic images of female candidates running in local elections sought to silence and discredit them.

The narrative of "harmful speech" or "bad actors" using platforms to wreak havoc proved immensely beneficial to BIg Tech companies, allowing them to evade scrutiny of their flawed and toxic business model. Social media companies rack up immense profits thanks to their engagement-based algorithm, which is at the centre of their business model and favors extremist,

viral and provocative content; the bigger the outrage and the hate, the higher the profits. Fact-checkers, whose number was never proportionate to the threat at hand, fought a losing battle against the deluge of disinformation and were prevented by Big Tech policies from fact-checking the politicians who were disinformation superspreaders. In this way, social media companies played a leading role in the corrosion of democracy, undermining and endangering the rights of minorities and vulnerable communities spanning the globe.

By focussing public debate on their receding investment in transparency and content moderation, tech platforms were able to effectively stay clear of any real discourse of algorithmic accountability measures to reverse this trend. **We desperately need to move the conversation squarely to the regulation of tech platforms and technologies.** Regulation must establish transparency and accountability, including explicitly addressing the internal recommendation engines which prioritise, elevate and promote some of the most harmful content–such as hate speech, misinformation, and material inciting racial violence–all under the guise of boosting "user engagement" and ultimately driving profit.

Introduction

Our movement, the Global Coalition for Tech Justice, was born against a backdrop of growing cuts to online safety and the historical neglect of tech harms to the Global Majority world. We call this the "global equity crisis" at the centre of tech accountability, whereby social media companies headquartered in the global north – and the regulators responsible for them – have been negligent when it comes to dealing with their impacts everywhere but particularly in the Global Majority. We now bring together over 250 organizations, networks and experts across 55 countries making it the world's largest movement devoted to holding tech giants to account. Focussing on the extremely uneven, profit-driven and iniquitous responses to harm by tech platforms – from disinformation, hate and incitement to manipulation of democratic processes – the Coalition articulated a set of demands as part of its Year of Democracy campaign which centered on the biggest elections year in modern history, with well over 60 countries and regions going to the polls in 2024, including some of the world's most populous – Indonesia, India, the United States (US), the European Union, and Brazil. This report looks at the performance of platforms run by Big tech companies in their responses to the speech and safety challenges posed by the election cycle of 2024, with special regard to elections held in Global Majority countries.

Background

The changes to its fact-checking and hateful conduct policies announced by Meta, which owns Facebook, Instagram and WhatsApp, in the first week of 2025 caps off over two years of rapid and steady dilution of trust and safety practices in the world's largest digital technology companies. These developments bookended a historic year in 2024 in which elections were held in over 60 countries. The impact that these decisions at large Big Tech companies have had on elections around the world has been significant.

The effects of layoffs have been more pronounced in non-English-speaking regions. As the Global Coalition for Tech Justice warned, while platforms like Meta and TikTok planned to implement strict safety measures in the US, they neglected other regions, allowing the unchecked spread of disinformation and hate speech. **This disparity reflects a prioritization of content moderation in English-speaking markets, leaving regions like South Asia, Africa, the Middle East and Latin America particularly vulnerable.** The lack of investment in linguistic and cultural expertise has further exacerbated the challenges faced by these regions during the 2024 election cycle.

This should not have happened. The 2016 US elections and revelations about the role of Facebook in facilitating violence in Myanmar in 2017 were important inflection points in the growth of an entire ecosystem of trust and safety professionals serving Big Tech companies. They focussed on election related online content on the one hand, and researchers, non-profit organisations and charitable funding designed to investigate and mitigate online harms, on the other, to emerge. While the performance of the companies was extremely uneven across the world, there were clear trust and safety gains in the form of better content moderation practices in the cycle of elections in 2019-20. But the companies chose to gradually withdraw from platform safety efforts since then, despite a growing understanding of the tech harms they caused or facilitated.

In the US, from a recognition of the scant and inadequate trust and safety protocols at platforms run by Big Tech companies in 2016, to the improvements in these standards culminating in the deplatforming of outgoing President Trump

in 2020, the rapid rollback and backsliding of these processes over the last two years led to an almost complete capitulation on moderation of illegal or harmful speech. **This tells a story of the agility with which Big Tech companies shift their positions in response to geopolitical expediencies with an alarming degree of moral flexibility.**

Beyond the US, the short history of platform accountability has been marked by Big Tech companies eager to publicize non-binding commitments to election integrity, but withholding critical details and data that would allow the public and independent observers to assess the effectiveness and fairness of those efforts. In their transparency reporting, Meta and Google provided broad overviews of election-related activities while concealing specifics about how they were executed in specific nations or regions. On the other hand, X performed even worse, providing singularly few updates on its election integrity measures, a sharp decline compared to past efforts.

There is extreme inequity in the response of Big Tech companies to trust and safety challenges across world regions. To illustrate, while Meta says it has invested more than US $13 billion in its platforms' safety and security measures since 2016, leaked documents demonstrate that in 2020 the company spent 87% of its global budget for time spent on classifying misinformation in the US, even though 90% of its users live outside English-speaking North America. **The investments and resources that Big Tech companies spend in the Global Majority towards election integrity represent an inequitable fraction of their expenditures in the Global North and the US in particular.** Company sources report that this is because of the more consequential reputational and legal risks they have faced in their home markets of the US and across the Global North where courts and institutions have greater enforcement capacity.

In this report, we tell a more comprehensive story, which covers the intersection of Big Tech platforms, in particular Meta's Facebook, Instagram and WhatsApp, Google's Youtube, X and Tiktok, and elections in a number of countries in the world, and the trends we saw in 2024 and beyond. Elections are not perfectly discrete events, much less a year with so many of them. How platforms respond to them, and then reorient themselves after it in the world whose political economy has been altered by the election are events often in flux. To that end, the scope of this report will look beyond 2024, and where needed look to events before and after it.

Despite promises by platforms to tackle online harmful speech such as Meta and Google's announcements in late 2023, there was little change or improvement in the approaches they adopted. As cases of online harmful electoral speech emerged over the election year, online platforms were largely found wanting in their responses.

To prepare this report, we relied on desk research review of publicly available transparency reporting by the platforms; investigations, request for information and audits of platforms carried out by civil society organisations and independent researchers; a review of academic literature on online safety and election integrity; and deep-dives into specific elections conducted over the course 2024 with interviews of experts and insights from local coalition partners and civil society organizations.

2024 Year of democracy

Digital Action, convenor of the Global Coalition for Tech Justice ('the Coalition'), began listening to civil society organizations in a global consultation in 2022-2023 about how Big Tech had harmed them, their democratic processes and human rights. The Global Coalition for Tech Justice was co-created with these civil society organizations in the run up to the historic 2024 elections megacycle, in

the midst of drastic cuts to trust and safety and election integrity in Big Tech companies.

In 2024, elections and referenda took place in 78 countries and territories on 110 unique election days, almost a third of the year, one of the biggest and busiest election years in history. This included close to 60 countries conducting their parliamentary or presidential election. One hundred and four national executive and legislative bodies were elected in 119 elections. Citizens also weighed in on 20 national referenda worldwide. The electoral calendars of several of the world's most populous regions aligned, including India, the United States, Indonesia, Pakistan, Bangladesh, Mexico, South Africa and the European Union with its 27 member states. As a result, more people were affected by and eligible to vote in elections this year than ever before in a single year. The total population of the nations holding elections in 2024 was around 4.16 billion–more than half of the global population. **This coincides with a time where there are 5.24 billion people on social media platforms globally, accounting for nearly 64% of the world's population, and 95% of the people online. For a large proportion of these people, social media is a prominent, if not primary source to access news and information.**

Meta operates the world's largest social media platforms by far – Facebook, Instagram, Whatsapp, and Threads – with a global user base of well over 3 billion people, mostly outside its home US market, alongside Google's YouTube with 2.5 bn users, TikTok's 1.5 bn users and X's 200-250 million users. **The Coalition wrote to these Big Tech companies in July 2023, with a series of demands for 2024's Year of Democracy**, including additional investments and transparency in election integrity plans, and for locally appropriate, rights-respecting content moderation and meaningful safety efforts. **While Meta and TikTok responded in letters containing general information and met with the Coalition on a small handful of occasions; X and Google ignored our calls and refrained from any meaningful engagement over the course of the next 18 months. None of the companies stepped up to meet the demands we set for fully-resourced and transparent elections plans for all countries going to the polls.**

Key Trends: Big Tech's impact on 2024 Elections Megacycle

As part of our campaign, we covered elections and corresponding unregulated illegal or harmful speech online speech, which went largely unaddressed by Big Tech platforms. The range of problematic online speech that platforms failed to adequately respond to are not uniquely new, and clearly indicate that there should have been better preparedness in the form of fact-checking and content filtering measures to identify speech in violation of domestic laws and platform policies, changes to algorithmic design and mechanisms for speedier responses.

On the next pages, we look at some key trends that emerged in how Big Tech platforms responded to the challenges posed by 2024's long election cycle.

Key Trends in Big Tech's Impact On 2024 Elections

1. Poor Regional Commitments and Resourcing
2. Failures to stop the online spread of incitements to violence and harassment, particularly targeting minorities
3. Facilitating Online Gender-based Violence
4. Failures to address coordinated disinformation undermining electoral integrity
5. Algorithms promoting the most harmful content
6. Politically motivated censorship at Meta
7. Abysmal application of rules and inadequate transparency on political advertising
8. Selective non-compliance and non-cooperation with state laws and institutions
9. Platforms obstruct independent scrutiny
10. Poor transparency and response amidst growing Generative AI risks

1 Poor Regional Commitments and Resourcing

Since 2020, major technology companies such as Meta, Google, X (formerly Twitter), and TikTok have implemented significant workforce reductions, particularly affecting their trust and safety, election integrity, and human rights teams. The reduction in trust and safety personnel has weakened platforms' abilities to tackle hate speech, disinformation, and influence operations effectively. Meta announced substantial investments in AI-based moderation tools and human moderation teams, highlighting the engagement of 40,000 personnel worldwide to ensure content integrity during elections. However, this appears to be a cumulative number of staff engaged since 2016, not accounting for layoffs post 2022, let alone regional or linguistic breakdowns, complicating the evaluation of the fairness of resourcing.

In July 2023, the Global Coalition for Tech Justice wrote to Meta (which owns Facebook and Instagram), Google (which owns YouTube), TikTok, and X (formerly known as Twitter), asking them to establish transparent, country-specific plans for the upcoming election year, in which more than half of the world's population would be going to the polls across some 55 countries. Notably, announcements in late 2023 from Meta and Google focused mainly on US elections to be held a year later largely ignoring the scores of elections across the world before that. While Meta and Google released rudimentary country specific blogposts for some countries such as Indonesia, Taiwan, India, Pakistan and Europe which count as its bigger markets, the wide disparity in their approach even in communicating clearly about all the countries holding elections is a stark reminder of the unevenness in their commitments across countries and markets. On the other hand, after the takeover by Musk, X altogether stopped communicating about its plans.

Google's activities were evident in the United States, where election-related measures included greater fact-checking on YouTube and easier access to verified election information. These projects demonstrated the platform's potential to create tools that effectively battle misinformation in well-resourced settings. However, this success did not spread evenly across all regions. In emerging democracies such as South Africa, Google's actions were significantly less constant. Disinformation in local languages spread unchecked in these places, exposing systematic resource distribution and support disparities. This unequal deployment of safeguards prompted concerns about Google's commitment to resolving election-related issues in low-priority areas.

TikTok's Election Integrity Center, which launched in 2023, has shown potential as a tool for real-time monitoring of election-related activity. In regions such as the United States, the platform made strides in combating disinformation by partnering with fact-checking organizations to report and regulate inaccurate content. However, TikTok experienced substantial issues in countries with lesser fact-checking infrastructures. Its poor ability to resolve flagged content in other locations reflected a lack of local responsiveness and highlighted the challenges of scaling election protections to meet the needs of varied electoral settings. Similarly, TikTok disproportionately catered right wing disinformation with young voters in Germany ahead of the elections. During the Romanian elections, Global Witness found Tiktok's recommendation algorithms to be "deeply one-sided" , consistently promoting content supporting one candidate – Georgescu – at a much higher rate than the other candidate. Global WItness also reported

pro-Georgescu content which was in violation of Tiktok's policies but remained online for an extended period of time.

X, adopted a noticeably different strategy under Elon Musk's leadership, which began in October 2022. The network depended mainly on its Community Notes feature, a user-generated tool for addressing misinformation. While this tool enabled users to provide context for erroneous posts, it lacked the size and proactive involvement required to address broad election concerns. X failed to do larger risk assessments or execute extensive counter-disinformation efforts, exposing it to criticism. Observers reported a significant drop in the platform's efforts to maintain election integrity compared to prior years, raising concerns about its ability to counter challenges in the present electoral scene successfully. Most of the staff at X's only African office in Accra were fired shortly after Musk's takeover.

These differences in efficacy suggest a patchwork approach to election preparation across platforms, with some regions, particularly the United States, benefiting from specific interventions and others predominantly across the Global Majority being left with highly inadequate safeguards comparative to the risks present.

The 2024 election cycle exposed major weaknesses in Big Tech's content moderation, particularly in local staffing, stakeholder collaboration, and algorithmic effectiveness in multilingual settings. Meta, Google, TikTok, and X faced criticism for failing to invest in culturally and linguistically appropriate moderation, often relying on flawed automated systems. This is a longstanding issue, which has been repeatedly called out by civil society across the world.

While some positive examples of collaboration emerged in 2024 –such as Google's partnership with South Korean election officials–most efforts were inconsistent. Delayed content removal and unclear moderation policies weakened trust in platforms' ability to ensure election integrity, across a diversity of countries and governance contexts, from Venezuela to Turkey.

Additionally, the rollback of content moderation initiatives in 2024 undermined public safety, particularly in regions with high linguistic diversity. AI systems failed to properly detect misinformation in languages outside of English, leading to biased enforcement. In Hungary, for instance, Meta's AI disproportionately flagged criticism of the ruling party while allowing falsehoods about the opposition to thrive. Similarly, in Indonesia, AI-generated election fraud videos circulated widely before human moderators intervened. **"I can see an increase in hate speech and misinformation related to the election, especially in local dialects on social media platforms like Facebook groups, TikTok, X, and YouTube. Like in the previous election, propaganda, including hate speech and misinformation, may influence p ublic o pinion and political behaviours, including voting," Nuurrianti Jalli of Oklahoma State University** told us shortly before Indonesia's elections in February 2024.

There was a dearth in culturally competent content moderators across Global Majority countries. During Jordan's elections in October 2024, one activist, who requested that their name be withheld for fear of retaliation, told us: **"Content moderators in Meta's team aren't aware of the context. People from different countries m ight not be aware of the nuances in Jordan, even if they had been raised in the region".**

In Pakistan, which has had periodic political conflicts during elections, platforms, and the Election Management Body struggled with language-based disinformation directed at ethnic

groupings. Twitter, now X, played an important role as a vehicle for political discourse, but the shrinking of its moderation personnel left most of the damaging information untouched. This led X to be blocked in Pakistan since the February elections over "national security concerns".

The timeliness of removing harmful content also varied. Platforms committed large resources to the United States, Brazil, and India, while emerging democracies and less important markets such as Bangladesh and Sri Lanka received little support. In Bangladesh, weaponized disinformation went mostly unchecked on Facebook, contributing to increased tensions and leading to a student-led uprising. Similarly, in Sri Lanka, X became a hub for organized disinformation campaigns propagating false allegations about the opposition's claim of voter fraud, but the platform responded minimally due to a lack of on-the-ground moderating experience.

2 Failures to stop the online spread of incitements to violence and harassment, particularly targeting minorities

Perhaps, the most significant form of harmful speech content found on online platforms during the election year was speech targeting religious, ethnic and linguistic minorities, which in India particularly led to deaths of innocent people.

Online xenophobic and inflammatory rhetoric by India's ruling BJP leaders and supporters led to killings of Muslim men by radicalised Hindu nationalists. According to journalist Sirshi Jaswal:

> ***"These disinformation campaigns, this lack of regulation is deadly in India. The people believe in the political disinformation and tend to become radicalized. This is something I have seen on the ground quite a lot. I saw how the far right Hindus, radicalized Hindus were harassing Muslims just because they had seen disinformation online, saying that Muslims seduce Hindu daughters in order to increase the Muslim population. And people do believe in this and then the Muslim men had been beaten to death."***
>
> *Sirshi Jaswal, Journalist*

Human Rights Watch analyzed all 173 campaign speeches by Indian Prime Minister Modi, many of which were propagated online, after the election code of conduct took effect on March 16th 2024. The code forbids appealing to "communal feelings for securing votes." In at least 110 speeches, Modi

made Islamophobic remarks apparently intended to undermine the political opposition, which he said only promoted Muslim rights, and to foster fear among the majority Hindu community through disinformation. Modi regularly raised fears among Hindus through false claims that their faith, their places of worship, their wealth, their land, and the safety of girls and women in their community would be under threat from Muslims if the opposition parties came to power.

Platforms generally exercised inadequate oversight over political hate speech and disinformation, often monetising and subsiding paid channels and handles, in a barely regulated payment space. An investigative report by CheckMyAds documents Google's continued monetization of Hindu Nationalist Media site OpIndia despite its repeated violation of Google's policies on incitement of hatred and disinformation.

In South Africa, the UN had warned that the country was **"on the precipice of explosive xenophobic violence" since 2022, after a social media hate campaign** had spilled over into the streets of Johannesburg and elsewhere, unleashing violent protests, arson of migrant-owned businesses and leading to the murder of a Zimbabwean national. In the leadup to South Africa's elections in May 2024, members of Jacob Zuma's MK party ramped up threats of violence should they not get their way at the polls, or should the court decide to disqualify Zuma from the race. Provincial leader Visvin Reddy faced legal proceedings for charges of inciting public violence, and this is one of many examples of incitement of violence made by MK party members in public and shared on social media platforms, including X (formerly Twitter) and TikTok.

LRC and Global Witness tested Facebook, TikTok and YouTube ability to detect hate speech and incitement to violence targeting non-nationals. The groups prepared ten adverts – which were withdrawn post-approval and never published – based on real-life content in English and translated into Afrikaans, Xhosa and Zulu. The ads called on the South African police to kill foreigners and encouraged violence through "force" against migrants. The ads were approved by all three social media platforms, with the exception of only one ad in English and Afrikaans rejected by Facebook. During the elections, **Yasmin Rajah head of Refugee Social Services in KwaZulu-Natal, a coastal South African province, told us: "I don't think social media platforms do much to stamp out hate speech. It seems like anything goes in South Africa."**

In Tunisia, the authoritarian regime of Kais Saïed used Facebook as its own hate speech megaphone building out a sustained social media campaign dehumanising and targeting human rights defenders and activists, among others. In May 2024, in his speech shared on Facebook, Saïed referred to rights groups helping migrants as "traitors" "[foreign] agents" and "rabid trumpets driven by foreign wages". The video was played almost 290 thousand times and was shared 1.7 thousand times, with some users explicitly repeating Saïed's inciting language. **"It's a matter of time before this turns into real world violence," said one Tunisian who we interviewed on conditions of anonymity.** "This narrative isn't only dangerous in terms of them [NGOS] getting charges but these campaigns are successful at convincing the public opinion and removing the public support from civil society. Even my mom believed it. "

Siwar Gmati of civil society group I Watch, which is part of Meta's trusted partner programme – an initiative that Meta says taps into the expertise of local groups to "address problematic content trends and prevent harm"– flagged incendiary Facebook posts targeting right groups to the tech giant. In response, Gmati said, Meta told

them there's nothing it could do because posts accusing rights groups and their employees of being "traitors" working against the interest of the Tunisian state are just an opinion.

> ***"Facebook is in a way responsible for democratic backsliding and it's not doing anything to protect human rights defenders," Gmati told us. "They [Meta] claim that they're defending freedom of expression. But sorry this is not just an opinion if someone says that you're a traitor, that you're not a patriot. What is going to happen offline to an activist who is labelled as a traitor? We see double standards in treatment. When it comes to countries and consequences – i.e. they take down pro-Palestine content or against Ukraine or pro-Russia content."***
>
> ***Sirshi Jaswal, Journalist***

In Indonesia, research and analysis of online content found hate speech against six vulnerable minority groups: Shia, Ahmadiyyah, Christians, LGBTQ+, Indonesian Chinese, and people with disabilities, and Jews.

In the US, election disinformation blamed immigrants for a rise in violent crime and that immigrants were causing a rise in unemployment for people born in the U.S. These claims were completely unfounded but anti-immigrant disinformation was rampant in both English and Spanish, such as widespread rumours that crime rates skyrocketed in New York due to increased immigration. After the presidential debate in September 2024, in which Donald Trump alluded to disinformation that members of the Venezuelan gang Tren de Aragua were taking over a Colorado apartment complex, the complex's residents said "they feel unsafe [...] and they fear being stereotyped as criminals." Similarly, in response to online rumors amplified by Republican leaders, including Trump, that Haitian immigrants were engaged in crime and 'eating people's pets', they received threats and the town of Springfield, which was at the centre of this controversy, received more than 33 bomb threats in a short period after the US Presidential debate where these rumors were repeated.

The pace at which platforms responded to hate speech was alarmingly slow even in the US where platforms had committed most resources. Global Witness carried out an investigation reporting a sample of comments to Facebook where they assessed that the post may have breached a Facebook community standard. After more than three days, 13 of 14 posts reported had not been reviewed by Facebook moderators. One comment, which made explicit sexual claims about Kamala Harris, was reviewed in this period and was determined by Facebook not to violate Community Standards. After Global Witness reached out to Meta for a comment on this investigation, they reversed this position, and removed the post for violation of Community Standards. To give you an example of the severity of content reported by Global Witness, they included posts "Wishing death upon a political candidate", "Saying that people of a specific religion or ethnic group "are plagues," or that they are "inbred" and "parasitic", "Calling women political candidates an "ugly whore", a "slut" and saying that "Someone should just spit in her ugly face" and "Using homophobic slurs against political candidates."

3 Facilitating Online Gender-based Violence

The 2024 election cycle was also notable for the hate speech, abuse, offensive speech and misogyny directed towards women and the LGBTQ community. In the US elections, women of color and African American women candidates in particular were subject to more offensive speech overall, and specifically to more hate speech, than other candidates.

The ready availability of AI based tools, cheapfakes and image manipulation have been used to depict women candidates in fake sexual situations, misogynistic terms like "whore," doxxing of their private photos and videos and attacks on women for how they choose to dress, much of this content was spread via Facebook and Instagram.

This was a well documented problem prior to the 2024 election cycle. In Iraq in a 2022 report, where, given the ultra conservative nature of the society, these tactics are very detrimental to women and their participation in public life including elections. In Tunisia, in 2022 supporters of President Kaïs Saïed's authoritarian regime targeted human rights activist Rania Amdouni, subjecting her to a hateful social media campaign on Facebook that included derogatory comments, threats, and doxxing. Despite seeking help from the police, she was arrested and wrongfully charged with insulting a public officer, and had to eventually flee the country.

In South Africa LRC and Global WItness carried out an investigation in 2023 testing the ability of platforms to detect hate speech and abuse, by submitting ads offending their policies. The ads submitted for approval were inspired by real instances of abuse faced by women journalists. They contained violent, sexualized, and dehumanizing language, referring to women as vermin, prostitutes, or psychopaths, and even advocating for their assault or murder. Statements included phrases like "they're just all sheep and should be slaughtered" and "they all need to die." Despite the overtly extreme nature of these ads, which clearly violated the social media platforms' own hate speech policies, the majority were still approved across all four platforms – Facebook, X, Youtube and Tiktok.

Notwithstanding the clear issues with online gender-based violence, made easier with the rollout of new Big Tech-backed generative AI tools since 2022, tech platforms did little to prevent it becoming an election feature affecting women's public and political participation in 2024. In Brazil, during the mayoral elections of October 2024, women were the primary targets of online violence. In São Paulo, candidates Tabata Amaral and Marina Helena experienced a surge in digital attacks, facing three times more harassment on YouTube and X than their male counterparts. According to research by Democracy Reporting International (DRI) and Fundação Getúlio Vargas (FGV), over 80% of gender-based violent posts aimed to discredit women's participation in politics.

Additionally, FGV found that left-wing women candidates were targeted more frequently than those on the right. Misogynistic and transphobic content targeting women candidates spread across YouTube in both urban and rural areas of Brazil. A study by MonitorA revealed that most attacks sought to depict women as inferior, reinforcing misogyny and questioning their intelligence.

In Pakistan, political leader Azma Bukhari's sexualized deepfake video was circulated widely on social media platforms in November 2024. Before that, prominent women journalist Meher Bokhari had been the victim of technology-facilitated gender-based violence (TFGBV) and Generative AI with circulation of morphed images of her.

In the US, the American Sunlight Project released findings after the presidential elections that 25 members of the Congress including 24 women had been victims of sexually explicit deepfakes, thus accounting for 16% of US Congresswomen.

> ***"The cost of producing a deepfake during elections against a woman is zero in this country. We presented all the evidence, documented, and the most we managed to achieve was that some videos were removed after a few days. These videos were already in people's phones and in their WhatsApps. It's impunity that explains all of this. And do I think that the social media platforms are responsible for this process? One hundred percent!"***
>
> *Tabata Amaral, Member of the Chamber of Deputies, Brazil*

In particular, the online discourse around the elections reflected an organised resistance to the gains that have been made on gender equality across the world. The New York Times called the US elections the Gender Election noting a stark new gender divide has formed among the country's youngest voters, and the support amongst Gen Z men for Donal Trump. Similarly, the rightward tilt in the EU Parliament was accompanied by a corresponding tilt towards more male candidates.

4 Failures to address coordinated disinformation undermining electoral integrity

Another form of harmful speech witnessed across countries was spurious speech designed to compromise trust in the electoral process and opposing candidates.

In Indonesia, analysis into the February 2024 election looked at the wide prevalence of black campaigns – campaigns maliciously aimed at destroying the character of a competitor through the spread of false information, slander, and accusations without any evidence. Their data collection highlights several examples of disinformation which aimed to create confusion about the voting and electoral counting process, and a combination of misinformation, cheap-fakes and manipulated media to spread false information about candidates. Some of the extreme acts of disinformation included "the front runner in the upcoming presidential election speaking fluent Arabic; a long-deceased president praising the incumbent, and a presidential candidate being scolded by one of his political backers." Several of these posts went viral on social media, with the take down responses extremely late or non-existent.

In **South Africa**, X became the main conduit for conspiracy theories undermining the integrity of the country's election commission and falsely claiming that the election had been rigged in favour of the ruling African National Congress (ANC). Duduzile Zuma-Sambudla – the daughter of ex-president Jacob Zuma and leader of newly formed MK party – who has over 300 thousand followers on X was the top disinformation superspreader. One of her posts, shared days before the election, showed pictures and videos of what appeared to be ballot boxes with a caption accusing the ANC of "stealing votes". The post, which was viewed almost 650 thousand times, remains on X at the time of writing with no associated community fact-check note.

In November 2024, a little known ultranationalist candidate Călin Georgescu won the first round of a high stakes presidential race in Romania thanks to a TikTok campaign, which was similar to Kremlin-run influence operations in Ukraine and Moldova, according to declassified Romanian intelligence documents. On 6 December 2024, Romania's top

court annulled the results of the election, amid warnings of an "aggressive" Russian hybrid attack on the Eastern European country. EU regulators have launched a probe into whether TikTok breached the bloc's digital rulebook by failing to deal with risks to Romania's presidential election.

5 Algorithms promoting the most harmful content

Engagement-based algorithms, called recommendation systems, have long played a role in spreading health misinformation, political disinformation, hateful rhetoric, and other harmful content worldwide. Mozilla's Youtube Regrets project demonstrated that Youtube's algorithm is responsible for 70% of total watch time–amounting to an estimated 700 million hours daily. Given YouTube's vast reach, such videos have significantly influenced audiences, contributing to issues like radicalization and societal polarization. Similarly, engaged-based algorithms have been at the root cause of harm at Meta. **A leaked internal Facebook document states: "We have evidence from a variety of sources that hate speech, divisive political speech, and misinformation on Facebook and the family of apps are affecting societies around the world. We also have compelling evidence that our core product mechanics, such as virality, recommendations, and optimizing for engagement, are a significant part of why these types of speech flourish on the platform."** Meta's platforms continue to incentivise angry, polarising content including hate speech, violent speech and misinformation knowingly since it leads to more engagement.

TikTok's algorithms are actively driving radicalization, polarization, and the spread of extremism, contributing to societal instability. Findings indicate that users encounter extremist content through multiple pathways, with a significant portion being promoted by the platform's recommendation system, which acts as a radicalization pipeline. Rather than merely providing personalized content, these algorithms play a direct role in fostering radicalism, social violence, and division.

Problems related to engagement-based algorithms have been long recognised by experts. From the beginning of the 2024 elections cycle, it was noted how Meta, TikTok and YouTube's algorithms were favouring sensationalist, polarising and toxic content in Taiwan, helping push out misleading and false information to potential voters. ***"The media ecosystem, including the algorithm, prefers low quality information over high quality information. Low quality like disinformation, misinformation, fake news, sensational stuff, rumours, hoax,"*** states Eve Chiu, the CEO of Taiwan FactCheck Center, whose organisation was part of Facebook's Third Party Fact-Checking program at the time. This phenomenon was in evidence across countries in 2024.

Outside of AI-driven recommendation engines on online platforms, the use of AI to analyze demographic, household, and personal data drawn from a variety of sources, make inferences about the political motivations of citizens, and help campaigns manipulate them remains another big factor. The potential for this form of AI use to grow further along with the digital electioneering infrastructure is significant.

6 Politically motivated censorship at Meta

Tech platforms, and Meta in particular, have a years long track record of content moderation failures in the Arabic language and of silencing pro-Palestinian content from across the globe. In Jordan, where the moderate Islamist opposition made significant gains, in part due to anger in the country over Israel's latest war in Gaza, Meta's censorship policy is no different. At least 90

Jordanian journalists reporting on Palestine and the protests in Jordan in support of Palestine and against Israel's ongoing genocidal campaign in Gaza, had their Instagram or Facebook accounts blocked and/or removed, according to an interview we conducted on 20 August 2024 with a Jordanian human rights activist, who asked that their identity be concealed for fear of reprisals.

When a human rights activist working for a trusted flagger organisation intervened on behalf of some of the journalists, Meta allegedly said it had blocked or removed the accounts because they violated Instagram community guidelines. The journalists allegedly violated the rules on dangerous individuals and organisations, which Meta has been using to censor pro-Palestinian voices for years, most notably during the 2021 Gaza war. According to the human rights activist, some accounts were removed due to false reporting.

> ***"Since the war on Gaza started, Meta has been restricting freedom of speech. Journalists had their posts removed, their [Facebook] pages are getting blocked, pages of political parties are getting removed, student groups have been affected by these things,"*** *a human rights activist speaking on conditions of anonymity told Digital Action.* ***"We can't talk about democracy in Jordan if Meta is cherry picking the opinions that are going online. That's affecting how social media companies are shaping public opinion that is also shaping our political life."***

7 Abysmal application of rules and inadequate transparency on political advertising

Advertising transparency frameworks, such as Meta's Ad Library, faced criticism for their limitations. The Ad Library provides some visibility into political ads but fails to adequately track AI-generated content or disinformation campaigns, especially in under-resourced regions.

In India, where pre-publication certification of political advertising by the Election Commission, the country's Election Management Body (EMB), is mandatory by law, neither the platforms nor the EMB showed interest in implementing it for online political ads. India Civil Watch International (ICWI) and Ekō documented Meta's approval of AI-manipulated political advertisements during India's election that spread disinformation and incited religious violence. To evaluate Meta's mechanisms for detecting and blocking political content that could prove inflammatory or harmful during India's ongoing elections, they created several such instances of political advertisements and submitted them on Meta's ad platform. The report claims the adverts were based on existing examples of real hate speech and disinformation prevalent in India. In all, they submitted 22 adverts in English, Hindi, Bengali, Gujarati, and Kannada to Meta, of which 14 were approved by Meta's review mechanisms. Global WItness tested the efficacy of the ads review system of Facebook, Tiktok and Youtube prior of the US elections and found that Tiktok approved 50% of the ads despite its policy explicitly banning all political ads, Facebook approved one out of the eight ads, and Youtube disallowed them without more identification as they referenced elections.

In April 2024, Mozilla and Check First conducted stress testing of ad repositories of 11 platforms

including X, Apple's App Store, Google, Meta, TikTok and LinkedIn. The study revealed that these tools frequently deliver incomplete data, have malfunctioning search features, and are challenging to navigate efficiently. Among the technology giants assessed, X performed the worst, offering minimal useful data for both watchdogs and users.

Organized influence campaigns on social media have become a key strategy for political and economic elites to shape public opinion to their advantage. Poor application of review mechanisms for political ads by platforms play a huge role in allowing such actors to thrive.

In India, monetized actors created online pages, often appealing to Hindutva actors, but also those critical of Hindu nationalism. Their primary motivation was to earn money by building an online following, which could later be monetized through advertisements. There were also actors whose ideology aligned with their business interests.

In certain cases, there are no clear financial ties between political parties and the diffuse actors working for them. Academic Sahana Udupa analyzed social media content and network interactions through purposive sampling. According to her findings, one category of vote mobilizers was the 'techie-turned-ideologue', which primarily consisted of technically trained, English-speaking volunteers proficient in social media. These individuals did not receive any financial compensation from the party and engaged in their work out of passion. These actors create complex regulatory questions for platforms, and poor campaign financing laws, which facilitate the opacity of monetary exchanges between formal political actors, influencers and other diffuse actors making it harder to implement political advertising rules.

8 Selective non-compliance and non-cooperation with state laws and institutions

The election cycle was also a stark reminder of the immense geopolitical power amassed by a handful of Big Tech companies, as their transnational presence put them in circumstances where they had to navigate the regulatory and enforcement powers of national governments and institutions. The ability of national institutions to regulate and enforce their national laws against the platforms appears to be emerging as a function of the size of their market. This is made apparent, more than anything else, by the ease with which some election management bodies were able to strike agreements for co-regulatory efforts during their elections. Agreements between election management bodies (EMBs) and platforms which established a direct line of communication between the two, and imposed some positive obligations on platforms emerged as the primary form of election related online content regulation. Aside from the US and EU, large social media markets like India and Brazil found it relatively easy to convene platforms and ensure their participation in structures and mechanisms formulated to govern the behavior of platforms on issues of content takedown, moderation, transparency and accountability.
The contrasting degree of attention that platforms diverted towards different countries is apparent also from their public statements, which favored large markets. Tiktok, for instance, only published separate posts on its plans for the elections in Indonesia, Bangladesh, Taiwan, Pakistan, the European Parliament, the US, and the U.K. Meta announced its plans for the UK, South Africa, India, Brazil and EU. Google had listed its plans for elections in India, EU and the US. Aside from these large markets, other smaller countries had very limited negotiating power to ensure election integrity efforts from platforms in their regions.

There were singular examples where active participation between EMBs and platforms did take place. In Brazil, the Superior Electoral Court (TSE) spearheaded one of the most comprehensive initiatives, signing Memorandums of Understanding (MOUs) with Meta, TikTok, Google, and others. These agreements established clear mechanisms for reporting and removing harmful content, including disinformation, hate speech, and electoral manipulation. A notable aspect was the TSE's 24-hour takedown policy, which required flagged content to be removed promptly. Meta strengthening its local moderation teams and creating alliances with fact-checking organizations. For example, it deleted vast amounts of flagged disinformation about political personalities and electoral procedures. Similarly, Google implemented stronger regulations on YouTube, enhancing its ability to combat election-related misinformation. YouTube's updated search criteria prioritized authoritative sources, ensuring that users received accurate information during the election.

Brazil was, however, an outlier, both in its ability to negotiate a robust framework with platforms and ensure its reasonable implementation. In some cases like South Africa, the records were mixed. While the election management body, the Electoral Commission of South Africa (IEC) had success in bringing Meta to the table to make commitments towards combating misinformation, its data protection regulator's request of Google Meta and X for access to records regarding the classification of elections, risk assessments of South Africa's electoral integrity and the application of its global policies to the country were denied, leaving them with little regulatory recourse. Similarly, when public interest law firm LRC submitted two access to information requests to Meta, Google and TikTok on their election action plans asking for substantive information on content moderation and emergency tools available in respect of the South African elections, all three platforms refused to divulge any details, indicating that South African access to information laws do not apply to them.

In India, the Election Commission coordinated with Meta and Google to streamline the reporting of harmful content, establishing hotlines to facilitate real-time intervention. These efforts also included measures to address regional and vernacular content, though their success was mixed. TikTok's ban in India reduced its involvement, while X, following substantial layoffs, participated minimally, relying primarily on its Community Notes feature to crowd-source content moderation. By contrast, Indonesia's election management body, the KPU, faced challenges securing formal agreements with platforms.

In most cases these agreements were non-binding and left large aspects of platform activity including online political advertising unregulated. While platforms implemented broad election-related policies, localized protocols were limited, particularly in addressing linguistic diversity and regional-specific risks. At the time of the Mexican elections, Agneris Sampieri, Latin America policy analyst at Access Now, told us: "it is important to note that these agreements [between platforms and election management bodies] are not reflected in any changes to the platforms' policies. On the contrary, companies interpret their policies extensively to implement these collaborations without acknowledging any legal accountability or obligation between the platform and the authority, no binding responsibilities to users".

In Africa, the Association of African Electoral Authorities developed the Principles and Guidelines for the Use of Digital and Social Media in Elections in Africa, however, we did not see a proactive approach by African states to utilise these principles in the management of elections.

In the United States, platform-EMB cooperation was somewhat fragmented, owing to the country's decentralized election management system. State-level election boards collaborated with platforms to combat disinformation about voter registration and ballot security. In Pennsylvania, for example, the state's election office worked with Meta to create a trial tool that highlighted election-related falsehoods to local election officials. However, the decentralized model resulted in inconsistent reactions. States with fewer resources or less tech-savvy election officials were unable to fully employ platform technologies, making their citizens more susceptible to misinformation.

9 Platforms obstruct independent scrutiny

An obvious hurdle to studying, researching and understanding the way platforms work and influence elections and the enjoyment of fundamental human rights is their private nature, which places them outside of freedom of information requests and public auditability and accountability structures. To respond to this structural problem, over the last few years different mechanisms had been developed to seek access to platform data. However, the state of data access for researchers, civil society organizations, and election monitors has seen significant changes during the 2024 elections. These shifts, driven by platform policies and evolving priorities, have had profound implications for transparency, accountability, and electoral integrity. Meta's decision to dismantle CrowdTangle in August 2024 placed a substantial obstacle in the way of independent scrutiny of Facebook and Instagram. Despite Meta's claims that alternative tools would be introduced, academics have extensively criticized the replacements for lacking functionality and real-time capabilities. These disruptions significantly hampered their capacity to assess disinformation or other key trends, with 88% of surveyed researchers reporting disruptions to their work.

Other platforms have also restricted access for journalists and researchers. X introduced expensive API fees, making it difficult for independent researchers to analyze platform activity. This measure, implemented under Elon Musk's leadership, curtailed watchdogs' access to evaluate platform behavior and trends in real-time. In Poland, where elections were influenced by disinformation targeting LGBTQ+ populations, civil society organizations failed to prove the coordinated amplification of negative themes as a result of X's stringent access standards.

While Google maintained some transparency through its Ad Transparency Center, the data provided was either insufficient or confined to specific countries, resulting in significant gaps in understanding the scope of political ad campaigns in Africa and Southeast Asia. US tech company Reddit, a growing hub for political conversation in nations such as Australia, imposed stricter API access in 2024. This change hindered academic researchers' capacity to investigate coordinated manipulation tactics, such as those involving fringe political groups supporting voter suppression myths.

Platforms lack standardized protocols for data sharing, resulting in fragmented research efforts and incomplete analyses of online disinformation networks. Without standard data-sharing methods, these fragmented efforts resulted in an uneven environment for independent monitoring of the influence and impacts of tech platforms on electoral integrity.

There was strong regional inequity in access to platform data. Whereas European researchers could rely on data access provisions within the EU's Digital Services act, researchers across Global Majority regions often faced insurmountable obstacles to their independent scrutiny of Big Tech platforms. In response, the African Commission on Human and Peoples' Rights adopted a resolution in November 2024 calling for greater efforts to promote and protect access to data across Africa. This emerged with the backing of the civil society African Alliance for access to data.

10 Poor transparency and response amidst growing Generative AI risks

At the beginning of the 2024 election cycle, there were wide fears about the role AI, particularly generative AI, would play through the year. In February 2024, several large technology firms including Adobe, Amazon, Anthropic, Arm, ElevenLabs, Google, IBM, Inflection AI, LinkedIn, McAfee, Meta, Microsoft, Nota, OpenAI, Snap, Stability AI, TikTok, Trend Micro, Truepic and X signed up to the Tech Accord pledged to develop "detection technology" and "open standards-based identifiers" for deepfake content and watermarks. The idea was to make sure platforms and generators shared the same tools to spot and remove fake content when it harms electoral processes. At the end of the year, there is general consensus that the elections were not the generative AI disaster everyone thought they would be, despite some damaging instances of it. The use of AI and deepfakes to target women politicians and journalists was also problematic across regions, as noted above (see section on "Facilitating Online Gender-based Violence").
Political deepfakes cropped up in many elections, suggesting these will now become a permanent feature of future elections, in the absence of concerted action.

In Slovakia (where AI-generated audio recordings purported to show a top candidate boasting about rigging the election, which he would go on to lose) and Pakistan (where a video of a candidate was altered to tell voters to boycott the vote). There were also several cases of use of AI for coordinated campaigns.

In Rwanda, Kagame's government launched a coordinated online campaign that used AI to generate and spread political propaganda pushing propaganda narratives about the Congo conflict,

fighting the government's critics, and boasting of Kagame's successes and accomplishments. Bangladesh saw several cases of the pro-Hasina content generated by AI such as an opposition leader equivocating over support for Gazans.

In Indonesia, around 28 percent of the Indonesian population belongs to Generation Z, while millennials make up around 26 percent. Video-sharing platforms like Tik Tok and Instagram became battlegrounds for the youth vote. An AI-modified TikTok video of a social media post by musician Taylor Swift purportedly thanking Prabowo Subianto – the candidate expected to win the presidential race – "for helping me". The video had been in circulation at least between 31 January 2024 and 7 February 2024, and was taken down following a viral tweet that sounded alarm about the hoax. Before it disappeared from TikTok the video had racked up over half a million views. Do individual posts matter? **According to Rizka Herdiani, researcher at Indonesia's Center for Digital Society:**

> ***"Some voters choose a certain candidate for bite-size "information" or clips of the presidential debates that are posted on platforms such as TikTok. So, yes, it can influence the results of the elections"***

Nonetheless, a report by Knight First Amendment Institute analyzed every instance of AI use in elections collected by the WIRED AI Elections Project which tracked known uses of AI for creating political content during elections taking place in 2024 worldwide, concluding that half of them were not deceptive, and for the remaining half the cost of creating similar content without AI was modest.

At the same time, in 2024, experts found that China's domestic censorship had spilled over into AI-generated content on US Big Tech platforms, suggesting they were not taking adequate steps to mitigate risks. According to Tzu-wei Hung, a research fellow at Taiwan's Academia Sinica: "ChatGPT and Microsoft Bing answers differently in Mandarin and English to questions involving sensitive keywords like the Tiananmen incident or the Uyghur genocide. Google Bard repeatedly says it is unable to answer in Mandarin about Falun Gong, organ harvesting, and the Dalai Lama".

However, it was not entirely clear what steps were taken by most Big Tech companies to combat the mala fide use of AI in elections. The Brennan Centre noted that "many signatories did not report their progress despite being provided multiple forums to do so – and despite "transparency to the public" being one of the accord's pledges. Several failed to report any progress, and those that did often left out key commitments when assessing their own performance. Even when companies did report on their progress, several failed to provide much detail to back up their assertions." Thus, while companies were able to claim the goodwill that came with signing up for the accords, it was not accompanied by corresponding accountability measures.

The elections have been instructive on the ways in which synthetic content may be deployed in future elections and the kind of impact it might have. The use of AI in personalizing voice calls by political campaigns or live voice translations of speeches in regional languages were clear examples of the use of technology to make political messaging more effective. In countries like South Africa, reliance on older technologies attracted more news coverage than generative AI, such as traditional mis- and disinformation using false headlines, allegations of voter fraud, out-of-context images. A report by the Centre for Media Engagement found that X played a central role in both the spread and, more importantly, the longevity of misleading AI-content, and was notably absent from the

A Dire Outlook: Big Tech and Elections in 2025 and beyond

framework of cooperation that the Electoral Commission of South Africa signed with other platforms (X chose not to sign).

In 2017, Facebook was accused of enabling hate speech that contributed to genocidal violence against the Muslim Rohingya minority in Myanmar. In response, the company acknowledged its failures and implemented several measures, including a human rights impact assessment to evaluate its role in the crisis. It strengthened content moderation by hiring more local language moderators, refining its hate speech policies, and collaborating with international organizations such as the UN's Independent Investigative Mechanism for Myanmar.

While the company has historically addressed high profile international scandals by tightening policies, investing in third-party fact-checking, and committing to greater accountability, there was a dramatic reversal at the beginning of 2025. Meta CEO Mark Zuckerberg's announcement on 7 January 2025 to turn Meta's back on fact-checking, relaxing hate speech policies and limiting uses of content filters, effectively played to the incoming Trump administration's demand for lower Trust and Safety standards. Zuckerberg cited oft-disproved allegations that they have operated by a bias against right-wing content, limiting the US constitution's First amendment protection of Americans.

For platforms, whose algorithms already prioritise,

elevate and promote some of the most harmful content–such as hate speech, misinformation, and material inciting racial violence–all under the guise of boosting "user engagement" and ultimately driving profit, stepping back from existing efforts to identify harmful content is significant.

It is useful to note that this dramatic reversal on internal policymaking by Big tech companies in response to expedient political circumstances is not a new trend, if we look at the records of the platforms in other key markets. In India, for instance, platforms have largely acceded to government's demands to take down content or accounts, and block content not just locally but also globally, at the government's behest. Facebook whistleblower Frances Haugen had detailed structural problems where staff responsible for negotiating with governments on regulation and national security, as well as managing media relations, were also given the authority to influence discussions on creating and enforcing Facebook's content policies globally, creating a conflict of interest designed to ensure political influence over content moderation decisions. Similarly, Facebook gave in to censorship demands from Vietnam's communist government, removing more than 2,200 posts between July and December 2020. Such accession to government demands have in the past also been documented in a variety of countries, including India, Vietnam, Turkey, and Israel.

As tech companies branch out into related industries, they have created opportunities for the government to leverage political content moderation as a negotiating tool. During times of business consolidation and efforts to diversify revenue beyond advertising, Google and Meta have become more compliant with government requests. Over time, we have seen governments use a complex combination of legal, economic, and political forms of coercive influence to shape platforms' moderation of political content. Overall, the extent to which governments are able to exercise coercive power over Big Tech companies depends on the strength of institutional and civic mechanisms that counter arbitrary government regulations, the importance of a country as a key market for platforms, and the development of a nation's domestic tech sector in relation to its government's economic dependence on U.S.-based technology companies.

Rolling back fact-checking

Fact-checking has been around for more than a quarter of century in the form that we experience now but only experienced a financial boost in 2017 after it was revealed that Russian operatives had used fake accounts on Big Tech social media platforms during the 2016 US presidential election. In response, Meta started employing external fact-checkers to assess certain content across its platforms, including Facebook, Instagram, and WhatsApp. Meta's fact-checking ecosystem depended on over 90 third-party organizations accredited by the International Fact-Checking Network. Given the vast and overwhelming volume of content flowing through social media, fact-checking was never intended to address more than a small fraction of the misinformation circulating on major platforms. Additionally, fact-checkers had no authority to enforce their findings–it was entirely up to social media companies to take action. Social media platforms responded in two primary ways: by down-ranking and labeling posts proven to be false. While false content was not removed entirely, it became less visible in users' feeds. If users did come across questionable posts, they might see a label linking back to the fact-checkers' conclusions.

However, over time, the fact-checking industry got caught in a conspiracy theory peddled largely by right wing commentators in the US that they were unreasonably biased against them. Without

any real proof, these claims were picked up by Zuckerberg in January 2025 while withdrawing support for fact-checking in the US. He justified this move by adopting the baseless narrative that factcheckers "have just been too politically biased and have destroyed more trust than they created, especially in the US." This withdrawal is limited to the US, for now, however, it does not bode well for other countries, as Meta repositions itself in response to an assertive Trump administration. The African Commission on Human and Peoples' Rights responded directly to these developments in March 2025 underscoring the inefficacy of solutions like Community Notes, and asking the Special Rapporteur on Freedom of Expression and Access to Information in Africa to develop guidelines to enable countries to effectively monitor the platforms' performance in order to inform of the efforts to advance information integrity online, including the role of independent fact-checking in the African context.

Dialling down Content Filters

The January 7th 2025 announcements by Zuckerberg justifiably received a lot of press coverage globally. While this press coverage was focussed on the roll-back of support for fact-checking programmes in the US, there were perhaps a couple of other developments which signify even more important shifts. The first is the changing position of the use of content filters. For some years, social media companies, particularly the platforms run by Meta, have actively invested in, developed and implemented automated and AI-driven content moderation systems, diverting significant resources in their direction. Going forward, as Zuckerberg mentioned, this will no longer be the case. He said: "We used to have filters that scanned for any policy violation. Now, we're going to focus those filters on tackling illegal and high-severity violations, and for lower-severity violations, we're going to rely on someone reporting an issue before we take action." This effectively means a move from proactive content filtering, where automated systems would constantly flag policy violations to a reactive system which would wait for a violation to be reported by a user. Content filters would be limited to a much narrower set of illegal and high-severity violations. This puts the entire onus on consumers, allowing Meta platforms to shirk their responsibility.

As academic Rebecca Hamilton succinctly explains Meta's two-fold card trick. Relying on users to govern platform content presents significant challenges. In Myanmar, Meta depended on users to report hateful content, resulting in Facebook's "significant role" in spreading content that fuelled the country's genocide against its minority Rohingya population, and now it is using the same approach globally. However, the fundamental issue remains unchanged: the flagging system simply reflects the prevailing norms of the majority within a given user community. For Zuckerberg and Musk, who first deployed Community Notes on X, this is an advantage–but for marginalized groups, it is a dangerous flaw. Second, facilitating an online space which defaults against marginalized communities is a giant looping mechanism with online activity and offline contexts fuelling each other.

Mainstreaming hate speech

The second significant move by Meta at the start of 2025 was to dramatically relax the content and even the name of its Hate Speech Community Standard, now called the Hateful Conduct Community Standard. The updated policy now allows the use of derogatory language and calls for exclusion based on gender, sexual orientation, and national origin. For instance, it explicitly permits claims of mental illness or abnormality related to gender or sexual orientation, citing political and religious discussions on transgender issues and homosexuality. It also explicitly allows for calls "for sex or gender-based exclusion from spaces commonly limited by sex or

gender, such as restrooms, sports and sports leagues, health and support groups, and specific schools."

The language of the address by Zuckerberg itself employs an intentional distortion, fashioning 'content moderation' as censorship, thus dishonestly framing Meta as a protector of free speech through these changes.

A whole-of-society approach and its limited replicability

Early in 2024, the elections in Taiwan offered an interesting model to respond to disinformation online. In Taiwan, a key techno-political challenge of the last decade has been to respond to Chinese propaganda. The public and government both now view the threats to trust posed by disinformation as existential threats from China and the government has responded with a 'whole of society' approach.

One prong of this strategy is to increase transparency with a host of initiatives, including seeking the participation of the innovative g0v (gov-zero) project. This team joined the government to create the Public Digital Innovation Space (PDIS) and launched the consensus-building project, vTaiwan and its graphic avatar, Polis, which found a structured way for public forum discussions to inform policy decisions. This digital infrastructure was vital in responding to disinformation, where an unconventional team, including graphic designers and comedy writers inside the government, would create memes directly responding to fake news. This model of response is in direct contrast to more authoritarian responses which involve censorship and takedown of content, something that a country which only thirty years ago emerged from four decades of martial law is, with good reason, uncomfortable with.

The strategies adopted in Taiwan, particularly by the government actors, do offer plenty of implementation lessons for a successful counter-disinformation campaign. However, it is important to note the combination of very unique circumstances which made this approach feasible. The methods that succeed in a country like Taiwan will have limited successes in more complex, larger, multi-ethic polities. Second, the identification of a common foreign adversary posed due to existential Chinese threats are strong factors getting buy-in from different stakeholders for a whole-of-society approach. Third, a state of independent and highly vibrant democracy, a working constitution, a transparent government, and a free and fair judiciary along with high levels of digital literacy are strong contributing factors, not replicated in most other countries.

A failure of self regulation

Mark Zuckerberg announced a partnership with the incoming Trump administration to fight against the efforts of other countries to regulate US tech companies. The US administration has made threats against various states, including the European Union and Brazil, to retaliate if they try to tax and regulate US Big Tech companies.

This follows roughly a decade of community guidelines by Big Tech platforms, partnerships with fact-checkers and content moderation practices representing a long experiment with self-regulation as the primary mode of regulating speech online. However in light of the failures of these to prevent high profile tech harms, such as online radicalization and hate speech spilling over into violence, calls for regulation strengthened over time. In response, Big Tech companies have spent an enormous amount of capital lobbying governments and policymakers across the world. A recent report by LobbyControl, and jointly published with Balanced Economy Project and Global Justice Now, noted that the tech industry as a whole has increased its lobbying expenditure from

97 million euros to 113 million euros in the European Union, with the big five companies (Google, Amazon, Meta, Microsoft and Apple) accounting for 33 million euros. The powerful lobbying efforts of the tech industry, combined with its vast market dominance and monopoly influence, are at odds with democratic values. As economies and societies grow increasingly reliant on the products and services of private tech companies, governments become more susceptible to the sector's influence, often prioritizing corporate interests over the public good. Tech billionaires are leveraging this dependence to expand their businesses and advance their political agendas, ultimately weakening the fundamental democratic principle of equal representation for all voters.

So far, most of the social media platforms have managed to avoid significant regulations across jurisdictions on the basis of the arguments that they self regulate themselves, through content policies. There are strong arguments for supporting self-regulation for media companies to create barriers for state censorship, however, for self-regulation to be meaningful and effective, several factors are necessary. They require self-regulatory bodies which are independent from government, commercial and special interests, established via a fully consultative and inclusive process, have a robust complaints mechanism and clear procedural rules to determine if ethical standards were breached in individual cases, and have the power to impose sanctions. It is noteworthy that none of the above factors are satisfied in the form of self-regulation that Big Tech platforms practise.

Conclusion

As we navigate the complexities of issues involved in the governance of online speech on digital platforms during elections, it is easy to miss the forest for the trees. With the emergence of social media-driven news practices, it was believed in the 2000s that this was a significant shift towards greater democratization of news. The mounting loss of faith in mainstream media led many to believe that this would limit the ability of editors, compromised by economic and political compulsions, to play the role of gatekeepers of news. It was hoped that public accountability would emerge from the networked nature of the new media. Several examples of citizen journalism enabled by social media were hailed as harbingers of a new era of news. This vision of social media as a democratizing actor was based on the ideal that it would be open, neutral, egalitarian and also enable genuine public driven engagement. Google News, Facebook's News Feed which tries to put together a dynamic feed for both personal and global stories, and Twitter's trending hashtag feature, have brought forward these services as key drivers of an emerging news ecosystem. Initially, new media was hailed as a natural consequence of the Internet, which would enable greater public participation, allow journalists to find more stories and engage with their readers directly.

Over time, it has been realized that far from being open, neutral or egalitarian, social media platforms introduce their own parameters to shape how information is accessed, which only amplify issues plaguing the mainstream media. This new knowledge logic in effect replaces human judgement (as earlier exercised by editors) to some kind of proxy decision-making based on data. There is little evidence to suggest that the latter is any more democratic in its character than the former and creates new problems of its own. Research has clearly demonstrated that Facebook's feed, X's trending topic and Youtube's recommendations, all prioritize extremist stories over other kinds of content. For instance, the algorithm for the trending topics depends not on the volume, but the velocity of the tweets with the hashtag or term. It could be argued that given this predilection, the algorithms will rarely prefer complex speech or content of a more complex nature.

For a democratic society to thrive, individuals need to be active participants in discourse and not passive recipients of information. Social media platforms view us primarily as consumers, and not citizens. Their single-minded drive to appeal to our basest and narrowest set of stimuli may make good business sense, but that does no favours to the cause of democracy. As citizens, we need to be exposed to more than the most agreeable or extreme form of our

still-evolving opinions. The signal we give to algorithms through likes and clicks are often only a fleeting or tentative take on an issue. A democratic society needs media and platforms that allow us to explore different perspectives and arguments before we make up our minds. Instead, algorithms seize on our half-baked opinions and hasten their crystallization. It is bad enough that our online selves drive this propaganda, but lately, politically aligned actors are making creative use of such platforms to inundate us with misinformation, hate speech and polarizing content designed to manipulate.

Research and investigations over the last decade have clearly established the nature of algorithmic recommendation engines deployed by platforms which favor extremist, viral and provocative content, at the very heart of the content dissemination problem, with clear impacts on democratic processes. By focussing public debate on their receding investment in transparency and content moderation, the large platforms were able to effectively stay clear of any real discourse of algorithmic accountability measures to reverse this trend.

We desperately need to move the conversation squarely to the regulation of tech platforms and technologies. Regulation must establish transparency and accountability, including explicitly addressing the internal recommendation engines which prioritise, elevate and promote some of the most harmful content–such as hate speech, misinformation, and material inciting racial violence–all under the guise of boosting “user engagement” and ultimately driving profit.

Recommendations

The 2024 elections megacycle and subsequent alignment of Big Tech CEOs with the incoming Trump administration in the United States demonstrated the limits of voluntary self-regulation by Big Tech to address the harms their platforms and technologies cause to democracy and human rights across the world.

There is a need for a democratic and human rights-based regulatory architecture with binding accountability mechanisms, transparency requirements, and enforceable standards across operational contexts. We make first recommendations on a series of components for such architecture here. These will necessarily be built upon and updated on an ongoing basis, as our collective understanding of what constitutes effective tech regulation evolves, including localisation to different governance and institutional contexts.

For this to happen, there is a need for civil society to mobilise across regions, and for there to be continued support for the research and media ecosystem fundamental to an understanding of the relationship between digital platforms and technologies, democracy and society.

As the Global Coalition for Tech Justice, we call on defenders of human rights and democracy worldwide to unite in establishing accountability for Big Tech platforms and technologies:

To democratic policymakers and regulators

Establish mandatory rules for transparency and accountability of digital platforms and technologies through democratic and human rights-based platform governance mechanisms. This should include:

• Measures to address content policy, product and algorithmic design risks and impacts on elections and human rights. Particular emphasis should be placed on addressing recommendation systems that may amplify illegal or substantively harmful content, disinformation, and polarization, mitigating for differential impacts across linguistic, cultural, and regional contexts.

• Data access frameworks for researchers and journalists, ensuring free or affordable data access across platforms.

• Electoral integrity protocols, including rapid response to emerging threats, cross-platform coordination and post-election continuity provisions to address post-electoral risks such as violence.

• High standards of transparency and compliance obligations for advertising and monetisation.

• Mandatory labeling, real-time detection and reporting obligations for platforms in relation to AI-generated content, particularly in election periods and in respect of deceptive deepfakes.

Foster normative convergence on platform governance standards across geographies to prevent inequitable outcomes and mitigate the risks of regulatory arbitrage. This may be assisted by inclusive multistakeholder processes, internationally recognized and enforceable standards for platform governance during electoral periods, and establishing oversight mechanisms that systematically evaluate implementation consistency across diverse global contexts. These standards should uphold international human rights laws and conventions, including protections against targeted harassment of marginalized communities and historically vulnerable groups, with particular attention to contextual application of safeguards across different sociopolitical environments. Electoral protection mechanisms should require platforms to implement a full spectrum of electoral integrity tools across all operational jurisdictions, irrespective of market prioritization.

Require binding cooperation agreements between Electoral Management Bodies and Big Tech platforms, and resourced partnerships for electoral integrity, including with fact-checking organizations, independent media entities, civil society groups, and electoral integrity bodies–including requirements for respect for the independence of partners, standardized engagement reporting, and cross-platform collaboration mechanisms to optimize trust and safety investments.

Establish gender-sensitive electoral safeguards requiring Big Tech to implement specialized mechanisms addressing technology-facilitated gender-based violence in electoral contexts, with particular emphasis on expedited review protocols for synthetic media targeting candidates based on gender characteristics. Such regulatory interventions should mandate transparent reporting on enforcement metrics disaggregated by gender to facilitate analysis of implementation efficacy and potential disparate impacts across diverse electoral environments.

To the European Commission, continue strong enforcement of the EU's digital rulebook, including the Digital Services Act, Digital Markets Act and European Media Freedom Act, and use all available tools to allow a return to a healthier online environment, not just in Europe but globally. We call on you and all EU institutions to stand in solidarity with other democratic countries that implement rights-respecting legislation on Big Tech.

To the United Nations, OECD, international and multi-stakeholder bodies

Work with civil society to secure binding business and human rights obligations backed by international penalty regimes for Big Tech platforms, particularly where they operate in authoritarian, democratically challenged, fragile and conflict-affected states. All international institutions and bodies should work to reduce dependency on Big Tech and build democratic, equitable and inclusive digital infrastructure.

To Advertisers and investors

Use your combined power to pressure Big Tech platforms into maintaining global fact-checking partnerships, restoring and strengthening policy protections for vulnerable groups, investing in global platform safety and compliance with business and human rights standards, and establishing transparency to the ad-tech system. Tech-facilitated destabilisation of democracies and abuses are bad for business and antithetical to corporate social responsibility – play your part to stop this. In addition, we call on you to invest in and support independent media as well as digital infrastructure consistent with human rights and democratic values.

To Civil Society Organizations, activists and users of digital platforms

Continue mobilizing and campaigning for globally equitable and inclusive corporate tech accountability, supporting vulnerable groups targeted as Big Tech weakens protection, advocating for corporate accountability standards across different regions, providing direct support to vulnerable groups affected by harmful platform practices, and continuing to report platform enforcement inconsistencies.

To Researchers and Independent Media

Advance evidence-based understanding of Big Tech effects by researching, investigating and exposing Big Tech's impacts on democracy, human rights and society worldwide, provide evidence-based research, investigative reporting, and scholarly analysis to expose tech-enabled harms, create shared knowledge infrastructures on platform governance disparities, and develop interdisciplinary methodologies for analyzing algorithmic amplification of electoral manipulation.

2024 Country Election Case Studies

JANUARY

Taiwan

Taiwan's January 2024 elections opened the year's election megacycle. The country's presidential and parliamentary races were the testing ground for China's evolving prowess in online disinformation and electoral interference.

Taiwan, the self-ruling island democracy that China claims as its territory, headed to the ballot to elect a new president and members of parliament on 13 January 2024. Incumbent President Tsai Ing-wen from the Democratic Progressive Party (DPP), concluded her second and final term in May 2024 and could not run again.

Candidates from three main parties were running for president, with the ruling DPP presidential candidate, Lai Ching-te, leading in the polls. The established opposition party Kuomintang (KMT) and the relatively new Taiwan People's Party (TPP) were fighting for second place, experts said.

The ballot came at a time when China had escalated military activity in the Taiwan Strait and other nearby waters. As in previous election cycles, China tried to influence the election result, at the time framing Taiwan's presidential race as a choice between "peace and war, prosperity and decline."

Social media: Key facts and trends

Taiwan has a population of over 23 million, and some estimates put the number of active social media users in 2023 at a whopping 21.5 million.

Facebook dominates the social media landscape, with almost 50% of adults using the app, as of 2023. However, Facebook's messenger was facing serious competition from LINE, the instant messaging app, operated by Japanese internet company LY Corporation. It was the second most used app in Taiwan, followed by Instagram – attracting less than 20% of users – Twitter and the Taiwanese online discussion forum PTT.

When in November 2023, Meta announced general plans for how it was going to keep elections it became clear there were almost no changes in its approach, compared to previous years. The one new policy Meta said was going to be applied globally was the requirement for advertisers to disclose when they use AI or digital methods to create or alter a political or social issue ad – and only "in certain cases". Meta had no country-specific plans for ensuring the safety and integrity of the Taiwanese ballot.

While Facebook was most widely used, YouTube was used more frequently on a daily basis, with Taiwanese internet users spending almost 1.5 hours on the video-sharing platform each day. At the time, YouTube's owner, Google, only published a plan for how it's going to approach the 2024 US ballot.

TikTok and Instagram were most popular among younger internet users. In December 2023, TikTok launched its in-app "2024 Election Guide" for Taiwan. The project was a partnership with MyGoPen, a Taiwan-based organisation certified by the International Fact-Checking Network (IFCN). The short video app also provided channels to report content that might breach electoral rules.

A recent history of disinformation in Taiwan

Chinese disinformation and/or influence operations have become a fixture in Taiwan. And the 2024 election was no different. For decades, China has tried to sway Taiwanese voters, including through online campaigns, with the aim of one day peacefully annexing Taiwan. In the past such

campaigns sought to portray Beijing in a positive light, appeal to Taiwanese voters to vote for pro-China candidates, and push Taiwanese voters to not vote at all – a strategy that has had little effect.

But in the 2024 election cycle China's approach became more subtle and, some say, more effective. Rather than a blunt "don't vote for Tsai" message, China's misinformation campaigns had focused on stirring up scepticism about the United States' sustained support for Taiwan's continued independence.

More specifically, there was an increase in AI-generated content, including deep fake videos, targeting pro-independence candidates and stoking fears of conflict, disinformation from bots and Beijing-friendly online influencers trying to persuade voters that the U.S. was not a dependable partner. Mandarin-language friendly TikTok became the main source of AI-generated deep fakes and election-related disinformation videos, according to Eve Chiu, the CEO of Taiwan FactCheck Center.

"There's been rumours of election fraud in Taiwan [on TikTok]. TikTok is not the most popular social media app in Taiwan but Taiwanese people share TikTok videos on Facebook and Line. Many of the election and politics-related videos have come from TikTok recently," she said. "I think it will affect election results because many people still believe in this hoax."

Making matters worse were Meta's and Google's algorithms which favour sensationalist, polarising and toxic content, helping push out misleading and false information to potential voters. "The media ecosystem, including the algorithm, prefers low quality information over high quality information. Low quality like disinformation, misinformation, fake news, sensational stuff, rumours, hoax," said Chiu, whose organisation was part of Facebook's Third Party Fact-Checking program at the time. Chinese disinformation had also piggy-backed on domestic issues, like power outages, attacking the Taiwanese state and its democratic processes.

One of the election-related misinformation campaigns had already kicked off in May 2022 on Facebook, TikTok and YouTube, according to a December 2023 report by research firm Graphika. The operation involved a network of over 800 fake accounts and 13 pages about Taiwanese politics. Though Graphica wasn't able to identify who was behind the campaign, the suspect social media posts promoted the opposition KMT party, that's seen as pro-China, and slammed its opponents, including the ruling DPP, which favours Taiwan's independence.

"The content closely tracked Taiwan's news cycle, quickly leveraging domestic news developments, such as controversies surrounding an egg shortage and the alleged drugging of toddlers at a kindergarten, to portray the KMT's opponents as incompetent and corrupt," Graphika researchers wrote.

Other false narratives peddled across Taiwan's social media platforms in the past have played up food safety and vaccine concerns, war and the possibility of a conflict across the Taiwan Strait.

Researchers have also warned that domestic platforms are increasingly being used to spread falsehoods, with half of the false narratives originating inside Taiwan, the other half in China. They identified PTT, the Taiwanese online discussion forum, as one of the favourite platforms for China-borne disinformation.

Taiwan's National Security Bureau reportedly tracked at least 1,800 pieces of online disinformation in 2023 – up from 1,400 in the same period in 2022. All of the items came from social media platforms,

both in Taiwan and abroad, including TikTok, Instagram, Facebook, and YouTube.

Mandarin-language monopolisation

Next to disinformation, China's Mandarin knowledge monopolisation – which toes China's Communist Party line – is also undermining Taiwan's democratic integrity and polluting the information space of Taiwanese voters, according to Tzu-wei Hung, a research fellow at Taiwan's Academia Sinica.

The aim of this two-pronged approach is for China to peacefully annex Taiwan further down the line. "Monopolising Mandarin knowledge is a long-term infiltration of Taiwanese teens and children, such as using TikTok to create a China-friendly information space. TikTok may lead to digital addiction and affect children's mental health, making it easier to manipulate or echo Xi Jinping's dog whistle of "the great revival of the Chinese nation," Tzu-wei Hung said.

This was reflected in Google search results, which reached the majority of Taiwanese internet users thanks to its 90% share of the market. And so, Tzu-wei Hung said, Google search in Mandarin will generate less diverse and numerous results than the same search in English, dishing out a pro-China narrative to Taiwanese users. "Pro-Beijing websites often dominate Google's search results in traditional Chinese because they are well-managed in search engine optimization and comply with Google's anti-content-farm spam guidelines. Google also derives considerable ad profits from the internet traffic they generate," he said.

In his research, Tzu-wei Hung also found that China's domestic censorship has spilled over into AI-generated content in Silicon Valley. "For example, ChatGPT and Microsoft Bing answers differently in Mandarin and English to questions involving sensitive keywords like the Tiananmen incident or the Uyghur genocide. Google Bard repeatedly says it is unable to answer in Mandarin about Falun Gong, organ harvesting, and the Dalai Lama," he said.

FEBRUARY

Indonesia

On 14 February 2024, Indonesia, the world's fourth largest democracy and the most populous Muslim country, held general elections. Over 164 million people cast their ballot in the world's biggest single-day ballot to elect a president, vice president and almost 20,000 representatives to national, provincial and district parliaments.

The presidential election was a three horse race between Prabowo Subianto, the defence minister under the incumbent president Joko Widodo, known as Jokowi, is leading in the polls, Ganjar Pranowo, a former governor of one of the country's most populous provinces and Anies Baswedan, the ex-governor of Jakarta. The race came on the heels of a decade-long rule by Joko Widodo, who stepped down.

In the end, largely thanks to an elaborate social media campaign, Prabowo Subianto, an alleged human rights abuser with ties to Indonesia's most famous dictator won almost 60% of the vote, becoming the country's next president.

The island nation is Southeast Asia's largest economy and a key partner for the United States in its ambition to thwart China's influence in the region.

Social media: Key facts and trends

Indonesia has 275 million inhabitants and some estimate that around 224 million people were

accessing the internet in the country in 2022. The figure is expected to grow to about 270 million by 2028.

Facebook is the preferred social media platform, with almost 42% of Indonesians using the app as of 2022. This is followed by Instagram, which is used by 29%, YouTube accounts for 10% and TikTok almost 9%. Indonesia is also home to TikTok's second largest user base – after the US – with an average scrolling time of 29 hours per month.

An economic powerhouse, Indonesia has become an attractive destination for Meta, TikTok, X/Twitter, Google, YouTube and other tech companies, all of which have signed up to Indonesia's strict content law, which was announced in 2020. Campaigners have warned the rules threaten freedom of expression and amounted to a compromise by the big tech sector seeking to retain access to an important market. Authorities can order content that disturbs "society" or "public order" to be taken down and demand access to company data.

AI-generated content

Like in the Bangladesh and Taiwan elections, which took place in January 2024, in Indonesia too AI-generated deepfakes became a fixture across social media platforms in the leadup to the ballots. The most startling of all, and possibly most popular, included the AI-generated video of General Suharto, the Indonesian dictator who passed away in 2008.

"I am President Suharto, the second president of Indonesia, inviting you to elect representatives of the people from Golkar," the digital Suharto said in the video, which was posted on 6 January on Instagram and X/Twitter by Erwin Aksa, the party's deputy chairman. Reportedly created by the Golkar party, the video has been viewed over 4.5 million times.

But the Suharto deepfake aside, all three presidential candidates and their running mates appeared in AI-modified videos, which some said had the potential to influence the outcome of the election.

Despite the deluge of AI-generated disinformation on social media, the country's General Elections Commission (KPU) declared that it has no jurisdiction to regulate AI technology. The hands off approach has made the situation worse, according to Rizka Herdiani, researcher at Center for Digital Society (CfDS). "Since the General Election Commission or KPU didn't establish any regulation on the use of AI in the election, its presence has accelerated and amplified the spread of misinformation and/or disinformation on social media," she said.

And Big Tech companies were not doing enough to curb the spread and impact of AI-generated disinformation, said Nuurrianti Jalli, Assistant Professor of Professional Practice at the School of Media and Strategic Communications at Oklahoma State University. "Platforms should label AI content, proactively and strictly implement mis/disinfo policies, and be accountable for failing to immediately react to misuse of their platform during politically charged periods such as elections."

Pandering to Millenials

In an effort to appeal to young voters – over 52% of Indonesians were aged between 18 and 39 years old – all three presidential candidates took to social media long before the official campaign kicked off in November 2023, with some of their videos racking up millions of views.

Video-sharing platforms like Tik Tok and Instagram had quickly become battlegrounds for the youth vote, according to experts. Analysts also pointed to the widespread use of buzzers known as trolls or

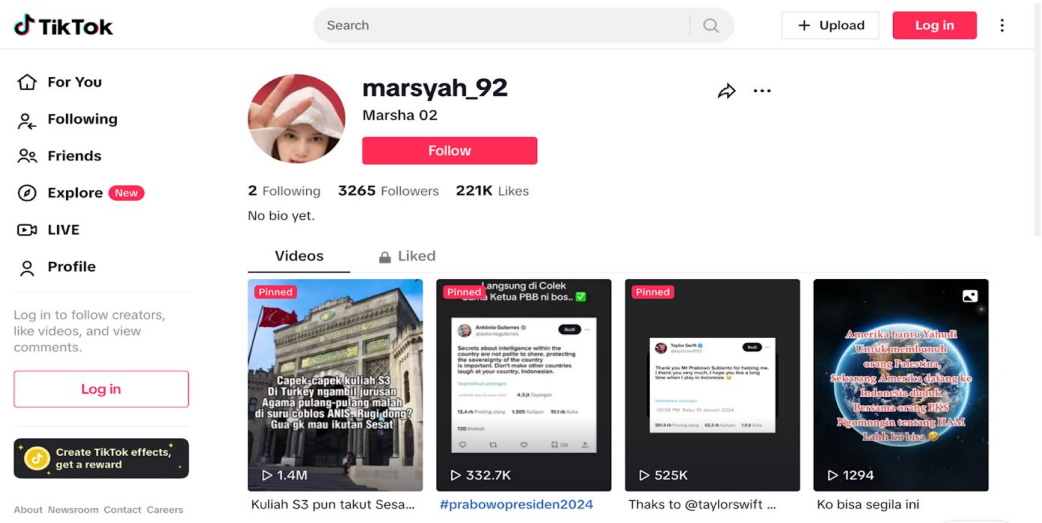
cyber troops, who were paid to spread falsehoods and tarnish the public's opinion of the candidate. This included rumours that the educational credentials of Prabowo's running mate – the son of the incumbent president Jokowi – were fake.

TikTok has also become a powerful image-changing medium, best demonstrated by the makeover of presidential frontrunner Prabowo Subianto, the leader of the Gerindra party – the third largest party in parliament – and the sole contender with ties to the Suharto dictatorship which ran from 1967 and 1998. Prabowo is a former special-forces commander, dismissed from the military following allegations of involvement in torture, and Suharto's former son-in-law.

But that's most likely not how he's being viewed by younger voters who know nothing about his past. To most of Indonesia's TikTok users he's become a "cute" grandpa dancing awkwardly in videos – some of which have gone viral. Experts have warned that the ubiquitous use of TikTok has reduced politics to memes and videos that put personality over policy, further raising concerns about the state of Indonesia's democracy, which has suffered under president Jokowi.

"Some voters choose a certain candidate for bite-size "information" or clips of the presidential debates that are posted on platforms such as TikTok. So, yes, it can influence the results of the elections," said CfDS's Rizka Herdiani.

The video sharing platform is also being used to spread AI-generated disinformation on a mass scale. One example included an AI-modified TikTok video of a social media post by musician Taylor Swift purportedly thanking the 72-year-old Prabowo Subianto – who won the presidential race – "for helping me". The video had been in circulation at least between 31 January 2024 and 7 February 2024, and was taken down following a viral Tweet that sounded alarm about the hoax. Before it disappeared from TikTok the video had racked up over half a million views. [Link to TikTok profile with the video here]



Screenshot of TikTok profile, which shared a video of a social media post by Taylor Swift purportedly thanking Subianto is third from the right.

Copyright: Digital Action

Screenshot of TikTok profile, which shared a video of a social media post by Taylor Swift purportedly thanking Subianto is third from the right. Copyright: Digital Action

"Based on this case alone, it's the younger generation that is becoming prone to misinformation and/or disinformation instead of the older generation. Voters of a particular presidential candidate are easily swayed by this "clickbait-y" content that even if there are other users (in this case, users in the X platform) who provide substantial information that is fact-checked, they are not willing to review them," Herdiani said.

TikTok has reportedly been working with Indonesia's elections oversight body to stop the spread of falsehoods. While officially TikTok doesn't allow political ads on its platform, before the 2022 US midterm elections it had approved 90% of ads containing false information about voting, submitted by the rights watchdog Global Witness.

Disinformation in minority languages

Like elsewhere in the world, in Indonesia's elections minority languages were also proving a stumbling block for social media platforms with potentially disastrous consequences, experts have warned.

"I can see an increase in hate speech and misinformation related to the election, especially in local dialects on social media platforms like Facebook groups, TikTok, X, and YouTube. Like in the previous election, propaganda, including hate speech and misinformation, may influence public opinion and political behaviours, including voting," said Nuurrianti Jalli of Oklahoma State University. "Tech companies have implemented various efforts, but there is still room for improvement, particularly in dealing with regional languages and social contexts. Collaboration with local/regional experts during the election could help significantly."

Transparency disclosures by Big Tech companies in the European Union (EU) and Australia have shown that social media platforms have been chronically under-resourcing content moderation in local languages as compared with the English language. For example, only 8% of X's (formerly Twitter's) content moderators are proficient in an official EU language other than English. While at YouTube only 11% of EU language moderators were reviewing posts that weren't in English.

Other Indonesia experts were more critical, saying that Big Tech companies simply weren't ready to safeguard the democratic and information integrity of the February 2024 ballot, partly because of the companies' precarious relationship with the authorities. "[T]his election cycle is actually worse than the last cycle - platforms are not set up to handle problems, and they are not being responsive and proactive enough. And that's a very dangerous sign," Raman Jit Singh Chima, Asia policy director at advocacy group Access Now reportedly said about Big Tech platforms in Indonesia, India and Bangladesh

APRIL

India

The world's biggest general election kicked off on 19 April 2024 in India, where nearly one billion people were eligible to cast their ballots. The vote was preceded by a polarising campaign, during which the ruling Hindu nationalist Bharatiya Janata Party (BJP) sought to secure a third consecutive term. Spread over a seven-phase voting period, it concluded on 1 June 2024.

In power for a decade, BJP, led by Prime Minister Narendra Modi, was hoping to win by a landslide, thanks to welfare payouts to the most needy and amid a campaign steeped in violent rhetoric against Muslims. But the election results, announced on 4 June 2024, shocked the world as BJP lost its single-party majority in the Lok Sabha (the lower house of Parliament).

Months after the ballot, our analysis of evidence from a myriad of investigations and media reports, revealed that despite being lauded as a "successful" exercise in democracy in which the BJP failed to win its expected landslide, Muslim citizens lost their lives and election laws were broken in a context of Big Tech compliance and platform safety failures.

India's ballot and the post-election period were steeped in offline and online hate speech and accompanied by violence against Muslim voters. Such violence, which social media platforms have helped normalise, is not a hallmark of democracies.

Neither India's electoral laws, nor the tech companies' own policies appear to have mattered throughout. Even though India's laws prohibit references to religion and creating "communal disharmony" during election campaigning, this was fully ignored by the politicians and Big Tech companies. Social media platforms have operated with impunity in deciding which democratic processes they want to follow, and which they want to ignore – it is clear India's election wasn't a priority.

Platforms also failed to meaningfully engage with civil society groups throughout the election period and made next to no efforts at being transparent – there were no announcements on how election policies were being implemented and no stock taking after the ballot concluded.

These systemic failures are at the heart of tech harms to democracy and human rights across the globe. They're also at the core of the platforms' toxic business model, which incentivises harmful and hate-filled content and prioritises ad revenue over democratic integrity and human lives. Unless and until that model changes, the already devastating impacts of social media platforms on people and elections will only become more severe.

Social media landscape

Home to over 460 million social media users, India is among the largest and most significant social media markets globally. Platforms have played a key role in political campaigning since at least 2019, when a third of the country's population had access to social media for the first time.

Meta-owned WhatsApp is the most often used social network and messaging app in the country, followed by YouTube, Facebook and Instagram. With almost half a billion active users in 2024, India is the encrypted app's largest market. While all platforms have been employed by bad actors to spread falsehoods and incite violence, WhatsApp and YouTube were at the forefront of electoral disinformation and hate speech.

Some ascribe YouTube's prominence in India to the fact that the authorities banned TikTok

in 2020, which until then had 200 million users. This prompted many younger social media users to turn to YouTube and Instagram. India has also emerged as a petri dish for tech harms and their often life-ruining consequences – the social media trends and adverse impacts of tech platform failures to secure information integrity in India have been harbingers of things to come elsewhere in the world. This has been the case partly due to a number of factors, including 1) India's ever growing online population, 2) immense investment into 'digital' by the Indian government and both domestic and foreign companies, 3) India, being a low rights jurisdiction with poor on-ground protection of human rights, and 4) absence of robust regulatory hurdles/ controls such as data protection law, online harms regulation and health-tech and financial credit related regulations. Social media trends: Online hate, real-world violence BJP politicians and other Hindu nationalist hardliners have been inciting violence against India's Muslims offline and online for years, but their hate speech intensified in the leadup and during the election. The problems of online hate in India had pre-dated the 2024 elections and in some cases led to real-life violence. In 2013, for example, a misleading video in northern India incited riots and five years later a spate of mob lynchings were linked to messages that circulated on WhatsApp groups.

Additionally Big Tech companies have been a part of India's political landscape for almost a decade and in that time have gained awareness and understanding of the track record of the ruling BJP – a supremacist party with a documented history of anti-Muslim hate speech – in stoking social polarisation and weaponising vitriol to win votes.

Despite this knowledge, tech platforms routinely failed to enforce their own content policies. They neither stymied nor removed posts fomenting enmity or hatred between different classes of citizens on grounds of religion, race, caste, community, or language – which violated India's election laws. These failures have meant that videos, ads and other social media posts attacking and calling for violence against India's Muslims have become a fixture, fostering an environment of hate and societal approval for the dehumanisation and violence against Muslims. This in turn has led to Muslim men and women being violently attacked, and in some cases tortured and killed. Such violence is not a hallmark of democracies.

The impacts of Big Tech failures have been long-lasting, continuing well after t he election concluded on 1 June 2024, including in states ruled by the opposition Congress party, as part of a purported retaliation for BJP's loss of majority. Families had their homes razed to the ground and at least three men were beaten to death after being tortured. Like with past attacks against Muslims, the violence was often recorded and shared on social media, where it went viral.

While much of the anti-Muslim violence appears to be a result of a years-long environment of hate, offline and online, th ere is a gr owing body of evidence linking social media to specific harms perpetrated by those espousing supremacist Hindutva views. The attacks are often documented by perpetrators or onlookers and shared online, where they cause further distress and suffering to members of already marginalised communities.

In August 2024, vigilantes from the right-wing group Hindu Raksha Dal (HRD) assaulted Muslim families in two separate incidents in northern India – both recorded and shared on HRD's social media accounts. "They were so angry, went to everyone's house, and asked if they were Muslims or Hindus," a witness of one of the attacks reportedly said. "Hindus were spared, and Muslims were beaten brutally."

The violence took place after BJP politicians and other hardliners called for revenge on social media for violence against Hindus in Bangladesh

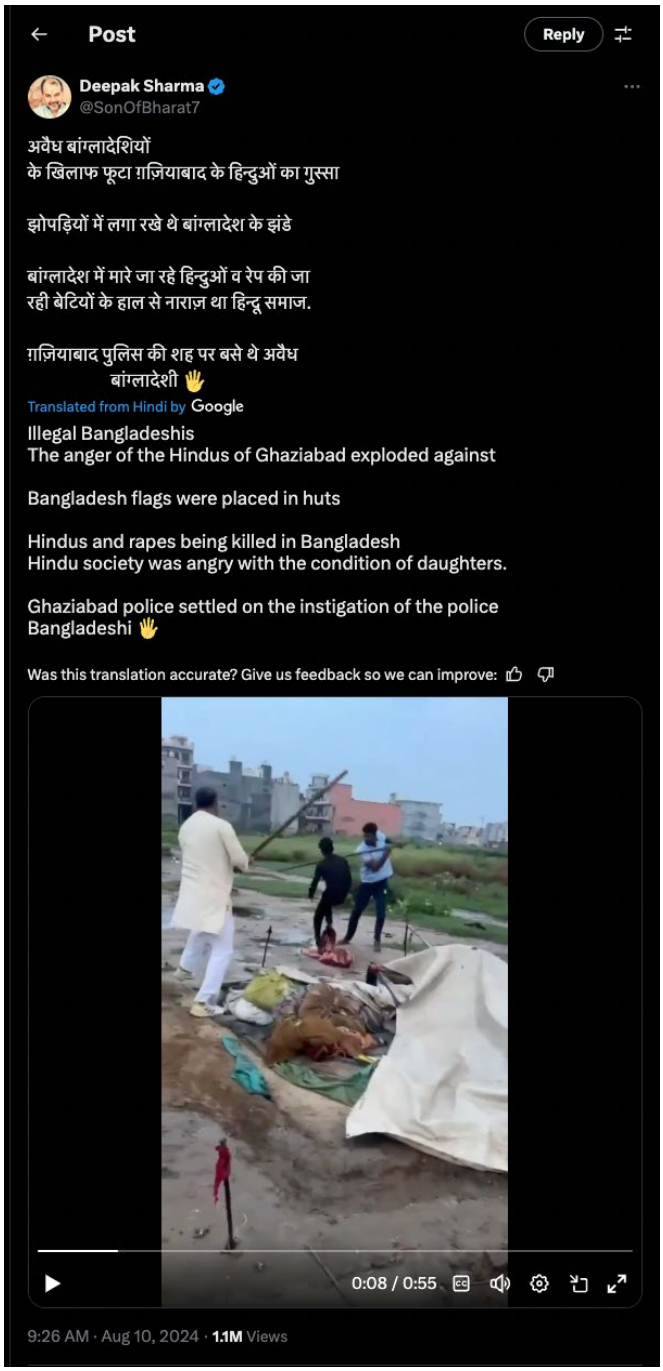
following the escape of Bangladesh's Prime Minster from the capital Dhaka on 5 August 2024. At the time a wave of deadly anti-government protests had swept through Bangladesh.BJP's Nitish Rane reportedly called for outright murder, tweeting: ***"If Hindus are targeted and killed in Bangladesh, why should we allow even one Bangladeshi to breathe here. We will also target and kill."*** Though at the time of writing the tweet appears to have been deleted.

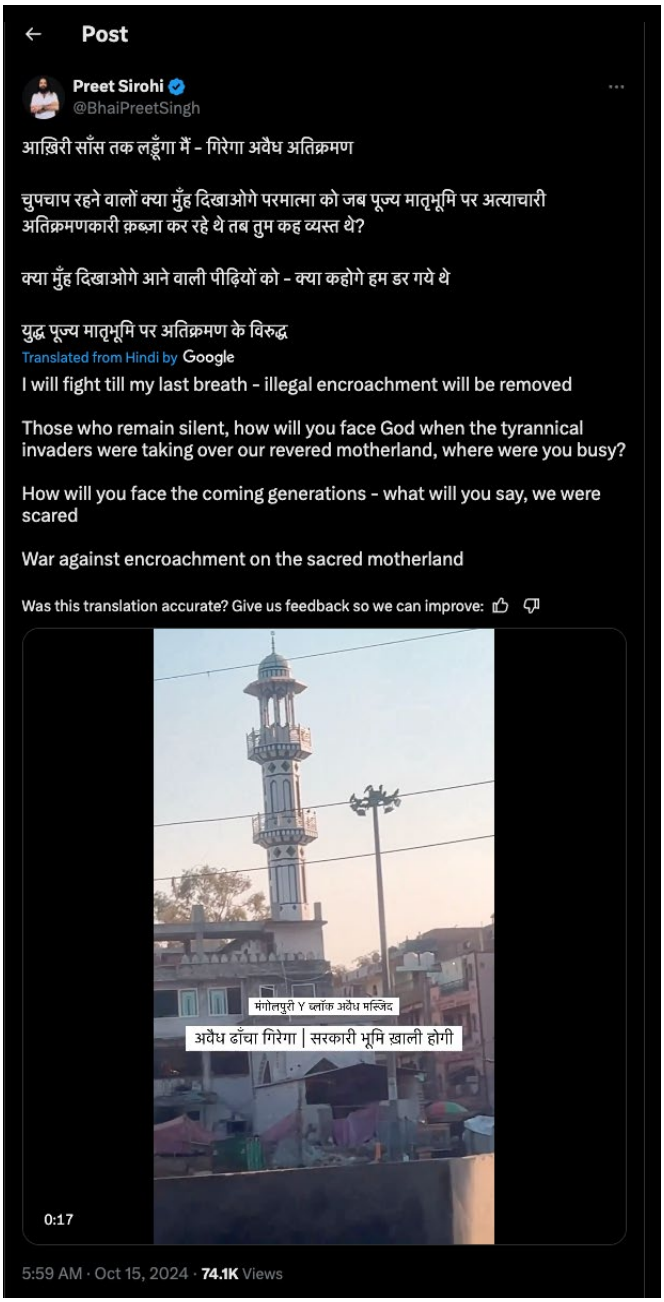
Another BJP hardliner, Kangna Ranaut, in her X post drew comparisons between India and Israel, saying that "now we are also covered by extremists. We must be ready to protect our people and our land." "Peace is not air or sunlight that you think is your birthright and will come to you for free...Pick your swords and keep them sharp, practise some combat form everyday," said Ranaut's post which at the time of writing had 1 million views.

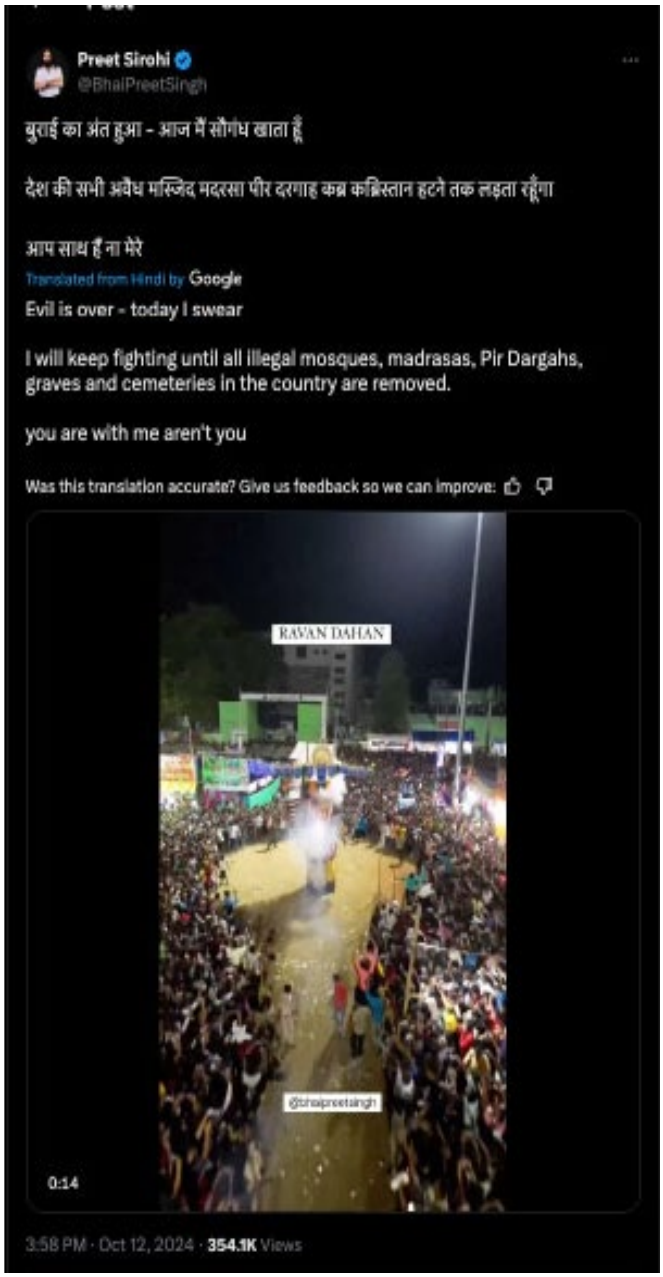
Five days later HRD thugs, led by Daksh Chaudhary were seen abusing the Muslim residents of Delhi's Shastri Nagar district. "Go to your Bangladesh!" Chaudhary shouted in the video which was shared on social media. (Chaudhary is known for hate speech and violence. In June 2024, he was arrested after the ballot for abusing the residents of the northern city of Ayodhy after BJP lost there.) The second assault by an HRD mob took place on 9 August 2020, when thirteen men attacked Muslim families in a shanty town of Ghaziabad, a city north of the capital Delhi. This attack was also filmed and reportedly first shared on the group's social media accounts, and by others. One video of the assault shared on X, garnered over 1 million views.

Other instances of violence include offline and online calls for the eradication of mosques by vigilante influencers like Preet Sirohi. Sirohi keeps filing complaints with local authorities, claiming mosques he wants demolished were built illegally. In June 2024, he was behind the campaign to demolish two mosques in Delhi in a week.

Sirohi's hateful campaign has continued at the time of writing. In October 2024, Sirohi vowed to keep on demolishing mosques, claiming they are illegal and sharing posts with inflammatory language, calling for the destruction of mosques. "Those who remain silent, what face will you show to God, where were you busy when the tyrannical encroachers were occupying the revered motherland?", his October 2024 tweet reads.


Post
Reply
Deepak Sharma
@SonOfBharat7
अवैध बांग्लादेशियों
के खिलाफ फूटा ग़ाज़ियाबाद के हिन्दुओं का गुस्सा
झोपड़ियों में लगा रखे थे बांग्लादेश के झंडे
बांग्लादेश में मारे जा रहे हिन्दुओं व रेप की जा
रही बेटियों के हाल से नाराज़ था हिन्दू समाज.
ग़ाज़ियाबाद पुलिस की शह पर बसे थे अवैध
बांग्लादेशी
Translated from Hindi by Google
Illegal Bangladeshis
The anger of the Hindus of Ghaziabad exploded against
Bangladesh flags were placed in huts
Hindus and rapes being killed in Bangladesh
Hindu society was angry with the condition of daughters.
Ghaziabad police settled on the instigation of the police
Bangladeshi
Was this translation accurate? Give us feedback so we can improve:
0:08 / 0:55
9:26 AM · Aug 10, 2024 · 1.1M Views


Post
Preet Sirohi
@BhaiPreetSingh
आख़िरी साँस तक लड़ूँगा मैं - गिरेगा अवैध अतिक्रमण
चुपचाप रहने वालों क्या मुँह दिखाओगे परमात्मा को जब पूज्य मातृभूमि पर अत्याचारी
अतिक्रमणकारी क़ब्ज़ा कर रहे थे तब तुम कह व्यस्त थे?
क्या मुँह दिखाओगे आने वाली पीढ़ियों को - क्या कहोगे हम डर गये थे
युद्ध पूज्य मातृभूमि पर अतिक्रमण के विरुद्ध
Translated from Hindi by Google
I will fight till my last breath - illegal encroachment will be removed
Those who remain silent, how will you face God when the tyrannical
invaders were taking over our revered motherland, where were you busy?
How will you face the coming generations - what will you say, we were
scared
War against encroachment on the sacred motherland
Was this translation accurate? Give us feedback so we can improve:
मंगोलपुरी Y ब्लॉक अवैध मस्जिद
अवैध ढाँचा गिरेगा | सरकारी भूमि ख़ाली होगी
0:17
5:59 AM · Oct 15, 2024 · 74.1K Views



The details: how platforms helped foster a violent environment

Content moderation and policy enforcement

Modi peddled an islamophobic conspiracy theory around the so-called "vote jihad", which echoed the alt-right "great replacement theory" – implying Muslims were in India to replace Hindus – further disenfranchising India's Muslims and dissuading them from voting. The narrative echoed across all social media platforms, with pro-BJP Facebook pages becoming a "ubiquitous part of Indian election campaigns as political parties and their leaders seek to directly connect with their followers," according to Usha M. Rodrigues, a professor at Charles Stuart University.

Facebook was at the heart of a countrywide hate-filled campaign. Researchers examined posts shared in 812 Facebook pages and in 15 Facebook groups between 1 March and 10 May 2024. In that sample alone, they found over 50 posts inciting hostility between Hindus and Muslims, with a large share of the content boosting Prime Minister Modi's speeches.

"In most cases, it was only after large and blatant policy failures were made public through reports and press articles that companies took action. In multiple cases, that action came too late to protect the election," the diaspora organisation India Civil Watch International (ICWI) said.

While Meta's blog post boasted having 20 local languages in India covered by content moderators – out of a total of 780 – that obscured the fact that the measures were neither new nor sufficient to protect the integrity of the elections. In fact, the tech companies' election measures have remained largely the same over the years, "despite mounting evidence that they must be adjusted", said corporate

accountability watchdog The London Story.

Similarly, WhatsApp appears to have failed to implement its content moderation policies, allowing BJP-affiliated accounts to spread conspiracy theories, anti-Muslim rhetoric and hate speech free from public scrutiny (The party reportedly operates at least 5 million WhatsApp groups in the country). This included messages glorifying police brutality, which spread like wildfire within minutes

Failures of the tech platforms also resulted in electoral misinformation, with conspiracy theories and falsehoods about politicians and opposition candidates circulating on Facebook, WhatsApp YouTube, and X.

YouTube failed to remove policy-violating content, according to a report by ICWI, Dalit Solidarity Forum, Indian American Muslim Council, Hindus for Human Rights and Tech Justice Law Project. The groups reported to YouTube that 26 videos, published between April and May 2024 by the far-right Sudarshan News YouTube channel, violated the platform's hate speech and misinformation policies – they contained conspiracy theories, hate speech dehumanising Muslims and misinformation about the opposition Congress party. Instead of taking the videos down, YouTube shared ad revenue with the channel.

The lack of meaningful engagement with civil society groups has been a hallmark of the 2024 election season in India and beyond. Multiple members of the Global Coalition for Tech Justice who have monitored the social media space in the leadup and during the ballot reached out to YouTube and Meta to act on violating and harmful content, and take it down. But in most cases, they said, platforms failed to take any action and when they did the hate speech or disinformation in question had already spread far and wide.

X's crowdsourced fact-checking programme Community Notes, which allows X contributors to write context notes for misleading tweets, also proved to be a complete fiasco, leaving tweets with outright falsehoods without any notes or disclaimers. This included a tweet by a verified X user, who falsely claimed that a non-existent Dubai-based Association of Sunni Muslims was financially supporting Muslims travelling to the southwestern state of Karnataka to unseat the BJP and Modi. There was no Note accompanying the tweet, even though fact-checkers flagged it as containing false information.

Failure to enforce ad policies social media platforms, including YouTube, also failed to implement their own ad policies. Corporate accountability rights groups tested YouTube's systems for ad review and ad approval mere days before the election kicked off, only to find out they didn't work.

Global Witness and Access Now submitted 48 ads in English, Hindi, and Telugu, all breaching YouTube's advertising and election misinformation policies – the ads contained disinformation meant to suppress voter turnout among youth and women, and content inciting violence against the Muslim minority. (It was based on already existing falsehoods specific to India.) Although YouTube purportedly reviews all ads prior to publication, it approved all the ads, which the groups withdrew post approval by the platform.

YouTube was at pains to defend itself, saying the ads would have been reviewed by a moderator before going live. But when Global Witness tested YouTube's ad approval system in English and Spanish ahead of the 2022 US elections, the video-sharing platform rejected all the ads at the first stage and suspended the host channel. Such glaring disparities in the treatment of harmful content by YouTube proves that the company has

the means to enforce its policies, but it chooses not to do so in Global Majority countries, even in major markets like India – something civil society groups have been flagging for years.

Breaches of campaign financing laws India's campaign financing laws may have been violated due to lack of transparency across social media platforms.

Political advertising in India, including on social media, is subject to a number of content and financing regulations. To skirt around the rules and evade the scrutiny of the Election Commission, political parties and actors allegedly resorted to surrogate and shadow advertising, thus undermining the integrity of India's ballot.

Surrogate and shadow advertising involves an individual or a group placing social media ads on behalf of a candidate or a political party without disclosing that the ads are directly funded by said party. Surrogate advertisers are typically identifiable by their tax number but shadow advertisers are less transparent and further obfuscate who's behind purchased advertisements.

Corporate accountability rights groups Eko, The London Story and ICWI reviewed Meta's Ad library in the leadup to India's vote and uncovered a network of 22 far-right shadow advertisers, who purchased ads with memes and videos attacking the opposition, fomenting violence against the Muslim minority and promoting the BJP. The adverts amassed more than 10 million interactions in over 90 days and appeared to breach Meta's own content and transparency policies.

The 22 shadow pages spent over $1 million on advertisements for BJP and in particular Prime Minister Modi, which accounted for almost 22% of the total sum of election advertisements during a 90 day period before the ballot.

The researchers couldn't establish who owned the pages due to lack of verifiable information. The pages had no active phone lines and the researchers' emails to addresses indicated on the shadow pages remained unanswered.

Meta's failures to implement its own policies on political advertising go beyond the national election. A group of rights groups, including Global Coalition for Tech Justice member ICIW, documented how Meta has allowed BJP and its shadow pages to violate India's electoral law and Meta's own policies ahead of the November 2024 local elections in the north-eastern state of Jharkhand. The researchers identified at least 87 shadow pages purchasing ads spreading BJP narratives in Meta. These pages have been churning out nearly five times more ads and garnering four times more impressions than the official BJP Jharkhand page.

AI-generative content and system failures

It should be noted that, despite widespread fears that AI-generated content would inundate social media with falsehoods, in India deepfakes were predominantly used to troll rather than to launch information warfare.

This was a blessing for tech platforms, which despite their pledges to prioritise and fight generative AI misinformation failed a simple stress test by civil society groups, which proved that some platforms can't detect AI-manipulated content.

In May 2024, corporate accountability groups Ekō and ICWI created inflammatory and Islamophobic political ads and submitted them to Meta's Ad library – the repository for ads published on Facebook and Instagram. Some of the ads contained AI-manipulated images and all of them were created based on real hate speech and disinformation prevalent in India. One ad, for

example, used inflammatory language targeting Muslims and read: "Hindu blood is spilling, these invaders must be burned".

Out of the 22 ads the groups submitted in English, Hindi, Bengali, Gujarati and Kannada, Meta approved fourteen – all containing generative AI images. The ads were approved despite containing harmful content, violative of the platform policies, and despite the fact that Big Tech giants pledged to prioritise tackling harmful generative AI content. (The adverts were submitted halfway through the six-week voting period and withdrawn post approval, so that they would not be circulated.)

Researchers monitoring social media throughout the six-week voting period have also raised concerns about the lack of clarity around the fact-checking of political advertising across Meta's platforms. Some ads that were flagged by fact-checkers, they said, remained on the platforms, while others did not contain labels indicating they had been fact-checked, creating further confusion.

MAY

South Africa

On 29 May 2024, some 28 million voters cast their ballots in what many describe as South Africa's most contested general election since the end of the apartheid regime in 1994. Overall, 70 political parties and 14,903 candidates were fighting for 887 seats in the national and provincial legislatures. And for the first time since taking power three decades ago, the governing African National Congress (ANC) lost its outright majority, receiving 40% of the vote.

The ANC had been under pressure due to high unemployment, which hit 32% in 2023, economic inequalities, corruption allegations, high levels of violence and frequent power cuts. The main opposition Democratic Alliance (DA) party saw this as a crisis, while the third largest party in parliament, the Economic Freedom Fighters (EFF) sought to attract voters with plans to redistribute land to the less well-off.

But it was the recently formed uMkhonto we Sizwe (MK) party, led by ex-president Jacob Zuma – ousted from the ANC amid allegations of corruption – which became the ultimate election disrupter. Named after the ANC's former armed wing, the party promised to create 5 million jobs and engaged in a vicious online disinformation and intimidation campaign. Despite the fact that Zuma spent time in jail after ignoring a court order and other legal problems, the MK party received 15% of the vote.

On 20 May 2024, South Africa's top court disqualified Jacob Zuma from running. The court ruled that he wasn't eligible to be a member of nor qualified to stand for election to the National Assembly. (The case stems from the decision of South Africa's Independent Electoral Commission

(IEC), which disqualified Zuma in March 2024, saying the Constitution bars anyone with a prison sentence from running for office. Zuma challenged that decision.) At the time, experts warned that the Constitutional Court's ruling would affect the results of the ballot and could lead to security issues. (More information on threats of political violence below.)

Social media landscape

In January 2024, South Africa had 26 million active social media users – accounting for 42% of the total population – and rapidly advancing artificial intelligence (AI) tools. Meta's WhatsApp was the most popular social media platform, utilised by almost 94% of active social media users, followed by Facebook (88%), TikTok (74%) and Instagram (67%).

Social media trends and threads ahead of the 2024 ballot

Electoral mis-and-disinformation

Civil society organisations and the Independent Electoral Commission of South Africa (IEC) had been sounding the alarm about election-related disinformation in the leadup to the 29 May 2024 ballot. Falsehoods that could harm or deter voters from casting their ballot have become a fixture over the years and have been proliferating on WhatsApp, Facebook, X (formerly Twitter) and Instagram, among others.

Some of the most common falsehoods on social media about elections included claims that if one is registered to vote but doesn't, their vote automatically goes to the governing party. Falsehoods also targeted the independence of the IEC. In February 2024, several political parties accused the IEC of frequently hiring members of the South African Democratic Teachers Union (SADTU) as voting station staff, alleging that, as SADTU is an alliance partner of the ANC, these staff would rig the election.

In February 2024, Bantu Holomisa, the President of the United Democratic Movement (UDM), tweeted: "The IEC uses this Union to run this country's elections, an affiliate of Cosatu which is in alliance with the ANC. This time in our meeting of the opposition parties on 26/2/24, we must take a resolution on this rigging of SA elections". Like elsewhere in the world, AI-generated disinformation, including deepfakes, has been on the rise in South Africa – including in the context of the 2024 election – and South Africans reportedly struggled to spot them. In March 2024, Duduzile Zuma-Sambudla, the daughter of former president, Jacob Zuma, shared a deepfake video on X (formerly Twitter), in which ex-US president Donald Trump urges South Africans to vote for Zuma's MK party.

"Greetings all South Africans. My name is President Donald Trump. I urge all South Africans to vote for uMkhonto WeSizwe on 29 May. The African National Congress of Cyril Ramaphosa has failed all South Africans. With this new black party, led by President Jacob Zuma, all South Africans will matter," Trump says. The post, which was shared over 500 times and garnered over one thousand likes on X is still on the platform at the time of writing.

AI-generated misinformation on Meta's WhatsApp messaging app, which has become the main means of communication across the country, has been of particular concern to experts. In rural areas, WhatsApp provides access to information on a range of issues, including reproductive health. But it has also become a major conduit for disseminating and consuming misinformation.

However the Trump video deepfake shared by Duduzile Zuma-Sambudla was only the beginning of her disinformation campaign. Zuma-Sambudla, who has over 300 thousand followers on X was also behind the most dangerous disinformation push on the social media platform, falsely claiming that the

election was rigged in favour of the ANC. One of her posts, shared days before the election, showed pictures and videos of what appeared to be ballot boxes with a caption accusing the ANC of "stealing votes". The post, which was viewed almost 650 thousand times, remains on X at the time of writing.

Calls for election violence South Africa has a history of political violence. While law enforcement was better equipped to handle instances of electoral and post-election violence than in previous years, authorities and experts had raised concerns about some politicians' threatening rhetoric.

"As the government, we want to issue a stern warning to anyone with intentions to disrupt the elections that the law enforcement officers will deal with them decisively and will put them behind bars," Minister of Defence and Military Veterans Thandi Modise reportedly said in April 2024.

In the leadup to the ballot, members of Jacob Zuma's MK party have ramped up threats of violence should they not get their way at the polls, or should the court decide to disqualify Zuma from the race. "If these courts, which are sometimes captured, if they stop MK, there will be anarchy in this country. There will be riots like you've never seen in this country. There will be no elections," MK's leader Visvin Reddy said in March 2024 in a video, which was widely circulated on social media.

Reddy's statement, for which he was charged with inciting public violence, is one of many made by MK party members in public and shared on social media platforms, including X (formerly Twitter) and TikTok. One TikTok video, since removed, reportedly showed a man wearing an MK shirt firing a pistol into the hills, followed by a camera pan to a table with a shotgun and assault rifles. MK's purported ties to violent individuals became very real in January 2024, when a group of over sixty men and women charged with instigating deadly riots in 2021 in KwaZulu-Natal, turned up to court sporting MK regalia.

The violence of 2021, in which over 300 people lost their lives, was sparked by the decision of a local court to imprison former president Jacob Zuma for contempt of court. In January 2024, South Africa's Human Rights Commission confirmed that social media platforms played a key role in fuelling the violence through amplification of inciting posts.

The investigation found that through "dissemination of inflammatory content, social media amplified grievances, stoked fear and anger, and mobilised individuals towards disruptive actions." "It was clear from the evidence obtained that mechanisms to gather information to counter the weaponization of these platforms are available. However, the responsible entities [authorities] did not take steps to improve their skills, neither did they have the capacity to do so at the time," the Commission said in the report. Concerns over political violence were raised again on 20 May 2024, after South Africa's top court disqualified the face of the MK party, ex-president Jacob Zuma from running in the upcoming elections. The court ruled that Zuma wasn't eligible to be a member of nor qualified to stand for election to the National Assembly. During the proceedings Zuma challenged the independence of some of the court justices.

"These narratives attacking the independence of the Constitutional Court are reminiscent of the narratives that played out both online and offline challenging the legitimacy of his contempt of court conviction which led to violent riots in July 2021 to stop his arrest," said Sherylle Dass, Regional Director of the public law firms Legal Resources Centre (LRC), adding that "there is a real risk that similar calls through concerted online campaigns may result in unrest, particularly in KwaZulu-Natal."

The MK party was also accused of stoking tribal divisions. In February 2024, Jacob Zuma reportedly referred contemptuously to KwaZulu-Natal residents, raising concerns that the newly formed party could be hijacked to promote Zulu nationalism. Zuma's comments were followed by social media posts allegedly shared by MK party supporters emphasising that the party represents the Zulu nation. At least two political parties have filed complaints – one criminal and one with the IEC – alleging intimidation by MK members. Though antagonism ran both ways, and social media videos appear to show MK supporters assaulted by ANC members. Allegations of intimidation were also made against other political parties.

Xenophobic statements and incitement to violence

Months before the 29 May 2024 ballot, rights groups and experts had been sounding the alarm about narratives conflating xenophobia with patriotism spreading across social media, and warning about the exploitation of anti-immigration rhetoric by political parties and possible violence.

Migration, and especially irregular migration, emerged as one of the central campaign themes, with politicians of all denominations resorting to incitement to violence. Candidates blamed South Africa's social ills on migration and made xenophobic statements. In December 2023, for example, leader of ActionSa, Herman Mashaba, alleged in a tweet – without providing any evidence – that foreign nationals who run tuckshops use them as illicit drug channels, destroying small businesses and disrupting entire communities.

Government officials also joined the xenophobic chorus. After an August 2023 fire in a building in Johannesburg's central business district that killed more than 70 people, Kenny Kunene – the deputy president of the Patriotic Alliance and member of the Mayoral Committee – called for the "mass deportation of illegal immigrants who are staying in abandoned buildings that are taking rent." In April 2024, the ANC Minister for Home Affairs approved a White Paper recommending that South Africa withdraw from the 1951 Refugee Convention, only to ratify it again with reservations.

"Politicians are using immigrants as pawns, without regard for their safety in an attempt to score votes ahead of the general elections," Nomathamsanqa Masiko-Mpaka, South African researcher at Human Rights Watch said in a May 2024 statement.

"We are noticing that the anti-migrant rhetoric is coming from political parties – particularly smaller ones – and some like the Patriotic Alliance have been particularly hateful, like the leader saying he will remove oxygen tanks from sick foreigners in hospitals," said Yasmin Rajah head of Refugee Social Services in KwaZulu-Natal, a coastal South African province. "I don't think social media platforms do much to stamp out hate speech. It seems like anything goes in South Africa."

By then the rhetoric already translated into real-world action in KwaZulu-Natal, with pamphlets warning foreign shop owners to close their businesses and a truck driver's forum mobilising to protest against foreign truck drivers, according to Rajah.

Social media platforms approving hateful ads

In June and August 2023, South Africa's Legal Resources Centre (LRC) and international non-governmental organisation Global Witness – both members of the Global Coalition for Tech Justice – tested social media platforms' safety efforts, by submitting extreme and hate-filled advertisements for approval. The June 2023 investigation tested Facebook, TikTok and YouTube ability to detect hate speech and incitement to violence targeting non-nationals in South Africa. The groups prepared ten adverts – which were withdrawn post-approval

and never published – based on real-life content in English and translated into Afrikaans, Xhosa and Zulu. The ads called on the South African police to kill foreigners and encouraged violence through "force" against migrants. The ads were approved by all three social media platforms, with the exception of only one ad in English and Afrikaans rejected by Facebook.

The platforms' approval of hateful ads was of particular concern, bearing in mind the violent and sometimes deadly attacks on migrant workers living in South Africa. Xenowatch reported 170 such incidents in 2022 and 2023 and 18 between January and April 2024.

South Africa has been a home for millions of African migrants, mainly from Lesotho, Nigeria, Malawi, Mozambique, and Zimbabwe for decades. While xenophobic sentiments have been simmering in the country for years, they came to a boiling point in 2022 with the launch of Operation Dudula. Meaning to "force out" in the Zulu language, this social media hate campaign spilled over into the streets of Johannesburg and elsewhere, unleashing violent protests, arson of migrant-owned businesses and leading to the murder of a Zimbabwean national.

Hate speech targeting female journalists

The LRC and Global Witness August 2023 investigation, in addition to those platforms tested previously, included X (formerly Twitter). The ads submitted for approval were based on real-world examples of abuse received by women journalists. These were violent, sexualised and dehumanising, referring to women as "vermin", "prostitutes", or "psychopaths" and calling for them to be beaten and killed. They included statements such as "they're just all sheep and should be slaughtered" and "they all need to die". Despite the extreme content of the adverts that clearly breached the social media platforms' own policies on hate speech, all four platforms approved the vast majority of the adverts. The findings indicate significant gaps in the platforms' ability to effectively moderate harmful content, particularly when it involves nuanced, culturally specific language.

Big Tech policies and plans for keeping South Africa's ballot safe

Ahead of the vote on 29 May 2024, Meta, TikTok and Google rolled out more or less the same measures they had resorted to in previous years, with some updates. All three companies have signed a voluntary agreement with the Electoral Commission of South Africa (IEC) and civil society group Media Monitoring Africa (MMA), undertaking to work together to combat disinformation and other digital harms ahead of the elections.

Though X (formerly Twitter) is mentioned in the agreement, the tech platform hasn't signed on – a move which aligned with changes implemented by free speech absolutist and billionaire Elon Musk, who bought the company in 2022.

The framework, which wasn't legally binding, set out to combat the dissemination of disinformation and relied on the good faith of the participants to work together to ensure free and fair elections. The agreement also encouraged the platforms to implement their own policies regarding removal of problematic content, publishing advisory warnings on harmful content and delisting of content. It ended on the date election results were announced.

As part of the framework, the tech companies committed to assisting IEC with media literacy programmes and to working with the IEC and MMA on misinformation complaints through platforms such as Real411.org and PADRE.org.za.

Real411 was an official website for reporting mis and disinformation in the lead up to the elections, which shared complaints directly with the IEC. The IEC then assessed each complaint and was supposed to take appropriate action. This included notifying social media platforms, which were expected to ensure "a diligent response," according to MMA director William Bird.

However, despite IEC's official commitment and efforts, some observers raised concerns about the Commission's lack of action around threats of election violence. The IEC never officially acknowledged the possibility of violence in the leadup to the ballot. Some alleged that the offline and online statements made by MK politicians violated the Electoral Code of Conduct by using language that provokes violence or intimidates candidates or voters.

The IEC was also criticised for not doing enough regarding inciting anti-immigration rhetoric propagated by political parties and candidates. "The Electoral Commission of South Africa, as an independent constitutional body which manages free and fair elections, should explicitly condemn the harmful rhetoric directed towards foreign nationals," said Nomathamsanqa Masiko-Mpaka, South Africa researcher at Human Rights Watch in a statement.

Past failures and 2024 concerns

In 2021, social media companies agreed to a similar framework for the purposes of safeguarding local elections, but they fell short of their obligations, a report penned by former MP-turned disinformation expert Phumzile van Damme has found.

Civil society groups had been trying to engage with the platforms for months leading up to the 29 May 2024 ballot, but the companies refused to do so, raising serious concerns about their commitment to safeguarding the election.

The South African National Editors Forum (SANEF) has expressed "anger at being 'ghosted' by big tech companies and our Parliament's Portfolio Committee on Communications", after the group's requests to discuss how to combat disinformation and hate speech during the 2024 ballot were met with silence. "Despite two reminders, by April no acknowledgement had been received from TikTok or X (formerly Twitter). Meta provided a vague response to revert in due course, but six weeks later had not done so," the group wrote in an April 2024 press release.

Concerned by the lack of transparency and concrete information explaining how Big Tech companies planned to safeguard the ballot, public interest law firm LRC submitted two access to information requests to Meta, Google and TikTok on their election action plans. In its request, LRC asked for substantive information on content moderation and emergency tools available in respect of the South African elections. But all three platforms refused to divulge any details, indicating that South African access to information laws did not apply to them because they weren't headquartered in the country.

"The Legal Resources Centre (LRC) is concerned by the lack of transparency and unwillingness to engage with civil society organisations seeking substantive information around their election plans. Despite TikTok's formal response, they did subsequently provide the LRC with some of the information requested albeit in very general terms," said Sherylle Dass, LRC's Regional Director. "The refusal by social media companies to provide information on their management of the South African information ecosystem, on the basis that the companies are registered in another jurisdiction is incredibly problematic and undermines democratic processes. It appears impossible to properly hold the companies to account for their acts and omissions relating

to the management ***of the online space during South Africa's elections,"*** said Bulanda Nkhowani Campaigns and Partnerships Manager for Africa at Digital Action, convenor of the Global Coalition for Tech Justice.

Social media companies' failures to prioritise platform guardrails in global majority countries have led to the amplification of content inciting violence and spreading disinformation, often with catastrophic consequences for the most vulnerable members of society. In Brazil, Myanmar, Ethiopia, Tunisia and South Africa, among others, platforms' failures to stamp out harmful content have translated to real-world violence – including attempts to undermine democracy – and even deaths. Companies like Meta, TikTok, Google and X (formerly Twitter) have known about the grave impacts of their failures to act for years and can't claim ignorance – they have a responsibility to keep people and elections safe.

"***This includes averting any amplification of violent inciting content pre, during and post-ballot, alongside increased public transparency in South Africa, including responding to institutional and civil society requests for information regarding the management of the country's information ecosystem,"*** Digital Action's Nkhowani added.

JUNE

Mexico's June 2024 Elections: Context and key facts

Some 100 million voters were expected to cast their ballot on 2 June, 2024, in Mexico's largest election to date. The voters, including a record number of first-timers, elected a new president, more than 600 members of parliament, 9 governors and over 20.000 local officials. It was a landmark ballot as Claudia Sheinabaum became the first woman to be elected president of Mexico, taking over the reins from her party colleague Andrés Manuel López Obrador from the governing Morena coalition.

Scheinbaum ran against former senator and tech entrepreneur, Xóchitl Gálvez of the main opposition alliance, Strength and Heart for Mexico. The third running mate was Jorge Álvarez Máynez from the Citizen's Movement, who polled in single digits.

But the ballot also made history for being Mexico's most violent elections yet, with 30 candidates murdered, over 70 threatened and 11 kidnapped. The violence has affected politicians from across the spectrum, and one study has found that over 200 civil servants, politicians and candidates had been killed in the leadup to the June vote.

Some commentators have raised concerns about the increasingly blurred lines between the state and organised crime, calling into question the governability of certain regions.

Social media landscape

As of April 2024, Facebook was the most often used social media platform in Mexico, followed by WhatsApp and Instagram – all owned by Meta. In May 2023, when WhatsApp was the most popular social media app in Mexico, 90% of internet users proclaimed having an active account on the messaging platform. But video sharing apps YouTube and TikTok are attracting ever more internet users.

Over 60% of Mexicans get their news and information from social media, Reuters Digital News Report has found, with YouTube and TikTok as the fastest growing platforms for news in the country.

Despite the rapid expansion of social media platforms, Mexico has been marred by a digital divide which means that not everyone can access social media. And while electoral disinformation has been a huge concern ahead of the 2 June vote, only 63 million of Mexico's 126 million inhabitants have internet access.

Social media trends

Electoral disinformation

As polarisation and tensions rose ahead of the ballot, online disinformation surged. One post which made the rounds on social media falsely claimed that the ruling Morena party wanted voters to use new identity cards linked to what it called "a Venezuelan fraud company." The London-based firm reportedly had no contracts in Mexico for the June election, and there were no plans to use a new ID card during the ballot.

In another instance, a photograph allegedly showed a tortilla shop in northern Mexico, being closed down after refusing to use wrappers with the logo of the ruling party candidate for governor. But the picture in fact featured an entirely different business, from another part of the country, and dated back to 2020.

"We see a deliberate strategy by all political actors, of all political campaigns, to exacerbate polarisation," Abraham Trejo, coordinator of the Hate and Harmony project at the College of Mexico, told France 24.

Only days before the ballot, some voters exposed to misinformation remained confused about how to mark their ballot or about whether the ink from the pens at voting stations could be erased post-ballot.

"It can be very serious in places where people don't have full internet access, only to WhatsApp and Facebook," and cannot verify what they receive, tech harms expert Martha Tudon of Article 19 told Radio France International.

Experts sounded the alarm about disinformation hitting the June 2024 elections hard already one year prior to the vote. In June 2023, when candidate selection at Mexico's ruling Morena party and the main opposition kicked off, The Associated Press press said its Spanish-language fact-checkers "found about 40 fake publications across social media platforms, favouring or discrediting members of both sides of the political spectrum."

At the time, experts had raised concerns that some of the falsehoods targeting the opposition were coming directly from President Andres Manuel Lopez Obrador.

Facebook groups became the gateways to WhatsApp and Telegram chat groups, where experts said falsehoods were widely disseminated with little to no oversight. The entire ecosystem consisted of publishers, bots and trolls which help amplify the falsehoods and ordinary – organic – users who took the bait without fact-checking the information they were sharing.

Disinformation targeting presidential candidates

Fact-checkers had their hands full, with Associate Press teams debunking numerous falsehoods about presidential candidates Claudia Scheinbaum of the governing coalition Morena and her opponent from the opposition Xóchitl Gálvez.

Some rumours included old videos or footage edited out of context and accusing Galvez of planning to scrap the social programmes rolled out by the outgoing president Andres Manuel Lopez Obrador. Social media posts targeting Scheinbaum, who has Jewish roots, alleged she was planning to make circumcision compulsory and turn the revered Basilica of Guadalupe into a museum.

Both women were also accused by internet users of having lied about their university degrees.

Anti-immigrant rhetoric

Immigrants were also dragged into the disinformation war, with some alleging that Scheinbaum bought them with promises of social assistance from the government, should she win. This claim was debunked by the National Election Institute – only a small fraction of naturalised citizens are able to cast a ballot.

Content take-downs and censorship

As the ballot approached, rights watchdogs, including Article 19, Access Now and R3D, warned that politicians and other powerful individuals have been weaponizing local laws and regulations to take down social media content, which exposed alleged wrongdoing or questionable behaviour. This form of censorship occurred when journalists sought to shine a light on matters they believed were of public interest, including politicians making lewd comments.

In February 2024, presidential candidate Jorge Álvarez Mayens of the Citizen Movement shared a video on his instagram account, in which he made derogatory comments about other politicians. After criticism online, Álvarez took down the video, but not before others had downloaded it and shared it across social media. Journalists who reported on

the video and used the footage received content take-down requests from YouTube and Instagram alleging copyright infringement.

The digital content production company Badabun enabled Álvarez to request the content take-down. Right groups alleged that the company had helped a number of politicians delete unfavourable content from social media, invoking copyright infringement laws created to protect commercial interests – but instead being deployed as a censorship tool.

Article 19, Access Now and R3D also raised concerns about the use of provisions meant to tackle "political gender-based violence" to silence voices criticising politicians, who as public figures should be subject to greater scrutiny and criticism than regular citizens.

One of the most notable cases involved Delfina Gomez, a candidate for governor of Mexico state, who in 2022 launched complaints with electoral authorities against a number of X (formerly Twitter) users, alleging their posts amounted to political gender-based violence. The social media posts in question referred to Gomez as an "electoral criminal" and a "thief". The electoral tribunal ordered X and other platforms to remove the posts about Gomez, citing threats to her political rights.

"When applying electoral regulation, it is necessary to adapt the principles and rules regarding freedom of expression, so that they are consistent with digital environments," said Martha Tudón of Article 19. ***"The following kinds of speech are especially protected by the human right to freedom of expression: Political speech and on matters of public interest; the speech about public officials in the exercise of their functions and about people candidates for public office; and speech that forms an element of the personal identity or dignity of the person expressing it. In these types of speech, the weight of freedom of expression increases and its limitations become even more exceptional, so any limitation on them must be subject to a level of strict scrutiny."***

Big Tech policies and plans for keeping Mexico's elections safe

Ahead of the 2 June 2024 vote, Meta, TikTok and YouTube published blog posts about the election, which for the most part mirror their previous election plans. But for a few additions, they highlight existing policies and emphasise information literacy and fact-checking initiatives.

Meta agreed to collaborate with the National Electoral Institute (INE) – one of the bodies tasked with ensuring free and fair elections. As part of this agreement Meta launched a chatbot on WhatsApp, allowing voters to report possibly false or inaccurate information about the election. Tik Tok has also partnered with INE and launched an Electoral Guide, which is meant to help voters access election-related information.

TikTok also signed an agreement in October 2024 with the Electoral Tribunal of the Federal Judiciary, which stipulated that the video sharing app will "contribute to the communication of relevant and truthful public electoral information...and discourage disinformation."

"However, it is important to note that these agreements are not reflected in any changes to the platforms' policies. On the contrary, companies interpret their policies extensively to implement these collaborations without acknowledging any legal accountability or obligation between the platform and the authority, no binding responsibilities to users" said Agneris Sampieri, Latin America policy analyst at Access Now.

But such agreements and cooperation between electoral authorities and social media platforms,

coupled with lack of transparency around the rules for such cooperation means that right groups ***"remain in the dark about their impact on electoral contexts,"*** Sampieri added.

Cooperation agreements between tech companies and electoral authorities as well as tech platforms' paltry policy updates remain a constant trend when it comes to Global Majority Countries. Like elsewhere in the world, in Mexico, Big Tech has also failed to prioritise platform guardrails and this inaction has resulted in more amplification of electoral disinformation and other types of harmful content.

Social media companies should be more proactive in terms of identifying and debunking electoral disinformation directed at candidates and about elections, as well as AI generated content, and also roll out 'break the glass' measures in case of violence escalation, as done in countries like Brazil and India.

"Mexico's ballot is yet another example of how much we lack proactive and meaningful engagement emerging from the companies themselves. Brazil has seen a similar movement with companies announcing policy changes - or reaffirming old ones - only after being provoked by the Superior Electoral Court. We need to move beyond the empty commitments to big tech companies actually helping safeguard voters and electoral processes," said Bruna Martins dos Santos, Global Campaigns Manager at Digital Action, Global Coalition for Tech Justice Convener.

SEPTEMBER

Jordan

Jordanians went to the polls on 10 September 2024 to elect the members of the lower house of parliament. The election saw more than 1,600 candidates fighting for the votes of over 5 million registered voters, of which 53% were women.

It was a historical ballot for a number of reasons; It resulted in the moderate Islamist opposition making significant gains, in part due to anger in Jordan over Israel's latest war in Gaza. The Islamic Action Front (IAF), the political arm of the Muslim Brotherhood in Jordan, won 31 out of the 138 parliamentary seats, tripling its representation in the lower house.

But it was also a test for the new electoral and political party law meant to encourage increased participation of women, minorities and political parties in general in parliament. Still, tribal politics and discriminatory attitudes towards women have reportedly resulted in some female candidates being prevented from running for office by their tribes and the Independent Election Commission.

The European Union election observers concluded that the election was "inclusive" and "well-run". Though only 1.6 million Jordanians cast their ballots amid low interest in the election, media and news sites failed to provide voters with the necessary information. The EU delegation also "observed that during the campaign, journalists operated under multiple legal restrictions to freedom of expression included in the Cybercrime Law and the Penal Code."

It also has to be noted that although Jordan's lower house is elected, it wields limited legislative power, which is still subject to the King's approval since Jordan is a constitutional monarchy ruled by the King. The monarch, who plays a dominant role in politics and governance, appoints all members of the cabinet and the upper house of parliament

While Jordan remains one of the more liberal countries in the Middle East, Freedom House described it as "not free" in its 2023 Freedom Report, and others have referred to the Kingdom as a liberal authoritarian state.

Social media landscape

Some estimate that 6.4 million people – over half of Jordan's 11 million population – have social media accounts. YouTube appears to be the most often used platform, followed by Facebook, Instagram and X (formerly Twitter).

TikTok was banned by Jordanian authorities in late 2022 after footage from truck drivers' protests, in which one policeman was killed, had circulated on the platform. The authorities justified the suspension – which was supposed to be temporary – saying TikTok had failed "to deal with publications inciting violence and disorder." But it is by no means the sole platform affected by censorship.

Freedom of expression has been heavily policed and curtailed in Jordan, including online. Jordanian authorities have a track record of instigating internet shutdowns, banning or blocking websites and social media apps. As of 2023, they banned around 300 websites, social media platforms and applications.

Social media platforms still played a central role in campaigning in 2024, allowing candidates to communicate with potential voters and sidelining traditional media. Platforms like Facebook, X and WhatsApp have emerged as essential campaign tools and, analysts said, will change the way politicians reach and influence voters in Jordan in the future.

Shrinking civic space

Civic space in Jordan has been shrinking year on year. Throughout 2023 and 2024, the authorities continued quashing dissenting voices, arresting and harassing journalists and critics of the government.

In August 2023, Jordanian authorities overhauled the country's decades-old cybercrime law. Packed with vague and undefined terms like "***fake news***" and "***online assassination of character***," the law includes criminal penalties for broadly defined online speech and introduces additional punishments for the use of circumvention tools, like VPNs.

The amendments came at the time when the authorities were ramping up persecution and harassment of citizens organising peacefully and engaging in political dissent.

Heavily criticised by domestic and international rights groups at the time of its promulgation as potentially repressive, the amended cybercrime law has lived up to its potential. It "***undermines free speech, threatens internet users' right to anonymity, and introduces a new authority to control social media, risking a surge in censorship***", Human Rights Watch said in a statement at the time.

Between August 2023 and August 2024, the authorities charged hundreds of individuals, including activists and journalists, under the law for social media posts criticising the authorities, expressing support for Palestine, criticising Jordan's peace deal with Israel, or calling for peaceful protests. Among those arrested and now facing trial is Ayman Sanduka, who addressed his October 2023 Facebook post to the King, criticising Jordan's diplomatic relations with Israel.

Ever since Israel began its genocidal attack in Gaza in response to the deadly 7 October 2024 attack on Israeli citizens by Hammas, thousands of Jordanians have taken to the streets to peacefully express their support for and solidarity with Palestinians in Gaza.

Big Tech failures

Despite heightened tensions in the region triggered by Israel's genocidal campaign in Gaza and a countrywide crackdown on peaceful protests, neither Meta, nor Google/YouTube or X rolled out or communicated about any plans meant to safeguard Jordanians and the September 2024 ballot. This stood in stark contrast to specific announcements Meta and Google issued about the measures the companies have taken in 2024 to protect people and ballots in the leadup to the European Parliament, USA or India elections. (X hasn't really communicated about its guardrails ever since Elon Musk took over the platform in 2022.)

This omission was yet another illustration of the fundamental lack of equity in how platform safety is being conducted by Big Tech companies. Publicly available election integrity plans are the exception rather than the rule for Global Majority countries, despite the significant user base and high risks of tech-related harms to people and democratic processes across many countries. In Jordan, tech harms that were already prevalent, escalated in the months leading up to the vote.

Censoring pro-Palestine and other voices

Tech platforms, and Meta in particular, have a yearslong track record of content moderation failures in the Arabic language and of silencing pro-Palestinian content from across the globe. In Jordan, Meta's censorship policy is no different. At least 90 Jordanian journalists reporting on Palestine and the protests in Jordan in support of Palestine and against Israel's genocidal campaign in Gaza, had their Instagram or Facebook accounts blocked and/or removed, according to

an interview we conducted on 20 August 2024 with a Jordanian human rights activist, who asked that their identity be concealed for fear of reprisals.

When a human rights activist working for a trusted flagger organisation intervened on behalf of some of the journalists, Meta allegedly said it had blocked or removed the accounts because they violated Instagram community guidelines. The journalists allegedly violated the rules on dangerous individuals and organisations, which Meta has been using to censor pro-Palestinian voices for years, most notably during the 2021 Gaza war. According to the human rights activist, some accounts were removed due to false reporting.

Jordanian activists who have been engaging with Meta, have raised concerns about the lack of transparency around the implementation of the tech platform's election and other policies. This also includes Meta's refusal to share how many employees are monitoring which countries in the region. The silencing of pro-Palestine voices and other speech across its platforms has led some to conclude that the company's content moderators for Jordan are not from the country and lack understanding of the local context. "***Content moderators in Meta's team aren't aware of the context. People from different countries might not be aware of the nuances in Jordan, even if they had been raised in the region***," said one activist, who requested that their name be withheld for fear of retaliation.

Activists also said that since Israel's most recent war in Gaza, Meta blocked and/or restricted Facebook pages supporting the Palestinian cause. Meta's transparency reports were silent on the matter.

"***Since the war on Gaza started, Meta has been restricting freedom of speech. Journalists had their posts removed, their [Facebook] pages are getting blocked, pages of political parties are getting removed, student groups have been affected by these things,***" a human rights activist speaking on conditions of anonymity told Digital Action. "***We can't talk about democracy in Jordan if Meta is cherry picking the opinions that are going online. That's affecting how social media companies are shaping public opinion that is also shaping our political life. In Jordan freedom of speech is restricted generally, you're getting arrested for protests and then when you're silenced on social media, the whole system collapses.***"

Endangering rights defenders and activists

Meta's transparency reports show that Jordanian government's requests for users' data have surged exponentially in 2023 – the last year data is available – rendering it the highest to date. The authorities have been seeking to gain more access to users' data year on year. Between January and December 2023 they requested data of nearly 2,300 users or accounts, as compared to 652 in 2020. In 2023, the tech giant produced information for up to 33% of government requests. In previous years, the rate varied between 50% to 60%.

This is important and highly problematic, according to a human rights activist interviewed by Digital Action, who asserts that Meta is failing to conduct due diligence when asked by the authorities to share information about accounts belonging to activists. "***Meta says they're doing their best not to give information about activist accounts to the government but in a recent meeting they said 'we can't be too sure***,'" the activist said.

Online gender-based violence and misogyny

The overall environment on social media is hostile towards women, with social media users regularly engaging in misogynistic comments, leading many women to resort to self-censorship across social media. Four individuals interviewed for the

briefing, including the Communist party politician Sara Abaza, said that social media companies and Meta, in particular, have been failing women in Jordan for the longest time.

Abaza recalled receiving abusive and misogynistic messages and comments from Facebook users in the past telling her to ***"burn"*** or ***"get back to the kitchen"***. All four interviewees agreed that while content with online misogyny and gender based violence is still circulating online, the volume of such content has decreased with the passing of the controversial cybersecurity law. ***"It's the only positive, an unforeseen consequence of the law, because people are now scared to make such comments,"*** said Abaza. It is noteworthy that, in the absence of platform safety measures, it took a draconian, illiberal law to stymie the deluge of harmful content targeting women.

Despite such a hostile environment, women have been leading pro-Palestine protests, which swept through Jordan post-October 2023, and female activists were allegedly disproportionately targeted by the authorities.

A social media campaign that appears to have been kicked off by male influencers sought to stop their sisters and mothers from taking to the streets, calling on women to stay at home. Likely backed by Jordan's security apparatus, the campaign was meant to create the impression that that was the dominant view, according to a human rights activist who asked that we withhold their name due to possible reprisals. ***"That's because the Palestinian cause unites people in Jordan and the authorities don't want this."***

Conclusions

Meta, Google/YouTube and X were not transparent about their plans to safeguard people and the September 2024 elections in Jordan at a time of heightened tensions in the country and the region. It is therefore near to impossible to gauge whether any of the tech platforms allocated appropriate resources.

Indeed tech harms that were already prevalent, escalated in the months leading up to the vote. In a context of increased government repression of civic space and dissent, the evidence suggests that Facebook and Instagram have not safeguarded their digital platforms as safe and free spaces for the legitimate expression of journalists, women, political dissidents and human rights defenders. The data and testimonies gathered for the briefing also suggest that Meta may be censoring some content at the request of the Jordanian authorities.

OCTOBER

Tunisia

Tunisia, the country known as the birthplace of the democratic revolutions (also known as the Arab spring) over a decade ago, held presidential elections on 6 October 2024. Unfortunately, the months leading up to the ballot confirmed what many have been warning for years: the North-African country is becoming ever more repressive and president Kais Saïed, in power since 2019, has emerged as a dictator. The election was neither free nor fair.

Receiving over 90% of the vote, Saïed won by a landslide, following a takeover of the election committee, decimation of the judiciary and a vicious crackdown on opposition, civil society groups and the media. The disillusionment of the electorate was evidenced by a low voter turnout. Barely 29% of the country's eligible 9.7 million voters cast their ballots, with the youth vote accounting only for 6%.

The incumbent ran almost unopposed. Tunisia's electoral commission approved only three candidates for the race, including Saïed and rejected the administrative court's order to reinstate three other presidential candidates. In the end, Saïed had to face only two other candidates; former parliamentarian Zouhair Maghzaoui and businessman Ayachi Zammel. Though the latter was arrested weeks before the vote. Others who submitted their candidacies for the presidential bid have faced judicial harassment and in some cases had been convicted on charges of "falsified ballot endorsements."

Days before the vote, the Parliament passed a new law stripping the Administrative Court of its jurisdiction in electoral matters, essentially removing checks and balances on electoral irregularities. As of October 2024, over 170 individuals were detained on ***"political grounds"*** or ***"for exercising their fundamental rights"***, according to Human Rights Watch. Among them were over 110 people with ties to the Ennahda opposition party. The electoral commission also denied accreditation to the media and election observers.

Parallel to the real-world crackdown, in the months leading up to the vote Saïed and his supporters had been targeting civil society groups and opposition figures on social media platforms, and Facebook in particular. The president's false and inflammatory remarks about rights groups helping migrants have set off an online smear campaign targeting all civil society organisations, resulting in what the UN described as a "general climate of hate speech".

Social media landscape

Meta's platforms, and Facebook in particular, are the go-to places for Tunisians for accessing information and news. As of April 2024, Facebook had almost 9 million active users, accounting for over 71% of the country's 12 million population (with more than half of the users being male) Some estimates put the video sharing app TikTok right behind, with 5.3 million users. Instagram and YouTube on the other hand have a tiny user base of a little bit over 4% of the population per platform, according to some data.

Facebook is also among the main channels of communication of the Kais Saïed regime with its electorate and the general public. The Facebook pages associated with the president and his office have some 3,6 million followers, in a country with over 9 million registered voters.

Consequently, the regime and its supporters have weaponized Facebook to silence Saïed's critics. And so, whenever the president publicly identifies

targets, his backers attack them online, according to journalist Amine Snoussi, who argued the attacks involve at least some level of coordination with state authorities. (More below).

Shrinking civic space

The Kais Saïed regime has been persecuting its critics – offline and online – since early 2021, when the first wave of anti-government protests sparked by the growing discontent among citizens towards the ruling political elite, further exacerbated by lack of employment opportunities economic mismanagement and the mishandling of the COVID-19 pandemic. By the summer of 2021, Saïed had shuttered the parliament with tanks, suspended the constitution and dissolved the Supreme Judicial Council. He later also took control of the country's electoral commission, further consolidating power.

The authorities have ramped up their crackdown on civil society again in 2023. That's when rights groups had spoken up against Kais Saïed's xenophobic comments – including on social media – in which he accused ***"hordes"*** of migrants from sub-Saharan African countries of bringing violence, alleging a ***"criminal plot"*** to change the country's demographic make-up".

Saïed's comment, which spread on Facebook like wildfire, upped racial tensions and led to real-world violence in July 2023, which saw one Tunisian man killed in an altercation between Tunisians and migrant workers.

The regime had weaponised local laws, including Kais Saied-issued Decree 54, which criminalises misinformation and provides for up to a five-year prison sentence and a fine of 50,000 Tunisian dinars (around 14, 800 euros). In the leadup to the 6 October 2024 ballot, around 60 activists, dissidents, opposition politicians and regular citizens – including a high-school teacher – were arrested and/or detained under the law, a Tunisian activist monitoring the situation on the ground told Digital Action in an anonymous interview.

"Decree 54 has been really terrifying. Everyone here is trying not to go to prison," they said. ***"It has become clear that everything you say in public and private spaces, that is critical, not only of the president, but of officials, even someone criticising the judge could get you in trouble. That has created higher rates of self-censorship."***

While dozens of individuals were persecuted under Decree 54, the actual number of those arrested, persecuted and imprisoned by the authorities is much higher, as they also made arrests under counter-terrorism laws. Between 12 and 13 September 2024, at least 97 members of opposition group Ennahda were arrested and charged with conspiracy and other charges under the counter-terrorism law.

The climate of fear has also permeated the media, with some outlets getting in the authorities' crosshairs because of statements made by guests on air. While independent media has faced pressure to reveal their sources, according to Tunisian activists who spoke to Digital Action.

In an effort to shield themselves from arbitrary arrest and imprisonment, activists have been enabling factory resets on their smartphones, erasing all data, settings, and applications that were previously stored on the device. But even that is illegal under the Decree and can result in further legal sanctions.

Meta undermining election integrity and safety of activists

Despite public warnings from UN agencies and rights groups, including the Global Coalition for Tech Justice, Amnesty International and Human Rights Watch, about the country-wide crackdown

on opposition, civil society and media, Meta never communicated about its plans to safeguard information integrity across its platforms. It also failed to prioritise the safety of activists.

Weeks before the ballot, Siwar Gmati of civil society group I Watch, which is part of Meta's trusted partner programme – an initiative that Meta says taps into the expertise of local groups to ***"address problematic content trends and prevent harm"***-- has flagged incendiary Facebook posts targeting right groups to the tech giant. In response, Gmati said, Meta told them there's nothing it can do because posts accusing rights groups and their employees of being "traitors" working against the interest of the Tunisian state are just an opinion.

"Facebook is in a way responsible for democratic backsliding and it's not doing anything to stop it or to protect human rights defenders." Gmati told Digital Action."***They [Meta] claim that they're defending freedom of expression. But sorry this is not just an opinion if someone says that you're a traitor, that you're not a patriot. What is going to happen offline to an activist who is labelled as a traitor?. We see double standards in treatment. When it comes to countries and consequences – i.e. they take down pro-Palestine content or against Ukraine or pro-Russia content***.

The social media posts and the hateful narrative in question were initiated by the president. On 7 May 2024, in his speech shared on Facebook, Kais Saïed referred to rights groups helping migrants as "***traitors***" "[foreign] ***agents***" and "***rabid trumpets driven by foreign wages"***. The video was played almost 290 thousand times and was shared 1.7 thousand times, with some users explicitly repeating Saïed's inciting language.

While the initial attack targeted organisations supporting migrants in Tunisia, including a group working with UNHCR, soon the online and offline persecution extended to pro-democracy groups and LGBTQI+ organisations. The president has set the narrative and his supporters, including influencers and state media have picked it up, becoming organic spreaders of disinformation.

Since then, content targeting members of civil society groups standing up to president Saïed and referring to them as "traitors" have become ubiquitous on Facebook. Digital Action reviewed eight screenshots of Facebook posts and comments shared by and targeting one civil society group, which requested anonymity fearing regime reprisals. The posts accused the organisation of serving foreign interests and being on the payroll of foreign agents, among other things.

Because they were made by the president, Saïed's inflammatory comments fall under Meta's political speech exception. While the deluge of Facebook media posts by regular users that mirrored the incumbent's hateful rhetoric appear to have been treated by the tech platform like an opinion – which likely doesn't breach Meta's misinformation

policies, nor is it subject to fact-checking or action like content take-down.

Meta failed to consider the local context of the dangers an activist labelled a "***traitor***" faces in a hate-filled environment. The president's rhetoric and the dehumanising language that flooded Facebook has put Tunisian activists at risk of real harm, according to I Watch's Gmati and two other activists interviewed on conditions of anonymity. Meta says it takes factors like individuals' safety into account when implementing its policies, but it appears not to have done this in the case of these inflammatory posts.

Since Saied shared the incendiary post in May 2024, dozens of employees, leaders of civil society groups had been harassed, interrogated or arrested, while organisations' funds had been frozen, sometimes without any notice or possibility to appeal the decision. I Watch, which is the country's leading anti-corruption group, was arbitrarily denied accreditation by the regime-controlled Election Committee, which said its decision stemmed from the fact that the organisation received "***suspicious foreign funding ... from countries with which Tunisia does not have diplomatic relations***".

"***Although NGOs and activists are being arrested, at some point people will be hurt physically because this is what you get from a dehumanisation campaign of activists online,***" said Siwar Gmati of I Watch, which has been taken by the authorities in the crosshairs and whose funds had been frozen in September 2024.

Days before the ballot, on 26 September 2024, Saïed doubled down. His Facebook page shared another consequential post – one that some activists say fully unleashed the inflammatory narrative of civil society groups "***pretending to support democracy***", and accusing them of "***betrayal***" and "***interference in Tunisia's internal affairs***". It wasn't long before the falsehoods were picked up again by his supporters, some of whom turned to Facebook live videos to accuse rights groups of conspiring against state security with entities based in foreign countries.

As pointed out by Gmati, Meta has been employing double standards when applying its content policies, allowing for freedom of speech endangering the safety of Tunisian civil society members, while censoring content supporting Palestine and opposing the genocide in Gaza. Rights groups, including Global Coalition for Tech Justice and Human Rights Watch, have raised concerns over Meta's inconsistent and biassed implementation of its content policies to Pro-Palestine speech, which have resulted in removal of posts expressing support of the Palestinian cause or content documenting human rights violations violations and war crimes against Palestinians that has news value.

"It's a matter of time before this turns into real world violence," said one Tunisian interviewed on conditions of anonymity. ***"This narrative isn't only dangerous in terms of them [NGOS] getting charges but these campaigns are successful at convincing the public opinion and removing the public support from civil society. Even my mom believed it."***

Facebook pages filled with mis- and dis-information

Between July and mid-September 2024, researchers at Mena Media Monitoring looked into posts across a number of Facebook pages and groups, discovering that misinformation has made its way into the content published, shared and often boosted online. The posts were meant to confuse, influence or manipulate public opinion through the dissemination of false information or incitement to hatred or violence among the Facebook users.

The pages and groups have tens of thousands of followers, post daily and have increasingly been generating more interactions in the leadup to the ballot. While some of the pages were managed from Tunisia, most were operating from abroad, further obfuscating who's behind them. They include:

- Tunis Today, with 69 thousand subscribers.
- Politiket, with 103 thousand subscribers.
- I voted Kais Said and I regret it, with 800 thousand subscribers.
- (Kais Said fans) with 373 thousand subscribers.

While attacking opposition members and presidential candidates, many of the posts had targeted Kais Saied and those seen as puppets of the regime, like the president of the election committee Farouk Bouasker. Posts pertaining to Mr Bouasaker appear to instil fear, confuse, incite violence and portray him as a "war criminal".

One post from August 2024 read: "As part of their master's fight against artificial intelligence, Farouk Bouasker proposes, in a correspondence sent on

August 15, 2024, the president of the Republic to cut off social networks including Facebook, TikTok and this until October 6 with the support of the Ministry of Technology."

Another August 2024 post alleged that according to "backstage info" Bouasakar and all election committee members are about to go to prison, while Saied is in the process of securing a safe exit from Tunisia, taking him to Razi Psychiatric Hospital and then smuggling him abroad.

Both posts are blatantly false, yet they remain on Facebook at the time of writing.

TikTok

Activists have also reported a widespread use of TikTok videos by Kais Saied supporters either to spread falsehoods about opposition figures or to incite hatred and violence against them and vulnerable groups, like migrant workers from Sub-Saharan countries. In some videos Saied's supporters were calling opposition politicians who defended migrant workers "rats" and "pests", according to one activist interviewed by Digital Action on conditions of anonymity.

While TikTok acted on reports of hate speech targeting migrants, they did nothing to address slanderous and inflammatory videos targeting the opposition, the activist added. TikTok was also used by influencers with ties to the regime to praise Saied. Such political influence campaigns could be a way to circumvent TikTok's ban on political advertisements on its platform. Though more research is warranted to determine whether this may be true.

Overall, both Facebook and TikTok appear to have failed to implement its policies and act on content that incited violence or put civil society members at risk of harm offline. This is likely caused by the fact that neither company has invested enough in content moderators who speak Arabic and understand the Tunisian context, and its automatic review systems continue struggling with language other than English.

OCTOBER

Brazil

Brazil's 2024 municipal elections, which took place on 6 October 2024, represented a watershed moment in the intersection of democratic processes, artificial intelligence (AI) governance, and the role of Big Tech in shaping public debate. With 441,350 candidates running for mayor and city council, these local elections took place amid new regulations aimed at governing AI use in elections, despite the failure of Brazil's Congress to pass its "fake news" bill (PL 2630/2020), following intense lobbying by tech companies, led by Meta and Google. The Superior Electoral Court (TSE) has stepped forward as one of the key architect of Brazil's digital guardrails, issuing groundbreaking resolutions that tackle two critical fronts: a stringent 24-hour takedown rule for electoral content that spreads demonstrably false or heavily distorted claims about Brazil's voting system and electoral process, and an outright ban on deepfakes – synthetic audio or video content designed to manipulate candidates' images or voices, regardless of authorization. These measures come as lawmakers debate the country's landmark AI framework law (PL 2338/2023), positioning Brazil at the forefront of regulating artificial intelligence in democratic processes.

While many expected a surge of AI-driven manipulation, the elections saw only a few notable incidents. The AI Observatory in Elections reported a handful of issues, like AI-generated jingles in low-budget campaigns and occasional deepfakes. However, the elections did highlight some serious concerns, particularly around gender-based violence, with deepnudes targeting women candidates in local races. Content verification challenges also became clear as audio deepfakes spread on messaging apps. Even major platforms like Google's Gemini and Meta's AI systems had trouble providing reliable election information.

This electoral cycle unfolded against the backdrop of unprecedented tension between Big Tech and Brazilian authorities, most notably in the month-long ban of X (formerly Twitter) from August to October 2024. X's systematic noncompliance with court orders about anti-democratic content, profile removals, and restricting access to certain content, combined with the company pulling out its legal representatives in Brazil, led to a major showdown with the Supreme Court and Judge Alexandre de Moraes. The stand-off with X, which culminated in the platform's compliance only after facing fines and asset freezes, laid bare the broader challenge of making global tech platforms bow to national regulations–a tension that intensifies during electoral periods when the stakes for democracy run particularly high. The clash also tied into ongoing investigations into threats to democracy and digital militias, while raising important questions about Brazil's platform liability laws and sparked a debate on the need to update the Civil Rights Framework for the Internet (Marco Civil da Internet). As a result, Brazil is becoming a key case study on how to balance technology and electoral integrity, showing that regulating Big Tech is about more than just AI–it's also about protecting democracy and holding corporations accountable in the digital age.

Another key development is the increasing involvement of the Superior Electoral Court and the Supreme Court (STF) in the discussion about regulating digital platforms. The Supreme Court has carved out an increasingly decisive role in Brazil's digital landscape, particularly after it began reviewing landmark cases on content moderation and platform liability at the end of 2024. Its active stance could reshape the country's internet laws, including Article 19 of the Civil Rights Framework, potentially setting new standards for how platforms must answer for the content they host. These deliberations, happening as Brazil grapples with electoral disinformation, signal

a deeper shift in how the country approaches platform governance. These changes could set new standards for platform responsibility, even during election periods.

Social Media Landscape

Brazil has the biggest population and number of online social media users in Latin America. It's the fifth-largest social media market worldwide. Meta's platforms are, by far, Brazilians' go-to for information and news, particularly WhatsApp and Facebook, but Instagram follows close behind. In May 2024, Facebook had 175 million active users in Brazil, accounting for 79.2% of its population and 55.3% being women.

However, despite Brazil's massive numbers of social media users, there remains a major connectivity gap in the country. According to the report "Significant Connectivity" on the quality of Internet access in Brazil, this scenario reflects the exclusion of important segments of the population, especially the most vulnerable communities. The majority of users in the country access the Internet exclusively via mobile phone, with limited data, impacting, among other things, their access to information. The proportion of those who fact check their information is much higher among those who use computers and mobile phones simultaneously (71%) than among those who use mobile phones exclusively (37%).

That being said, closed messaging platforms, particularly Telegram and WhatsApp, pose particular challenges to fighting disinformation and online harms. More than 96% of the population is active on WhatsApp. Given that social media platforms and messaging apps are widely used for election campaigning and public debate in Brazil, it comes as no surprise that electoral disinformation remains one of the biggest challenges in protecting democratic integrity.

Regulatory Framework: AI and Platform Governance

Brazil's journey into digital regulation has been groundbreaking, with major steps taken since 2014 to address issues such as platform governance, data protection, and the emerging frontier of Artificial Intelligence.

In 2014, Brazil introduced the Civil Rights framework for the Internet (Marco Civil da Internet), one of the world's first crowdsourced social media platform liability rules. This law requires platforms to comply with court-ordered takedowns of posts and profiles, while implementing a streamlined notice system for removing intimate content shared without consent as well as copyright violations. The framework was further strengthened in 2018 with the Personal Data Protection General Act, which set clear rules about how personal data should be handled by both individuals and companies.

The subsequent democratic crisis prompted new legislative action. This crisis involved the widespread disinformation campaigns during the 2018 election that circulated extensively across social media platforms and messaging apps, particularly WhatsApp, and culminated in the January 8, 2023 riots when thousands of Bolsonaro supporters stormed Brazil's Congress, Supreme Court, and presidential palace after rejecting the electoral results. In April 2020, Brazilian lawmakers introduced Draft Bill no. 2.630, dubbed the "Fake News Draft Bill," representing the country's most ambitious attempt to modernize platform liability rules. The proposed legislation would introduce a duty of care requirement and mandate more active moderation of harmful content, including anti-democratic messages, child abuse material, and terrorist content. It would also impose new transparency requirements on social media companies, forcing them to disclose advertising revenue and publish regular transparency reports.

Despite initial support from civil society groups like Coalizão Direitos na Rede, the bill stalled in late 2023 after facing fierce opposition from tech companies and far-right groups.

Brazil remains at the forefront of AI regulation with Bill 2338/2023. This human rights-based approach to AI governance acknowledges its potential risks, including discrimination, environmental harm, surveillance, and job displacement. It aims to foster innovation while maintaining crucial safeguards for human rights, setting a potential model for international cooperation. As a key step forward, the Senate approved the bill's substitute report and sent it to the Chamber of Deputies for further review. If changes are made in the Chamber and approved, it goes back to the Senate for final approval.

Brazil's presidency of the G20 in 2024 placed AI and information integrity at the center of international discussions. At home, debates continue over possible updates to the Marco Civil, especially with a planned constitutional review of Article 19 and other cases related to content moderation and platform liability. With presidential elections approaching in 2026, digital policy is set to take on even greater importance, particularly as lawmakers prepare to review the Electoral Code before the end of 2025.

Deepfakes during the municipal elections

Brazil's municipal elections this year marked an important milestone in AI regulation. For the first time, the country held elections governed by an electoral directive that prohibited the dissemination of deepfakes. The Superior Electoral Court (TSE) signed memorandums of understanding with Meta, TikTok, LinkedIn, Kwai, X, Google, and Telegram, securing the companies' commitment to work with Brazilian authorities to remove disinformation from their platforms. Research by NetLab and DRFLab highlighted the need to improve AI labeling, making it systematically collectable for research purposes. Their findings also showed that self-declaration of AI use on social media platforms is insufficient, and algorithmic review techniques are ineffective.

Spotting electoral deepfakes on social media requires constant monitoring. The platforms studied –X, Facebook, Instagram, and YouTube– don't offer the possibility to collect data from labels that indicate that the content was AI generated. One solution proposed by NetLab and DRFLab is to create a direct line to fact-checkers, such as Agência Lupa's public WhatsApp account, where people can send content circulating on social media for verification. Improving the internal search systems on platforms with tools like TrueMedia would also help identify AI-generated content more effectively.

While civil society is closely watching the use of AI for spreading disinformation, AI hasn't been used on a massive scale – yet. That said, it still had a serious impact, particularly on women politicians who were targeted by deepfakes. During this year's municipal elections, AI was used in several ways to spread false information, including creating deepfakes, generating political jingles, and making deepnudes of women candidates. This highlights the need for stronger regulations to prevent widespread damage.

Political Gender-Based Violence

During the mayoral elections, women were the primary targets of online violence. In São Paulo, candidates Tabata Amaral and Marina Helena faced an intensification of online attacks. On both YouTube and X, they received three times as many attacks as their male counterparts. Over 80% of posts containing gender-based violence, studied by Democracy Reporting International (DRI) and Fundação Getúlio Vargas (FGV), aimed to undermine women's roles in politics. FGV's research also found that left-wing women candidates were

targeted more frequently than their right-wing counterparts.

"The cost of producing a deepfake during elections against a woman is zero in this country. We presented all the evidence, documented, and the most we managed to achieve was that some videos were removed after a few days. These videos were already in people's phones and in their WhatsApps," said Tabata Amaral. ***"It's impunity that explains all of this. And do I think that the social media platforms are responsible for this process? One hundred percent!"***

YouTube hosted misogynistic and transphobic content targeting women candidates in both urban and rural areas of Brazil. A comprehensive study by MonitorA on online attacks across the country showed that most attacks against women candidates sought to portray them as inferior, while spreading misogyny and undermining their intelligence.

Several candidates, including Loreny Caetano and Suéllen Rosim, were victims of deepnudes generated using AI. A report by the AI Observatory in Elections highlights how deepnudes were used to reinforce the gender-based violence already faced by women in politics. As Yasmin Curzi, a professor at FGV Rio Law, warns, "Women won't feel comfortable participating in politics unless protective measures are put in place. This causes generational harm."

Curzi's research also points to a broader issue with content moderation: women in politics, including journalists and activists, often see their content blocked or deleted due to mass reporting. These coordinated campaigns aim to derail discussions and silence women politicians. An important first step, she says, is to implement effective measures to combat political gender-based violence in agreements with Big Tech platforms. These efforts should involve collaboration with local fact-checkers and civil society organizations.

AI is bound to play an even larger role in Brazil's 2026 presidential elections, and the use of deep nudes against women politicians must be addressed–not only by the TSE but also through stronger platform regulations, particularly more robust content moderation. Given that women–especially those who challenge the status quo, such as trans, racialized, and left-wing women–are disproportionately targeted by tech-driven harms during elections, platform policies must adopt a feminist approach. Failing to do so risks creating a future where generations of women politicians are blocked, discouraged, and harassed from participating in political spaces.

Big Tech Preparedness and Transparency

Big Tech giants are struggling to match their public promises with effective action to protect Brazil's election integrity, according to recent investigations by leading research institutions. Despite high-profile commitments and agreements with electoral authorities, companies like Meta, Google, and TikTok have shown significant gaps in their ability to combat disinformation and regulate AI-generated content.

"***The question of transparency is central***," said Bia Barbosa from Reporters Without Borders (RSF). "it's actually one of the main points that we in Brazil always include in regulatory attempts that exist – access to information for researchers and civil society." Barbosa highlighted how limited access to platform data has hampered effective oversight. These restrictions have grown more severe over time, with researchers noting that conducting analysis "***was more complicated in technical terms than two years ago because of changes in access to social networks***." This challenge became particularly evident in the platforms' response to transparency requirements during the election.

"What happened was that before the electoral process began, some platforms said '***well, then we won't allow [political content] so we don't have to create an ad library***,'" explained Carla Vreche from Conectas, describing how Google and others chose to restrict political content before the elections, rather than comply with transparency measures mandated by Brazil's electoral court.

Studies by NetLab UFRJ (Federal University of Rio de Janeiro), DFRLab, Aláfia Lab, and Data Privacy Brasil also uncovered widespread inconsistencies in the platforms' enforcement of their own policies. A stark example emerged when Google's AI system Gemini continued providing information about political candidates despite the company's announced restrictions on political content.

The municipal elections also revealed new challenges with paid digital influencers. The case of Pablo Marçal in São Paulo highlighted how platform monetization features can be exploited for political gain. Through his extensive digital following, Marçal organized viral content competitions where followers could win cash prizes, which could be seen as abuses of political power and media misuse by Brazilian electoral legislation. As Carla Vreche observed, "***He exposed the problem with the platform business model — hate speech and disinformation that affects our electoral system integrity are being monetized, with people producing this content getting paid by platforms because it goes viral***." This case showed how influencers-turned-politicians can exploit platform algorithms and reward systems in ways that traditional media figures cannot, since established TV personalities must observe a quarantine period before running for office. The campaign also revealed how third parties could bypass official campaign channels, as demonstrated when Marçal's wife make-up artist promoted campaign content, highlighting the difficulty of enforcing transparency rules in digital spaces.

On February 21, 2025, Brazil's Electoral Justice ruled Pablo Marçal ineligible to hold public office for eight years. Judge Antonio Maria Patiño Zorz of São Paulo's 1st Electoral Zone found Marçal guilty of abuse of political and economic power, improper use of media, and illicit fundraising during his 2024 São Paulo mayoral campaign. The ruling came after investigations revealed Marçal had sold his political support to city council candidates in exchange for R$5,000 (around $800 US dollars) campaign donations via Pix transfers, a practice documented in Instagram videos and promoted through registration forms. The court determined Marçal had spread misinformation about the electoral fundraising system and conducted negative campaigning against opponents.

The technical infrastructure for monitoring election integrity is fundamentally flawed, noted a recent analysis from NetLab UFRJ and DFRLab. The research pointed to critical weaknesses in platform APIs and monitoring tools, particularly in tracking AI-generated content and political advertising. Meta's much-touted Ad Library, while offering some transparency, lacks crucial functionality to identify AI-generated content. The situation has worsened with the shutdown of CrowdTangle, which researchers had previously relied on for in-depth analysis.

"***What concerns us is that they analyze content individually, but don't assess what can happen in terms of silencing journalists who receive 300 posts with this type of comment in one day***," said Bia Barbosa, highlighting how platforms fail to consider the cumulative impact of coordinated harassment campaigns. This systemic weakness particularly affected women journalists, who faced disproportionate attacks during the election period.

The challenge of identifying AI-generated content has become particularly acute in Brazil's political

landscape. A recent controversy in Fortaleza highlighted these difficulties when experts failed to conclusively determine whether a disputed audio message was artificially generated. In another revealing case, Salvador's candidate, Bruno Reis, posted AI-generated videos across multiple platforms but only labeled them as AI-created on Instagram, exposing the inconsistent approach to content transparency.

Agreements between tech companies and Brazil's Superior Electoral Court (TSE) have revealed structural weaknesses in oversight mechanisms. While these memoranda of understanding established procedures for handling complaints through the Integrated Center for Combating Disinformation (CIEDDE), they notably lacked requirements for proactive detection tools. NetLab UFRJ and DFRLab researchers concluded that Google's blanket ban on political content not only failed to achieve its intended goals, but actually reduced transparency by hampering systematic monitoring. For Bia Barbosa, the memorandum of understanding between the platforms and the TSE during the elections "***works very little, though it's better than nothing***".

The different approaches taken by the platforms have created an uneven playing field. Meta opted for a more permissive stance with enhanced transparency requirements, while Google and TikTok maintained strict bans on political advertising. This fragmentation in policy approaches, while each in line with TSE regulations, created potential gaps in the overall election information ecosystem that bad actors could exploit.

Despite significant investments in integrity measures, Meta claims to have deployed 40,000 people globally for security and integrity since 2016, but implementation has fallen short. TikTok's dedicated election monitoring systems and other platforms' safety measures have struggled to keep pace with evolving challenges, particularly in detecting and verifying AI-generated content.

The research findings come at a critical time as Brazil grapples with the intersection of artificial intelligence and electoral integrity. With deepfakes and synthetic media becoming increasingly sophisticated, the gap between Big Tech's technological capabilities and the challenges of maintaining election integrity continues to widen.

"***Even with this smaller number [of attacks compared to national elections], it's a serious situation that generates silencing from the press about external attacks that happen in campaign situations***," Barbosa noted, emphasizing how even reduced levels of harassment can significantly impact election coverage. The findings underscored the growing disconnect between Big Tech's stated commitments and their ability to protect democratic discourse in an evolving digital landscape.

Lessons for Brazil's 2026 presidential elections

2024 was a big year for tech-related debates in Brazil. Amid ongoing discussions in Parliament and the Electoral Court on updating the regulatory framework for platform governance and Artificial Intelligence, the government was busy with its pro tempore presidency of the G20 summit and introducing information integrity as one of the core themes in the forum's agenda. While the summit saw calls for more transparency from Big Tech, discussions on regulating AI, harnessing its potential for societal benefits, and combating climate disinformation globally, the domestic landscape still faces significant challenges.

InternetLab research coordinator Iná Jost points to the nuanced reality of electoral resolutions in Brazil's digital regulation landscape. "***Electoral***

resolutions end up innovating a lot in terms of electoral propaganda — they are both a charm and a liability," she noted, explaining how these rapid regulatory responses, while nimble enough to address the fast-changing tech ecosystem, raise concerns about democratic accountability since they're drafted within cabinet offices without broader public input. She highlights this year's AI regulations as an example of this dynamic, where the Electoral Court successfully implemented rules on deepfakes and AI labeling requirements while also demonstrating how the effectiveness of such resolutions heavily depends on the technical expertise of the officials crafting them.

Despite efforts by social media companies and electoral authorities to further regulate digital spaces during this year's elections, Brazil saw continued tech-related harms. Deepfakes and online attacks targeting women journalists and candidates were prevalent. As the country enters 2025, the question remains: What updates will be made to the Electoral Code – set to be reviewed by Parliament this year – and will the AI Act be approved? There are also potential amendments to the Civil Rights Framework for the Internet that could be proposed by the Supreme Court, addressing platform governance issues.

The failure to pass "fake news" legislation (PL 2630/2020) and X's months-long failure to comply with their obligations highlighted the ongoing tensions between tech platforms and Brazilian authorities. Carla Vreche pointed out that the "***judiciary is mostly acting because Congress failed to fulfill its role due to political disputes. If we had platform regulation, if Bill 2630 had passed, most of these details that need to be addressed in TSE resolutions or Supreme Court discussions would already be resolved.***" Without more robust legislation, platforms have relied on self-regulation and voluntary commitments, which have not been enough to protect the digital civic space. Furthermore, a scattered approach to content moderation across different platforms – from Meta's transparency requirements to Google and TikTok's inconsistent outright bans – created an uneven playing field that could affect future electoral cycles. This rule gap has forced judicial authorities to step in, showing the need for stronger enforcement tools and clearer obligations for platforms operating in the country.

Looking ahead to the 2026 presidential elections, civil society organizations are emphasizing the need for stronger institutional frameworks. "***There's a significant concern because the process of conducting hearings and gathering suggestions from civil society for electoral resolutions is not mandatory***," warnded Carla Vreche, alerting on how changes in electoral court leadership could affect civic participation in rule-making. On the other hand, the upcoming parliamentary review of the Electoral Code and the pending AI Act present opportunities to strengthen the country's digital democracy framework, particularly in addressing emerging challenges like AI-generated content and technology-facilitated gender-based violence (TFGBV).

The municipal elections also exposed gaps in Brazil's ability to track and verify digital threats to electoral integrity and press freedom. As researchers from RSF, the Internet and Data Science Laboratory (Labic) at the Federal University of Espírito Santo and the Institute for Technology and Society of Rio (ITS-Rio) as parto of the Coalition in Defense of Journalism (CDJor) found in their analysis on the municipal elections in 2024 to oversee freedom of the press situation:

"***When we use hashtags to find content against the press or when we use combinations of terms through artificial intelligence, we can conclude there's content attacking the press, but platforms analyze content individually, not systemically.***"

These continuing restrictions on content analysis and platform monitoring have made oversight increasingly difficult. Research institutions, civil society organisations and other stakeholders looking to gather information in electoral processes are losing direct access to platform data. ITS-Rio experience is telling, noted Barbosa – they "***used to do data collection themselves, but now with the changes that happened on X, they can no longer do it – forcing organizations to rely on a shrinking pool of institutions with platform access credentials***."

Looking forward, Brazil needs a balanced approach that combines stronger regulation with innovative solutions. Civil society groups are pushing for digital attacks to be recognized as serious threats to press freedom and democratic discourse, while calling for more robust enforcement mechanisms. As Barbosa said, **"When we analyze attacks on journalists during elections, we're not just looking at individual rights violations, but at attacks on society's right to information in the electoral context."** Similarly, political gender-based violence on social media targeting women politicians causes widespread, generational harm, with a chilling effect on women's political participation.

As Brazil prepares for the 2026 presidential elections, the lessons learned from last year's municipal elections will be crucial to shaping the country's democracy. While the government and civil society discuss and build the intersections of AI, platform regulation, and electoral integrity, digital rights advocates worldwide are keeping a watchful eye on what will unfold in Brazil. The path forward requires not just regulatory updates, but a comprehensive approach that balances innovation with protection, transparency with privacy, and technological advancement with democratic values.

NOVEMBER

United States

On November 5th 2024, more than 153 million ballots were cast to determine the United State's president, one of the highest voter turnouts since women were given the right to vote. Trump's campaign was fueled by anti-immigrant and white nationalist disinformation, in both English and non-English languages. The November 2024 ballot demonstrated the uniquely harmful ways in which offline and online disinformation campaigns target language-minority voters and the ways in which foreign information ecosystems impact U.S. domestic politics. It also exposed how Big Tech companies, which have failed to invest sufficiently in content moderation for the languages of the Global Majority, were also absconding their responsibility towards non-English speaking voters in the US.

Language differences in policy enforcement around election disinformation impacts millions of citizens. These Big Tech failures resulted in Muslim, Latinx and Black communities in the U.S. being affected by non-English and foreign disinformation campaigns.

Big Tech U.S. election policies 2024

While the deadly insurrection of the U.S. Capitol on January 6, 2021 brought to light how failure to moderate content on social media can undermine democracy, Meta, Google, TikTok and X all took a step back on their election integrity plans, as part of a growing trend to de-prioritize political content. This didn't begin in 2024. Big Tech had rolled back electoral integrity policies since the 2020 midterm elections despite promises to safeguard elections. 17 critical policies across Meta, X and YouTube were eliminated between 2022-2023, including Meta's political ad policy that mandated transparency and labeling. This happened alongside mass layoffs, particularly in the trust and safety, ethical engineering or responsible innovation, and content-moderation teams.

Musk's X cut half of its election integrity team just a week prior to the November 2024 election, after having pledged to expand it. It also reduced its resources to enforce its rules, as well as narrowing the categories of speech that violate them. For example, X previously banned posts containing claims like "***unverified information about election rigging***" or "***claiming victory before election results have been certified***." On top of that, the tech giant has reduced the penalties for breaking its rules. As part of its civic integrity policy, it encouraged users to add context to each other's posts using the crowdsourcing "fact-checking" feature, Community Notes. This feature mostly fails to combat misinformation, according to experts. X doesn't partner with independent fact-checkers and while its policy indicates that posts that violate its civic integrity rules will be labeled as misleading instead of removed, it gives no indication of whether those accounts will face consequences or not.

Since the 2020 U.S. election, Meta shifted its elections integrity policies by scaling back on fact-checking labels and has done less to promote voting process information on Facebook's information center. In response to the criticism it faced that its platform recommended people into "Stop the Steal" groups, Meta announced that Facebook would no longer recommend political or civic groups for users to join. Meta also cut down on election integrity employees, and while Facebook now requires labels on AI-generated content, there are still many unlabeled generative AI images being circulated. It also has more circumscribed policies regarding election misinformation, along with YouTube and X, weakening platform prohibitions against misinformation that delegitimizes

election violence. None of these platforms have committed to addressing election fraud rumors comprehensively.

TikTok played a bigger role in the presidential election than in any previous cycle. It's the only of the four platforms mentioned above that prohibits political advertising. It prohibits claims that Trump won in 2020 and uses independent fact-checking partners to review unverified election claims. NYU's Stern Center for Business and Human Rights found that "***TikTok has announced tough-sounding policies related to elections, but the platform's haphazard enforcement has failed to slow the spread of deniers' lies***."

Global Witness tested TikTok, YouTube and Facebook's robustness by submitting advertisements containing false election claims and threats a few weeks prior to this year's election. TikTok and YouTube approved 50% of the ads, although YouTube blocked the publication of the ads until the submission of formal identification. Facebook approved one ad containing harmful disinformation, which is an improvement from a similar test done during the 2022 midterm elections.

Spanish-language Social Media Use and Electoral Disinformation

Spanish-speakers in the United States are avid social media users. Two-thirds of Latinx people treat YouTube as a primary source for their news and information about politics and elections. Facebook is also widely used by Spanish speakers in the U.S., with over 28 million users, 69% of which use Facebook daily. Among Spanish-speaking Facebook users, bilingual Latinxs and Spanish-dominant Latinxs are more receptive to their platforms' video ads and are more likely to share advertising content.

Immigrant communities in the U.S. who speak languages other than English at home use social media in ways that make them more exposed to mis- and disinformation campaigns. WhatsApp and other encrypted messaging apps are often key places for political discussions. Latinx people in the U.S. are more than twice as likely to use messaging apps such as Telegram and WhatsApp than other groups. Encrypted messaging apps are more difficult to scrutinize and fact-check, making them particularly susceptible to disinformation campaigns.

As Roberta Braga, director of Digital Democracy Institute of the Americas (DDIA), states "***There's nothing inherent to Latino communities that makes us less accurate in our ability to identify false content online***." Rather, the ways in which Latinx communities use social media, coupled with the often racist and white nationalists motives behind disinformation campaigns, makes racially targeted disinformation in Spanish a pressing issue. Latinx communities are not more likely to believe disinformation than other groups, but they are more likely to be impacted by the violence and hate that such disinformation incites.

Immigration disinformation and its consequences

Public understanding of the relationship between online speech and offline harms has developed significantly in recent years. Research published in 2023 showed how anti-Muslim hate crimes spiked after Donald Trump posted anti-Muslim discourse on social media during the 2016 presidential primaries. This pattern w as in evidence again during the 2024 elections.

Most disinformation around immigration during this year's election revolved around portraying immigrants as responsible for a rise in violent crime and that immigrants are causing a rise in unemployment for people born in the U.S. Anti-immigrant disinformation was rampant in both English and Spanish, such as widespread rumors

that crime rates skyrocketed in New York due to increased immigration.

After September's presidential debate, in which Trump alluded to disinformation that members of the Venezuelan gang ***Tren de Aragua*** were taking over a Colorado apartment complex, the complex's residents have said ***"they feel unsafe [...] and they fear being stereotyped as criminals."*** Similarly, Haitian immigrants have received threats and Springfield has received more than 33 bomb threats since Trump's statements at the debate.

While immigrants, both documented and undocumented, are less likely to commit crimes than native-born U.S. citizens across various crime categories, the fact that a big part of Trump's campaign was built off of these tropes, and that so many people voted for him, shows how disinformation shaped voting behaviour.

Trump and his allies have heavily used anti-immigrant rhetoric about undocumented people registering to vote to continue the "Big Lie" narrative in an attempt to guarantee support for denying election results if Trump lost. As polls started showing Trump's lead, claims of a 'stolen election' on social media began to disappear. Donald Trump, JD Vance and Elon Musk were top spreaders of anti-immigrant disinformation, and although we'll never know to what extent Big Tech's failure to moderate harmful content contributed to Trump's re-election, it certainly played a part. This is partly due to the fact that platforms adopted a less aggressive stance on election integrity.

During the 2024 presidential elections, polling suggested that false claims affected how people saw candidates and their views on immigration, crime and the economy. Big Tech rolled back their policies meant to curb disinformation during elections, in a general attempt to depoliticize themselves and avoid similar scrutiny received in past elections. Policies around fact-checking political speech were absent, and while some argued that the risk of harm no longer outweighed the benefits of political dialogue, election disinformation caused fear, confusion, harassment against immigrant voters and even bomb threats.

"***What this moment has especially brought to my forefront***," says Sanaa, Director of the Digital Spaces Project at Muslim Counterpublics Lab, ***"is that I don't trust Big Tech to do this work***." The question that we need to be asking ourselves is: "***what kind of power can we develop and leverage as users to democratise power, so that it doesn't lie solely in the hands of Big Tech?***"

Meta's independent content watchdog, the Meta Oversight Board, said there were "***serious questions***" as to how the company deals with anti-immigrant content on Facebook. This, coupled with multiple above-mentioned studies of how Meta, Google, TikTok and X did not adequately enforce their election integrity policies (and even less so in non-English languages), is particularly troubling for minority rights and freedoms.

Anti-immigrant rhetoric is less about turning people against immigration and more about amplifying already held beliefs rooted in racism, according to Germán Cadenas, an associate professor at Rutgers University who specializes in the psychology of immigration. The biggest demographic to vote for Trump was white men, and white US adults are more susceptible to core stereotypes of Latinx immigrants being a threat to American society, according to a recent Rutgers University study.

Language minority voters are particularly vulnerable to disinformation campaigns that in turn have real life impacts, such as Springfield's bomb threats after the spread of the ***Tren de Aragua*** disinformation. Disinformation has also impacted

immigrant's access to factual information. Research on immigration misinformation during this year's election shows that mis- and disinformation has left many immigrants confused and fearful about using government benefit programs. Also, four in ten immigrants say Trump's rhetoric about immigrants has negatively affected them, including about half of Asian immigrants.

Big Tech failed non-English voters

Given that the Census Bureau projects white Americans to be a minority by 2050, the country's language, speech and culture will (and is currently) also shifting. Because these shifts are fueling nativist and racist disinformation and electoral strategies, domestic media and information ecosystems must also shift their priorities and meaningfully invest in non-English languages and cultures.

Research by the Center for Democracy and Technology (CDT) shows how content moderation practices on social media are not nearly as robust for non-English languages. The study highlights how language-minority voters sit at the intersection of culturally tailored misinformation and data voids. Big Tech companies such as Meta and Google have recently revealed plans to use automated content analysis tools to combat non-English disinformation. These tools have limited abilities to detect intent and motivation and perform especially poorly in languages that don't have much digital representation (such as Indigenous languages, different creole languages, etc.).

Prior to this year's election, 24 million voters were expected to rely upon language translations of voting materials to cast their vote. These languages cover Native American and Alaskan Native languages, certain Asian languages, and Spanish. Arabic isn't included as one of the language minorities protected in the Voting Rights Act, making Arabic-speaking voters particularly vulnerable to non-English electoral disinformation. Rima Meroueh, director of the National Network for Arab American Communities, said that "***community members are more concerned about whether they might face voter intimidation or be turned away at the polls, rather than not feeling confident in the system.***" Over 14,000 letters were sent to registered voters who are naturalized citizens threatening criminal prosecution for illegal voting. Widespread electoral disinformation about undocumented people registering to vote fueled voter intimidation this election cycle, in turn causing real-world harassment and fear among immigrants.

The Latinx community is also impacted by Big Tech's failure to moderate hateful content, especially in Spanish. University of Washington's Center for an Informed Public released research on how TikTok, Instagram and Facebook were not enforcing some of their own policies to safeguard against election misinformation in Spanish. In-platform searches revealed discrepancies in implementing election information policies. For example, searching for "***fraude electoral***" (electoral fraud in Spanish) on Instagram doesn't produce the voting information banner as it does in English.

Foreign influence campaigns affected the U.S. elections

While tracking the origins of disinformation can be challenging, there are clear patterns of non-English and foreign campaigns promoting white nationalist and racist agendas. For instance, the US Department of Justice found a Russian government-led disinformation network meant to influence elections which targeted "hispanic descendents," among other groups. The Russian state media firm, RT, is accused of using AI and bots to spread disinformation on immigration and crime prior to the U.S. elections this year. One notorious example was a fake video of a "Haitian man" (who wasn't actually Haitian) claiming to have voted in two counties after just arriving in the

U.S. This video, created in Russia, spread widely across social media and helped shape Trump's campaign narrative.

Onyx Impact, a nonprofit founded to better serve and empower Black communities by fighting the harmful information ecosystems targeting them, released a study prior to the elections called "The Black Online Disinformation Landscape" that found foreign actors to be one of six core networks spreading online disinformation impacting Black voters and Black social issues in the U.S. These include "***individuals or entities acting on the behalf of, have strong ties to, or may be inadvertently promoting talking points from foreign governments, organisations, or interests that seek to influence or interfere in US political discourse***." Concerns over foreign influence operations escalated further after reports that Meta allowed users around the world to buy and sell Facebook accounts that are authorised to run political ads in the U.S.

Big Tech alignment with Donald Trump

Over the past years, dozens of congressional hearings took place involving tech companies on issues around election integrity, online harms against children, privacy and content moderation. The Biden government had been investigating Meta, Alphabet, Tesla, Apple and Amazon, among others. In the run up to 2024, tech companies had also faced increasing political pressure from Republicans who portrayed content moderation as anti-conservative censorship. Alphabet, Amazon, Apple, Meta and Microsoft were subpoenaed by the House Judiciary Select Subcommittee on the Weaponization of the Federal Government in 2023 for information about their companies' communications with the executive branch over how their content is moderated.

Elon Musk, owner of X, endorsed Trump during the election campaign, spending more than $250 million on his re-election, and successfully shaped X into an echo chamber for Trump's supporters. By loosening content moderation policies on his platform, Musk allowed right-wing disinformation to flourish on X. Musk, who has 206 million followers on X posted false and misleading claims about the 2024 elections that were viewed two billion times. His tweet insinuating that Democrats were allowing undocumented people to cross the border in order to vote received 47 million views alone.

In response to civil society groups' research and advocacy against X's business model, Musk "*declared war*" and sued such groups multiple times. In particular, he sued the Center for Countering Digital Hate (CCDH), which recently looked into how X's crowdsourced fact-checking feature, Community Notes, failed to counter false claims about the U.S. elections.

Besides Musk, tech moguls in general bet on Trump to loosen government control over Big Tech regulation, in the hopes of avoiding accountability. Following the election, Meta's CEO, Mark Zuckerberg, met with President-elect Trump at Mar-a-Lago. He criticized the Biden administration's "***pressure to censor Covid-19 content***" during the pandemic and possibly sees potential in Trump's laissez-faire approach. Amazon's Bezos also announced he was "very optimistic this time around," after having refused to let his newspaper, The Washington Post, endorse Kamala Harris. While Trump acolyte and nominee for Federal Communications Commission President, Brendan Carr, declared a wish to smash Big Tech's "***censorship cartel.'"***

This has set the stage for the unabashed alignment between Big Tech CEOs and President Trump to dismantle content moderation efforts and campaign against the digital regulatory efforts of other countries in 2025.

Clicks, Code, and Consequences:
Big Tech's Gamble with Human Lives and Election integrity in the 2024 Year of Democracy

Global Coalition for Tech Justice