Digital Democracy Initiative

FUTURE-PROOFING ELECTIONS AGAINST DEEPFAKE DISINFORMATION

2025

CIVICUS

ACKNOWLEDGMENTS

Author
Phumzile Van Damme

Design and layout
Monica Molano

We would like to acknowledge and thank all the civil society actors, researchers, and practitioners who generously shared their time and insights through interviews and expert consultations for this research

Digital Democracy Initiative

The Digital Democracy Initiative (DDI) is a programme to safeguard inclusive democracy and human rights in the digital age. It focuses on supporting local civil society in the Global South, particularly in countries undergoing democratic regression and where civic space is under pressure.

For more information visit:
digitaldemocracyinitiative.net

CIVICUS

CIVICUS is a global alliance of civil society organisations and activists working to strengthen citizen action and civil society throughout the world.

For more information visit:
civicus.org

CONTENTS

EXECUTIVE SUMMARY

This paper examines how Generative AI (GenAI) is reshaping election-related disinformation and assesses how civil society can future-proof democratic processes against the next wave of manipulative technologies. While deepfakes have dominated global headlines as an existential threat to democracy, the evidence from the 2024–2025 "super election cycle," during which nearly half of the world's population voted, reveals a more complex reality. The much-feared "deepfake election" did not materialise; however, the convergence of cheapfakes, synthetic media, and algorithmic amplification continues to erode public trust in elections, journalism, and institutions.

Deepfakes - AI-generated video, audio, or images that mimic real people with a high degree of believability - represent a distinct medium of deception within the broader information disorder ecosystem. They differ from "cheapfakes" that rely on low-tech manipulation such as splicing or mislabelling. Deepfakes exploit the realism heuristic, namely the human tendency to place greater trust in visual over text-based information. Studies show that they reinforce pre-existing cognitive biases, increase uncertainty, and trigger lasting memory distortions, even after being debunked. Across the Global South, the most common threats remain cheapfakes, narrative manipulation, and coordinated inauthentic behaviour, which exploit a low level of digital literacy, linguistic inequities in moderating content, and polarised information ecosystems.

This report, produced by the Digital Democracy Initiative at CIVICUS, introduces a Deepfake Risk Matrix - a conceptual framework for assessing national vulnerabilities across five domains: political-institutional, social, economic, digital ecosystem and actor behaviour.

Using case studies from Namibia, Ecuador, Singapore and Germany, we find that the impact of deepfakes is less determined by their technical sophistication than by contextual factors such as media freedom, trust in electoral institutions, and the strength of civil society, as illustrated by the following points:

- **In Namibia**, cheapfakes targeting President Netumbo Nandi-Ndaitwah employed gendered disinformation to question her fitness for office, showing how manipulated media intersects with entrenched social biases.
- **In Ecuador**, AI-generated content formed part of a broader hybrid information war, amplified by violence, corruption, and bot-driven polarisation.
- **In Singapore**, deepfake regulation is among the world's most advanced, yet a high degree of state control limits open debate and can blur the line between regulating disinformation and censorship.
- **In Germany**, a robust regulatory and fact-checking ecosystem under the EU Digital Services Act and AI Act mitigated synthetic media risks, demonstrating the value of pre-emptive legal and civic action.

Among these cases, civil society organisations (CSOs) emerged as a crucial line of defence. In 2024 and 2025, CSOs globally expanded their monitoring, partnering with fact-checkers and launching deepfake literacy campaigns. Examples include BOOM Live's Deepfake Tracker in India, Mafindo's synthetic media monitoring in Indonesia, and Witness's pre-bunking initiatives in the United States. These initiatives collectively strengthened public resilience and pressured platforms to enhance transparency. However, disparities remain, especially among CSOs in the Global South which face limited resources, restricted data access, and potential political threats.

The report argues that regulatory readiness is uneven across the globe. Existing frameworks focus on reactive takedowns or content labelling rather than structural transparency or accountability. Many countries in the Global South lack specific legislation on synthetic media, while those that do often risk

criminalising freedom of expression. A human-rights-based approach to deepfake governance must, therefore, mitigate potential harm, while at the same time protecting free expression, creativity, and civic participation.

A key insight is that the deepfake problem is not technological in nature but systemic. It reflects long-standing weaknesses in media ecosystems, the commodification of public attention, and the political economy of content-generating platforms. Mitigating harm requires embedding human rights and accountability principles at every stage of the technology lifecycle - from data collection and model training - to deployment and platform moderation.

The advocacy strategy and recommendations (Part 5 of this report) call for a coordinated global response centred on:

- Mandatory algorithmic and data transparency from tech platforms.
- Human impact assessments and independent ethics boards for AI developers.
- Balanced, rights-based regulation of synthetic media that avoids overreach.
- Investment in digital literacy and prebunking campaigns, especially for women, youth, and marginalised communities who are often targeted with disinformation.
- Stronger South–South cooperation and funding for CSO-led monitoring and response networks.

Ultimately, the report concludes that deepfake disinformation should be viewed as a symptom of deeper structural inequalities in the global information space.

Technological defences alone will not suffice. The durability of democracy will depend on investing in informed, resilient, and empowered audiences and media consumers. Preparation - rooted in evidence, ethics, and equity - remains a powerful defence of democratic principles and human rights values.

INTRODUCTION

The launch of widely accessible, user-friendly Generative AI (GenAI) platforms has made deepfake production easier, faster, cheaper and more accessible than ever before. This has ushered in an era of disinformation - so believable that it could further undermine the foundations of civic life. Nowhere is this threat more acute than in the democratic arena, where deepfakes pose a direct challenge to elections, public trust, and the legitimacy of democratic institutions.

Platforms such as ChatGPT, Grok, Synthesia, Vall-E and thousands of other applications have significantly lowered the technical barriers that once restricted deepfake creation to those with advanced skills and resources. As the capabilities of GenAI continue to advance, the line between fact and fiction - already blurred in an era of contested truth - will become more obscured.

Indeed, the introduction of OpenAI's Sora in October 2025 illustrates how AI-generated content is becoming increasingly difficult to detect as fake. Earlier, GenAI applications, though hyper-realistic, often contained tell-tale signs that betrayed their inauthentic nature. Sora's deepfakes, in contrast, are "extremely real," as noted by The New York Times (2025).

What this level of believability of deepfakes could mean for elections slated in 2026 and beyond remains to be seen.

To better understand this new reality of GenAI and its impact on elections, we examine the lead-up to the 2024 "super election year," as the biggest election year in human history, when half of the world's population - 3.7 billion people - across 72 countries went to the polls.

The global electoral cycle in 2024 was described by Aspen Digital as "the first AI election" (Schiller & Harbath, 2025) due to its taking place in tandem with what was predicted to be a proliferation of "supercharged" AI-generated content (Freedom House, 2023).

Governments, policymakers, academics, and experts nearly unanimously agree that a massive wave of false information is approaching (Alanazi et al., 2025). The World Economic Forum declared disinformation and misinformation the primary immediate global risk for 2024, based on interviews with nearly 1,500 global leaders from academia, business, government, the international community and civil society.

Anticipating a surge in synthetic content, major platforms introduced policies, detection systems, and labelling initiatives to demonstrate readiness.

As the 2024 election year drew to a close, it was clear that the anticipated surge of deepfake disinformation had not materialised as anticipated. Evidence indicated that GenAI was predominantly used for entertainment, satire, and efficiency improvements rather than as a primary tool for widespread voter manipulation. Traditional mediums of disinformation, including societal elites and the mainstream media, had a greater impact (Simon & Altay, 2025). This was certainly the case in some countries in the Global South, where cheapfakes, not deepfakes, still dominated the information landscape.

This lesser impact, however, does not imply that the threat no longer exists. Rather, it emphasises the need to increase efforts to proactively prepare for and meet challenges posed by emerging technological threats to democracy. It also emphasises the need to continue addressing the root causes of information disorders, instead of shifting the focus to new mediums.

To that end, this paper pursues three aims. First, it traces the evolution of deepfake technology and its role in reshaping the global disinformation landscape.Second, it explores existing empirical research on the effects of political deepfake disinformation to build a solid evidentiary foundation for civil society to assess the real-world impacts of deepfakes.

The goal is to develop aknowledge and evidencebase that captures how deepfakes are influencing elections worldwide, with a particular focus on the Global South.

Third, it analyses the evidence-based scale of deepfake electoral disinformation between 2024 and 2025, up to the writing of this report. Four country case studies - Namibia, Ecuador, Singapore and Germany - are examined as each represents a distinct social, political, and economic digital ecosystem with unique challenges from which we can extract lessons that can strengthen future democratic resilience.

In addition, we introduce a comparative matrix framework for analysing deepfake disinformation. The matrix organises the case studies around key contextual and impact variables, enabling systematic comparison across digital ecosystems and socio-cultural and political contexts, particularly in the Global South. The framework enables a country-specific analysis within the matrix, allowing researchers to pinpoint issues that require targeted intervention. Its purpose is proactive: to anticipate and address emerging risks. Used reactively, it can also identify precisely where and how problems manifested.
Lastly, we present an advocacy strategy with a set of recommendations for civil society partners. To avoid the self-defeating narrative that exaggerates AI's threat to elections, we also highlight AI's potential to strengthen civil society and civic space worldwide with a list of use cases.

NOTE: For shorthand, "disinformation" is used in this paper to refer to both disinformation and its by product: misinformation.

PART 1: DEEPFAKES & ELECTORAL INTEGRITY

1.1 Deepfakes Explained

Deepfakes – also known as "synthetic media" – are videos, audio or images created or altered using machine learning algorithms. Machine learning, a subfield of AI, trains algorithms to identify patterns in data and apply that pattern recognition to make predictions or produce new outputs. As a subset of machine learning, Generative AI (GenAI) generates hyperrealistic images, audio or video. To produce text, GenAI models rely on Large Language Models (LLMs) to mimic human language and reasoning.

Textual representations of disinformation, even when AI-generated, are not considered deepfakes. That is because the phrase "deepfake" has as its origin in a combination of "deep learning" and "fake" and thus is primarily concerned with the use of deep learning technology to create hyperrealistic depictions of individuals saying and doing things that did not happen, i.e., visual realism and impersonation (Singh & Dhumane, 2025).

The phrase "deepfake" is often used as a catch-all term for all synthetic media, but there is a distinction. Falsified videos and images produced with basic editing tools such as Photoshop or Adobe software, or simple techniques such as cropping or splicing, are known as "cheapfakes" and have long been a part of the arsenal used by malicious actors.

Deepfakes, in their AI-generated nature and amplified by the rise of GenAI platforms, represent a new frontier of risk that requires specific focus.

The focus on deepfakes does not intend to diminish the dangers and risks posed by cheapfakes. Deceptive content undermines information integrity regardless of the technology used to create it.

1.2 The History of Deepfake Tech

The origins of deepfake tech can be traced to the 1990s, when researchers began experimenting with computer-generated imagery (CGI) to create lifelike representations of humans (Regan, 2024). Progress accelerated through the 2010s, driven by the emergence of larger datasets, improvements in deep learning, and access to greater computing power.

The introduction of Ian Goodfellow's Generative Adversarial Networks (GANs) in 2014 marked a landmark moment in the use of deep learning to automate the generation and refinement of deepfakes. This breakthrough laid the foundation for the modern GenAI system. GANs provided the architecture to enable much more realistic synthetic content. Gone were the days of the painstaking manual effort that CGI required. However, despite these advancements, deepfake production still required substantial technical expertise, coding abilities, and specialised hardware, keeping it largely within the capabilities of researchers and niche online communities.

The ultimate game-changer was the introduction of user-friendly, easily accessible GenAI platforms such as ChatGPT, Gemini, Synthesia, and now thousands of other applications that have made the production of deepfakes easier, cheaper, and faster.

1.3 Electoral Disinformation and Electoral Integrity

Within an electoral context, disinformation is false or misleading information intentionally created and disseminated to manipulate public opinion, suppress voter turnout, discredit opponents, and/or distort democratic processes and outcomes. It becomes misinformation when those unaware of its intent to mislead share it, thereby amplifying its reach.

Election integrity, according to the International Federation of Electoral Systems (IFES) - a nonpartisan, nonprofit organisation advancing democracy and free and fair elections worldwide - requires "comprehensive, fair, and

practicable legal frameworks; transparent and professional election administration; sound electoral operations and accurate election results; accessible and fair election dispute resolution mechanisms; inclusive and wide-ranging participation of voters and candidates; and professional and impartial media."

Online electoral integrity requires information integrity: open information ecosystems where voters can access accurate, trustworthy information without being misled by disinformation, misinformation, hate speech or foreign malign influence operations.
The proliferation of electoral disinformation across the globe does not bear repeating. It continues to have far-reaching consequences that have eroded information and electoral integrity. For civil society, autocratic actors continue to deploy internet shutdowns, online censorship, and cybersecurity laws to restrict civic space around the world.

The harm caused by electoral disinformation is most evident in its different tactics. These are:

Type of Disinformation	**Main Tactic**	**Targets**	**Key Effects**
Voter Suppression Disinformation	Spreading false information about voting dates, eligibility or safety at polling stations	Excluded groups, opposition voters	Intentional voter confusion, demobilisation, and undermining of electoral integrity and trust in democratic institutions
Election Denialism Disinformation	Falsely asserts elections are fraudulent or illegitimate	General public, election officials, and perceived political opponents	Erodes public confidence, delegitimises outcomes, fuels conspiracy thinking, encourages political radicalisation, incites threats or violence
Identity-Focused Disinformation	Exploits demographic, ethnic or religious differences through customised electoral disinformation	Targeted communities	Inflames prejudice, social fragmentation, and polarisation; provokes violence and harassment, and reduces political participation
Gendered Disinformation	Uses false or sexualised narratives to discredit, intimidate, and silence	Women, gender-nonconforming people	Reputational harm, psychological trauma deters women's political participation, weakened democracy
Violent Extremism Disinformation	Deliberate spread of emotionally charged, incitement to violence electoral disinformation	Public, opposition, dissenters	Disrupts elections, sparks division, legitimises violence, erodes trust, radicalises sentiment, suppresses dissent, paired with digital repression

It is within this volatile climate of manipulation, violence, and falsehoods that, for the first time, the tools to fabricate hyperrealistic deepfakes are now available to the public at virtually no cost.

OpenAI reported that, as of October 2025, ChatGPT had **800 million** active users weekly, an unprecedented scale of access, with nearly 1 billion people using ChatGPT alone. That is an 8th of the global population. In early 2025, Gemini and Anthropic reported 300 million and 18.9 million active users, respectively.

It is essential to understand not only the "how" of disinformation but the "why". At the core, disinformation campaigns are psychological warfare that weaponises existing social tensions and divisions to sow further division and deepen polarisation. They do so by targeting and exploiting human cognitive biases. In this way, false information achieves believability not through truth, but through psychological manipulation.

This negative effect is particularly potent in the context of electoral disinformation because politics cannot always be separated from emotion. It is through emotion that political attitudes and public opinion are formed, steering people towards certain types of content and influencing how it is interpreted through their inherent, and often deeply personal, cognitive biases. It is for this reason that disinformation is more prevalent, influential, and persistent on politically charged topics than on neutral or non-divisive ones (Zhou & Shen, 2024).

This is a principle that malicious actors are aware of and understand, namely that effective disinformation hinges on emotional resonance, particularly through base emotions such as fear, anger, outrage, loyalty and hope. When messaging taps into these emotions, it becomes more compelling, more viral, and more likely to override analytical scrutiny. As explained by Claire Wardle and Hossein Derakhshan in their foundational paper on the study of information disorders (2017):

> *The most 'successful' of problematic content is that which plays on people's emotions, encouraging feelings of superiority, anger or fear. That's because these factors drive re-sharing among people who want to connect with their online communities and 'tribes'. When most social platforms are engineered for people to publicly "perform" through likes, comments, or shares, it's easy to understand why emotional content travels so quickly and widely, even as we see an explosion in fact-checking and debunking organisations.*

What have been the impacts of deepfake disinformation since the launch of ChatGPT and similar platforms? This includes assessing how cognitive, emotional, and behavioural processes influence the believability of deepfakes. The need to carefully examine the psychosocial dynamics behind the spread of political disinformation, across all distribution methods, cannot be overstated.

Deepfakes cannot be understood solely through their technology. As an established principle, technological artifacts contain politics. Technology cannot be assessed solely for its contributions to efficiency and productivity, or for its positive and negative environmental side effects, but also for the ways in which it can embody specific forms of power and authority (Winner, 1980).

The impact of deepfakes is deeply contextual and depends on human factors such as adoption rates, the social norms and cultural values being disseminated, and the psychological mechanisms that govern how individuals interpret, accept, and share manipulated content.

It is at the intersection of technology, context, and human cognition where disinformation gains its persuasive power and spreads through societies. We will explore this in the next part of the paper.

As Access Now highlights, deepfakes should be analysed in a variety of contexts and not only within electoral periods:

> *For now, it isn't clear what, if any, additional new risk of GenAI poses in the context of election disinformation that were not already present before generative AI came on the scene. In the meantime, while the world panics over the as-yet-unproven unique impact of generative AI on elections, the very real and distinct societal harms… we would be remiss not to mention the serious systemic risks posed by generative AI models used to create and disseminate non-consensual sexual imagery and child sexual abuse material, for instance, and the heightened online threats to the human rights, safety, and dignity of women, LGBTQ+ communities, and other racialized and marginalized groups.*

PART 2: LITERATURE REVIEW: EMPIRICAL EVIDENCE ON THE IMPACT OF DEEPFAKES

Much of the popular understanding of the risks of deepfake political disinformation and its impact on electoral integrity rests on speculative, alarmist perspectives. This results from the inherent flaws of speculation as a thought process and from assumptions based on broad generalisations that assume uniform technology acceptance, use, and effects across the globe.

Fortunately, the last three years have seen a surge in empirical assessments of the risks posed by deepfakes. Researchers and academics have begun deploying experimental designs to measure how deepfakes influence voter perceptions, trust in democratic processes, and the overall information ecosystem. This includes assessing how cognitive, emotional, and behavioural processes influence the believability and spread of deepfakes.

The need to carefully examine the psychosocial dynamics behind the spread of political disinformation, across all its delivery methods, cannot be overstated. As a relatively new and potentially potent method of content delivery, analysing the impact of deepfakes must be grounded in empirical research to support evidence-based mitigation and advocacy strategies.

As researchers at Purdue University note, much of the existing academic literature on deepfakes has focused on their perceived credibility, examining whether viewers find them convincing, whether they alter beliefs or attitudes, and whether they can be detected in idealised laboratory settings using researcher-generated content. However, with the growing integration of GenAI into politics, there is an urgent need to expand inquiry into the real-world impact of political deepfakes (Walker, Schiff, & Schiff, 2024).

As a global civil society alliance, CIVICUS is leveraging the strength of its diverse membership to help civil society organisations respond to emerging digital challenges. Through this paper, we contribute to the growing body of empirical research on the impact of deepfakes on electoral integrity.

The insights offered here are intended to ground this empirical analysis, enabling civil society partners to test theoretical assumptions against real-world impact rather than speculative, fear-based narratives. By combining an overview of academic research with case studies from four countries and other examples, this paper aims to catalyse a research agenda that strengthens advocacy efforts, culminating in the recommendations outlined in the accompanying Advocacy Brief. Prior to the literature review, we will clarify some conceptual issues to start.

Not all deepfakes are intended to harm or are harmful. Deepfakes can and have been disseminated for legitimate political and social commentary in the form of memes, satire, and parody. Indeed, much of deepfake usage observed during the 2024 electoral cycle served as satire, educational purposes or political commentary, according to researchers (Nathan & Sanders, 2024).

Disinformation, however, is distinguished by its core feature, namely the intent to deceive. Therefore, a deepfake ceases to be simple commentary or satire if it is intended to deceive.

There has also been a tendency to describe the now broad access to GenAI platforms as the “democratisation” of AI. We caution against the use of this phrase. Access to technological tools alone does not make technology democratic. As highlighted by the Centre for the Governance of AI, Harvard University’s Berkman Klein Centre, and the Centre of Governance of AI, the concept of “democratisation” masks deeper questions of power, primarily who owns the infrastructure, who sets the rules, and who benefits? In all instances, the response to these three

questions of power is the Global North (Seger et al., 2023), to quote:

> *AI democratisation is a multifarious and sometimes conflicting concept that should not be conflated with improving AI accessibility. If we want to move beyond ambiguous commitments to "democratising AI" to productive discussions of concrete policies and trade-offs, then we need to recognise the principal role of the democratisation of AI governance in navigating tradeoffs and risks across decisions around use, development, and profits.*

CIVICUS continues to address the Global South–North digital divide as a key challenge affecting civic space and digital democracy. Our focus remains on strengthening digital resilience and access for civil society in the Global South. We advocate for genuine digital inclusion, which means ensuring access to technology, training, and tools that reduce inequality and empower participation. We also champion the use of digital technologies to expand democratic engagement and global solidarity, while protecting vulnerable communities from online repression.

2.1 What are the Psychological Effects of Deepfake Consumption?

Understanding the impact of disinformation requires understanding how the human brain processes information and forms beliefs. Cognitive biases have long been established as central to the spread and believability of disinformation, shaping how individuals interpret and accept information. And the interaction between deepfakes and cognitive biases is an area of emerging study.

HM Murtuza and MD Oliullah (2025) explored research on the impacts of political deepfakes on cognitive processing. While finding a Western dominated approach in many studies, they found evidence indicating that political deepfakes can trigger a wide range of psychological effects, including deception, uncertainty, loss of trust and shifting attitudes, thus reinforcing existing biases. They further include the vital proviso that belief formation is heavily dependent not only on the content itself but also on individual traits, political leanings, and the wider social and political environment. Deepfakes do not affect everyone in the same way. They interact with existing beliefs and predispositions, producing complex and varied outcomes.

In their study, Weikmann, Greber, and Nikolaou (2024) concluded that exposure to deepfakes has far-reaching consequences beyond simple deception, asserting that "people [could start to] no longer believe in what they see." Deepfakes affect both the perceived credibility of information and an individual's confidence in identifying falsehoods. In high-choice media environments, this uncertainty can be especially damaging, as it could amplify scepticism towards journalism and politics. Effects on memory have been equally substantiated. A Massachusetts Institute of Technology study confirmed that AI-generated content can distort memory and perception (Pataranutaporn et al., 2024). Participants exposed to AI-altered images were significantly more likely to report false memories, and this effect intensified with AI-generated videos, where confidence in these fabricated recollections was even stronger. The content not only misinforms in the present but also reshapes how the past is remembered.

Therefore, there appears to be scholarly agreement that deepfakes have deceptive power and can have the following consequences:

- Trigger a wide range of psychological effects, including deception, uncertainty, loss of trust and shifting attitudes, thus reinforcing existing cognitive biases.
- Affect both the perceived credibility of information and an individual's confidence in identifying falsehoods.
- Distort memory and the perception of events, not only in the present, but also in the past. This effect is more pronounced with deepfake videos.

- Its impact, however, is heavily dependent not only on the content itself but also on individual traits, political leanings, and the wider social and political environment.

2.2 Are Deepfakes more Persuasive than other Types of Disinformation?

Opinions vary over how persuasive deepfakes are compared to other types of disinformation. Ching et al. (2024) conducted a scoping review of existing empirical studies that have investigated the effects of viewing deepfakes on people's beliefs, memories, and behaviours. They found evidence suggesting that exposure to deepfakes can influence opinions about public figures, increase the believability of misinformation, and create false memories. However, it remained unclear whether deepfakes are more manipulative than other forms of misinformation.

HM Murtuza and MD Oliullah (2025) noted that while political deepfakes can lower trust in online news and increase distrust of the government, research offers mixed results on whether deepfakes are more credible or persuasive than other forms of misinformation. Sharing deepfakes depends on cognitive ability, confirmation bias, and social dynamics. In other words, deepfake content is not inherently more credible or persuasive than other forms of misinformation.

Sundar, Malino, and Cho (2021) arrived at the same conclusion, theorising that video can make disinformation seem more credible than audio or text, leading to greater possibility of sharing the content. The effect is stronger among users with less interest or knowledge of the topic and who are more likely to believe fake stories when presented in video format. The level of perceived realism increases both credibility and the intention to share the content on WhatsApp, for example.

These differing opinions highlight the need for more extensive and continuous research to fully understand the impact of deepfakes on information integrity and how they are reshaping media consumption habits.

In the research, there is a well-studied concept with broader agreement: the liar's dividend and deepfakes.

The phrase "liar's dividend" was coined by two legal scholars, Danielle Citron and Robert Chesney (2019), to describe a new social construct: how disinformation serves as a convenient and powerful rhetorical device for malicious actors, in this case, politicians. It allows them to plausibly dismiss real evidence as "misinformation" or "fake news." This is seen with deepfakes in information ecosystems, where the concept of truth is contested and it is hard to distinguish fact from fiction. Chesney and Citron explain:

> *Ironically, liars aiming to dodge responsibility for their real words and actions will become more credible as the public becomes more educated about the threats posed by deep fakes. Imagine a situation in which an accusation is supported by genuine video or audio evidence. As the public becomes more aware of the idea that video and audio can be convincingly faked, some will try to escape accountability for their actions by denouncing authentic video and audio as deep fakes. Put simply: a sceptical public will be primed to doubt the authenticity of real audio and video evidence. This scepticism can be invoked just as well against authentic as against adulterated content. Hence, what we call the liar's dividend: this dividend flows, perversely, in proportion to success in educating the public about the dangers of deep fakes. The liar's dividend would run with the grain of larger trends involving truth scepticism.*

2.3 Are Deepfakes more Convincing than Cheapfakes?

Hameleers (2024) conducted two experiments to compare the effects of deepfakes and cheapfakes on the perceived credibility of political disinformation. Using a between-subjects design, Dutch participants were shown manipulated videos of a conservative Dutch politician, falsely portrayed as delivering a radical, anti-immigration speech. The fabricated speech included statements such as "immigrants are responsible for most of our country's problems" and that people from "backward and retarded societies commit violent crimes." These claims were false and inconsistent with both the politician's real views and empirical crime data. The researchers deliberately constructed the video as disinformation.

In this particular experiment, deepfakes were not rated as more believable and credible than cheapfakes. That is a) on the specific topic of immigration, and b) in Dutch society - "on average, deepfakes are rated as less credible and believable than cheapfakes."

One could be tempted to generalise these findings and declare that cheapfakes are more potent than deepfakes. That is not so. The greater point in Hameleer's study is that less sophisticated modes of deception, such as cheapfakes, can be as credible as more sophisticated deepfakes. The power of both forms depends heavily on context.

This, however, does not discount the possibility that as deepfakes become more realistic and lack the tell-tale signs of falsity, cheapfakes will become more believable, even to disinformation-literate audiences.

Does this make deepfakes more powerful? Yes and no. It depends on the topic, the context, and the audience. Disinformation must always be analysed within context. We explore this issue regarding context when answering the next question.

2.4 How are Societal Norms and Cultural Values Influencing the Spread of Deepfakes?

To answer this question, researchers consulted 14 accomplished experts, each selected for their rich background and expertise across various disciplines (Alanazi et al., 2025). The experts highlighted the need to consider how factors such as region, personality, and social media use may affect attitudes towards technology and authenticity, leading to varied perceptions and impacts of deepfakes across the globe.

This is a crucial point. Social context cannot be detached from how disinformation is interpreted and shared.For example, the interplay between local customs, communal expectations, and ingrained values determines how disinformation resonates across populations, influencing not only those who are persuaded by it but also the motivations for sharing or rejecting the content.

To illustrate, African scholars studied the factors influencing the believability and dissemination of misinformation in six Sub-Saharan African countries (Madrid-Morales et al., 2021). The research revealed that a country's political culture and media system may affect how users interact with false information. For example, sharing political information, including misinformation, was an act of courage in Zimbabwe, a country with limited press freedom and ongoing authoritarian rule. In contrast, South Africa, which has an active media sector and a functioning democratic system, exhibits lower levels of motivation among consumers to share political news.

A different dynamic is present in some Asian countries. A report by the European Centre for Populism Studies (Yilmaz et al., 2022) of five Asian countries explains how governments in those countries often use religious values to justify digital authoritarianism, with little public backlash. Censorship, surveillance, and internet shutdowns are often justified as measures to safeguard religious and cultural values, with disinformation frequently cited as a rationale

for restrictions. The appeal to shared religious and cultural values can contribute to less opposition to such restrictions that would otherwise be considered authoritarian.

Sabhanaz Rashid Diya (2024), a computational social scientist and Executive Director of the Tech Global Institute, a policy lab working at the intersection of private technology companies, civil society, and government to reduce equity gaps in the Global South, explains that in the "Global Majority," voters are far more exposed to cheapfakes than to advanced deepfakes.

This distorts civic discourse, discredits candidates, and worsens misinformation in fragile or emerging democracies with low digital literacy and limited press freedom. Their ease of production makes their impact potentially larger in scale. However, cheapfakes have been a blind spot, as major platforms focus their manipulated-media policies mainly on deepfakes, neglecting the far more widespread cheapfakes. Diya suggests the adoption of more technology-agnostic frameworks that focus on harm, not just on the level of sophistication of the manipulation. We agree.

2.5 Key Insights from the Literature Review

Topic	Key Points
Cognitive and psychological effects	Deception, uncertainty, loss of trust, shifts in attitude, reinforce existing biases, distort memory and perception, recall false events with high confidence; effects mediated by individual traits and socio-political context; cultural norms and media freedom influence spread and interpretation
Persuasiveness vs. other misinformation	Deepfakes can influence beliefs; not consistently more persuasive than text/image misinformation; video realism heightens credibility and sharing intent among low-information audiences
Liar's dividend	Key effect of deepfakes is the liar's dividend as wrongdoers increasingly dismiss authentic evidence as fake; public awareness fuels truth; scepticism erodes trust in visual/audio media
Deepfakes vs. cheapfakes	Cheapfakes can be as believable as deepfakes; sophistication matters less than cognitive/contextual factors
Global North-South digital divide	Research is Global North-centric; assumes universal access/effects; CIVICUS positions research within Global South realities; cheapfakes dominate amid infrastructure, language, press-freedom gaps
Cheapfakes in the Global South	Dominant form of manipulated media; produced/shared more easily than deepfakes and equally harmful; distort civic discourse, discredit candidates, deepen misinformation/polarisation; danger lies in effect on trust, participation, and overall information environment

PART 3: PREVALENCE OF DEEPFAKE DISINFORMATION IN 2024/25 ELECTIONS

In this part of the paper, we use reliable data to quantify the prevalence of deepfake disinformation during the 2024 and 2025 global election cycles. Reliable data is needed to move beyond alarmism, anecdotal evidence, and speculation, and to reveal the true frequency, forms, and intensity of deepfake disinformation deployment during elections, as well as the range of methods, motivations, and their associated campaigns. This understanding is necessary to craft evidence-based mitigation strategies.

3.1 GenAI and Deepfake Landscape ahead of the 2024 Super Election Year

Ahead of the 2024 electoral cycle, GenAI platforms and their deepfake technology had been on the market for a little over a year. OpenAI's ChatGPT leads the global text AI market, while DALL-E leads the market for realistic image generation. Google's Gemini competes in both text and image synthesis, while Midjourney gains traction for highly realistic images used in both art and misinformation.

Video tools like Synthesia and DeepBrain AI allow for the creation of AI avatars and virtual anchors for media outlets. Apps such as Reface and DeepSwap popularise face-swaps and deepfakes among general users.

In audio, voice cloning tools have become mainstream. ElevenLabs dominates with hyperrealistic, multilingual voices, joined by open-source projects like TorToiSe and VALL-E, which enable voice replication from minimal data. While these technologies aid accessibility and creative work, they can also lead to impersonation scams and disinformation.

Governments, policymakers, academics and experts almost unanimously agree and are concerned that a massive wave of false information is flooding the information ecosystem.

A small number of outliers believe this to be alarmist. In what has become a canonical paper, Felix M. Simon, Sacha Altay, and Hugo Mercier (2023) argue that the threat of widespread misinformation was "overblown," and list their reasons as follows:

- Quantity of misinformation: GenAI only dramatically lowers the cost and effort required to produce manipulative content; the spread of misinformation has always been limited more by demand than by supply. A larger pool of misinformation does not automatically translate into greater impact if audiences are not receptive.
- Quality of misinformation: Realism is not the sole determinant of persuasiveness. Emotional resonance, narrative fit, and audience predispositions often matter more than surface quality. A slick deepfake might still fall flat if it does not align with what an audience is primed to believe.
- Personalisation of misinformation: While GenAI can lower barriers to producing such content, it does not fundamentally introduce new microtargeting capabilities. The mechanics of tailoring messages to digital identities already existed through online advertising and data analytics.

They were right. It is worth reflecting on their last point about the personalisation of misinformation. There were concerns that GenAI would enable greater microtargeting of disinformation, a perspective that may have been founded on alarmism. GenAI does not introduce new microtargeting methods; it merely lowers the cost and effort of tailoring messages to inferred digital identities, psychographics, and other similar data. Its current abilities are limited in producing synthetic media that can be tailored based on identified characteristics such as demography, income, location, etc(Simchon et al., 2024).

Microtargeting itself is not inherently wrong; political parties have used it as a tactic to reach

potential voters with tailored messaging before and after the advent of the internet. It is when digital identity data is obtained unethically or illegally and used to manipulate voters with personalised, fear-based disinformation, as seen in the Cambridge Analytica scandal, that trouble arises.

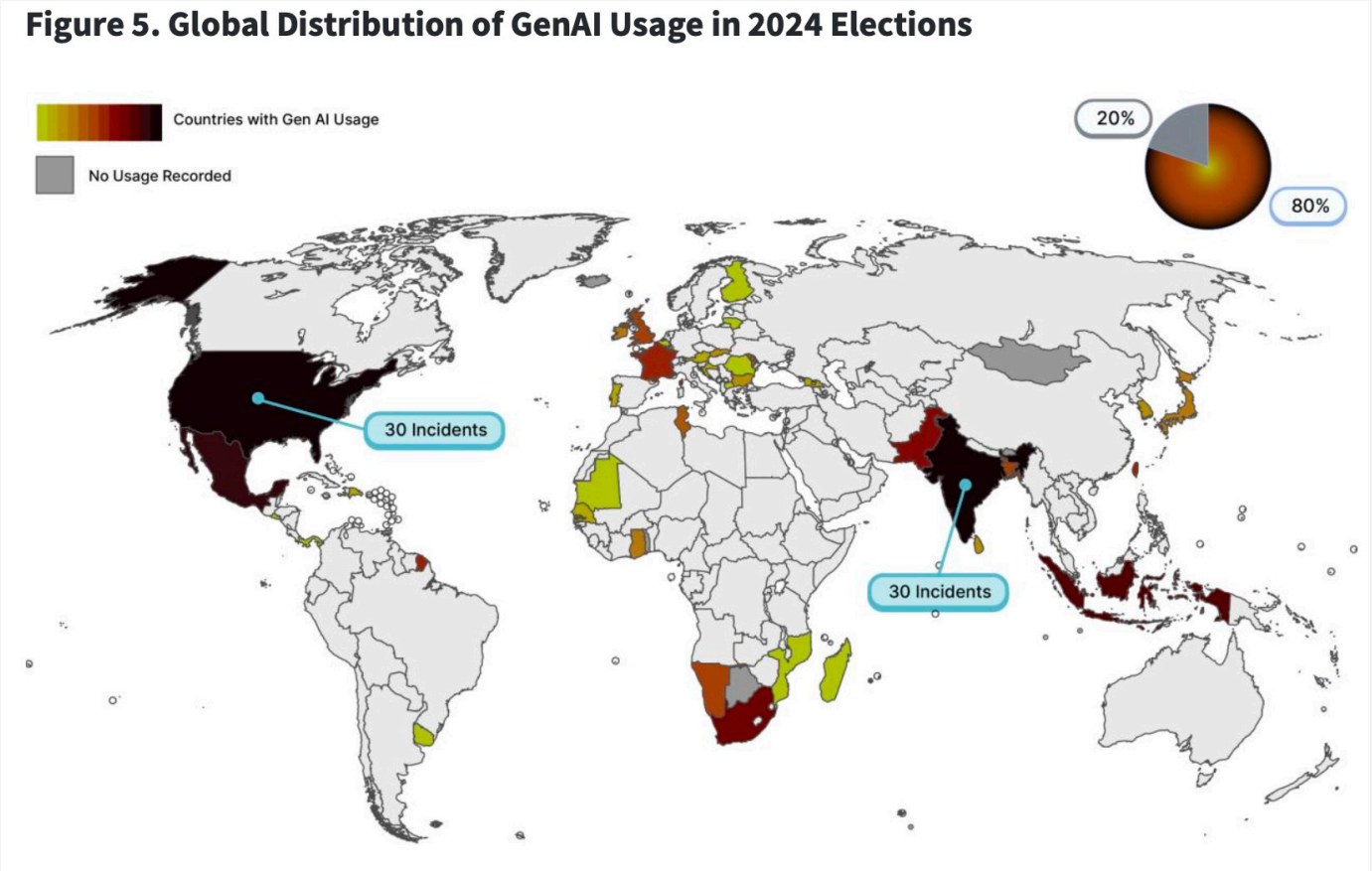
This, of course, is a GenAI ability that will need to be monitored in the coming years, and appropriate mitigation strategies will need to be devised and implemented.

3.2 Deepfake Disinformation during the 2024 Super Election Year

The International Panel on the Information Environment (IPIE) - an independent and global science organisation providing scientific knowledge about the health of the world's information environment - found just 215 instances of AI-generated deepfake electoral disinformation across all 50 countries with competitive national elections in 2024 (IPIE, May 2025).

Separately, the Knight Institute reviewed **78 cases** of AI use in the WIRED AI Elections Project,which tracked political AI content during the 2024 global elections, and found no deceptive intent in **39 instances** (Kapoor & Narayanan, 2024).

Figure 5. Global Distribution of GenAI Usage in 2024 Elections



This was much lower than expected, as confirmed by Meta's reports. Meta closely monitored the potential use of deepfakes by covert influence campaigns and found only incremental productivity and content-generation gains, accounting for **less than 1%** of all fact-checked misinformation on its platforms.

Unfortunately, other social media platforms did not publish reports detailing the full scope of the proliferation of deepfake disinformation. Even Meta's was a little thin on the details. Data access remains an area ripe for civil society advocacy. Platforms must publicly furnish complete datasets on the nature and prevalence of all disinformation, including deepfakes, to facilitate independent

monitoring and informed interventions. This would include granular data on content reach and engagement, amplification networks, audience exposure and reactions, as well as transparency in detection, labelling, and enforcement measures.

Despite what was relatively low deployment of deepfake disinformation across the globe, it is still worth reflecting on the incidents and their nature, as noted by the IPIE:

- **80%** of countries with competitive elections saw GenAI used during campaigns; India and the USA reported the highest number of incidents (**30** each). GenAI incidents were in double digits across eight countries (**16%**), ranging from **10 in France** to **30 in India** and the USA.
- The countries with the fewest instances (some as low as one) were Belgium, El Salvador, Finland, Lithuania, Madagascar, Maldives, Mauritania, Mozambique, North Macedonia, Palau and Panama.
- The remaining **20%** of countries with no identified GenAI usage data are primarily those with smaller populations.
- Explanations for the lack of recorded GenAI usage targeting elections in both groups of countries include less journalistic coverage of these countries' elections and lower internet penetration rates.
- An additional explanation could be restrictions on freedom of expression, which could lead to less user-generated content.
- The original source of GenAI-generated content was unknown in **46%** of cases. Of untraceable cases, **79%** involved suspected political manipulation.
- Foreign actors were identified in **20%** of cases, all linked to malign uses, such as Russian and Chinese coordination.
- Paid commercial actors were involved in **6%** of the cases. In **31%** of these cases, paid commercial actors worked in tandem with partisan groups, political parties and candidates, and/or foreign actors, whereas **69%** involved paid commercial actors acting alone.
- Constructive uses were also noted: **38%** of national party and candidate-related GenAI incidents were beneficial, with **16%** focused on civic outreach or accessibility.

Comprehensive data on the detection of electoral disinformation deepfakes during the 2025 election cycle is unavailable, as news coverage has notably declined. However, according to the CIGI (2025), the 2025 global elections confirmed key findings from the IPIE report, namely that:

> *Ultimately, AI is not a stand-alone disruptor but rather a powerful new layer in existing influence operations, with the potential to outpace rules and regulations if not managed appropriately.*

Altay and Mercier argued in 2025, as they had in 2023, that concerns about GenAI threats were exaggerated and cautioned that alarmism over GenAI could distract from ongoing issues, such as deepfakes targeting women and excluded groups. They caution:

> *By overemphasizing the risks of GenAI in the context of elections, we risk overlooking the broader, more insidious ways in which GenAI is misused, such as enabling targeted harassment and amplifying harmful biases. These include the harassment of women and minorities. The creation and distribution of AI-generated fake nudes, mostly targeted at females, is a form of gendered violence that seeks to silence women in public life and can be used to humiliate, discredit, and threaten women, which may have a chilling effect on their participation in politics. Similarly, minorities are targeted by AI-assisted harassment campaigns, including racially biased or xenophobic attacks that are amplified through social media. These targeted campaigns undermine efforts to build inclusive political spaces.*

3.3 Resilience Factors against Deepfake Disinformation

In addition to the reasons identified by the IPIE, such as low internet penetration rates, limited media coverage, and digital authoritarianism, as well as those highlighted by Simon, Altay, and Mercier, we explore other reasons that could have been resilience factors limiting the spread of deepfake disinformation.
These must not be seen as definitive explanations but rather as a summary of the collective measures implemented before elections that may have played a resilience role. The level of concern and corresponding preparation for disinformation threats in this context was unprecedented. It may signal the need for continued proactive planning.

3.3.1 Public Awareness of Deepfake Disinformation

There appears to have been high level of awareness across the globe of the threat posed by AI and disinformation, as revealed by a 29-country Ipsos survey (Ipsos, 2023). The survey data has limitations that extend beyond the typical constraints of opinion polling and it was conducted in mid-2023, so the conditions and perceptions may have changed since then.

In the survey - Global Views on AI and Disinformation – Ipsos sought to measure how people around the world perceived the risks posed by AI-generated misinformation. It surveyed over 21,000 people in 29 countries, with three-quarters believing that AI makes it easier to produce realistic disinformation and just over half believing it would worsen the problem of disinformation. Despite this, most were confident they could spot fake content, revealing a potential gap between perception and reality. Many also felt that political and media dishonesty have increased over the past thirty years, indicating a broader trend of less reliance on established institutions for information and a growing reliance on sources such as social media influencers for "alternative truths," which can lead to inadvertent consumption of and belief in misinformation.

The chart below contains the survey results for Global South countries, where respondents indicated to what extent they agree or disagree with these two statements: statements

- "Artificial intelligence is making it easier to generate very realistic fake news stories and images."
- "Artificial intelligence will make misinformation and disinformation worse."

Country	**Is GenAI making it easier to create realistic disinformation?** (% of respondents who agree)	**Will GenAI make disinformation worse?** (% of respondents who agree)
South Africa	77%	52%
Indonesia	89%	45%
Peru	82%	46%
Peru	81%	53%
Singapore	80%	46%
Thailand	78%	42%
Colombia	77%	58%
Argentina	77%	45%
Malaysia	76%	51%
Mexico	75%	42%
Brazil	74%	51%
India	66%	52%

It is unlikely that these figures accurately reflect the Global South region as a whole. For example, the survey only covers one African country, despite the continent's vast differences in internet penetration rates. That data, however, is instructive and gives a viewpoint on perceptions of GenAI harm.

3.3.2 Platform Readiness

Another factor that may have limited the scope of AI-generated deepfakes is the preparedness of social media platforms. Anticipating a surge in synthetic content, major platforms introduced policies, detection systems, and labelling initiatives to demonstrate readiness. However, the level of preparedness was uneven; some companies invested in AI detection and partnerships with fact-checkers, while others relied heavily on voluntary codes and vague commitments.

It must be noted that much of the information about these measures comes from the platforms themselves. Other than Meta, which released limited information, no other platform provided details on the success — or lack thereof — of its readiness measures. This raises questions about whether these strategies were designed to safeguard electoral integrity or to pre-empt regulatory scrutiny.

Platform	Preparedness Measures	Limitations
Meta (Facebook, Instagram, WhatsApp)	Updated misinformation policies, partnered with fact-checkers, invested in AI to detect synthetic media, and launched "Made with AI" labels in 2023	Detection capacity remains limited; enforcement is inconsistent; and the system is heavily reliant on self-reporting
TikTok	Introduced new rules banning harmful AI-generated content, added content labels for synthetic media, and partnered with fact-checking organisations in some regions	Limited transparency on enforcement; region-specific coverage is uneven
X (formerly Twitter)	Committed to "community notes" for misleading AI-generated media; voluntary alignment with the EU's Code of Practice on Disinformation	Scaled back trust & safety teams, enforcement is inconsistent, and collaboration with researchers has been reduced
YouTube	Announced it would require disclosure of altered or synthetic content and expanded misinformation policies to cover AI-generated deepfakes	Labels rely on uploader honesty, but their proactive detection capacity is limited.
Google (Search & Ads)	Banned political campaigns from using AI-generated content in ads without disclosure; improved detection for manipulated media	Enforcement is mainly ad-focused, not organic content; the impact on the broader ecosystem is unclear.
Snapchat	Rolled out AI content guidelines; partnered with third-party safety groups to track harmful synthetic media	Smaller scale than rivals; limited reporting on election-related enforcement

3.3.3 Regulatory Readiness

Ahead of the 2024 elections, many governments and policymakers acknowledged the dangers posed by deepfake disinformation and began developing strategies to address it. Efforts included detecting, labelling and, in some cases, restricting AI-generated content through both voluntary industry agreements and new laws. These varied initiatives signal growing acknowledgement that deepfakes pose serious risks to election integrity and democracy.

For comparative purposes, the following list of legislation and policies across different jurisdictions includes those that did not hold elections in 2024/25. The elements that worked and did not work could be extracted from each model applied or under consideration. This is another area where civil society advocacy will be particularly beneficial, namely informing of what successful legislation and policies exist or are under consideration.

Country/ Region	Legislation/Action Taken	Penalties & Enforcement	Effects & Impact
European Union	The EU AI Act (2024) was passed and regulates high-risk AI, including deepfakes and election disinformation.	Fines up to **€35M** or **7%** of global turnover for serious violations	Boosted transparency; inspired global policy models; platforms adapting rapidly
Ukraine	Proposed AI law on counter-disinformation; targets hybrid warfare and foreign interference	Criminal liability, content takedowns, and platform restrictions	Deter foreign manipulation; strengthen national security response
United States	State-level laws to regulate deepfakes in elections and impersonation	Civil/criminal penalties, fines, and imprisonment, depending on intent	Platforms preemptively restrict content, leading to growing pressure for federal action.
China	Deep Synthesis Regulation passed in 2023; tackles deepfakes by requiring all synthetic content to be clearly labelled and traceable. Providers must obtain explicit consent before editing personal attributes, and platforms are obligated to detect, disclose, and remove harmful or misleading AI-generated content.	Fines, platform bans, and criminal charges for harmful or political misuse	There has been an improvement in traceability, but critics warn of censorship risks under broad enforcement.
South Korea	Basic Act on the Development of Artificial Intelligence and the Establishment of a Foundation for Trustworthiness Protection law establishes national oversight and safety infrastructure to prevent harmful uses of AI, including deepfakes.	Expected fines and imprisonment; legislation is still pending implementation.	Raises public awareness; platforms preparing for stricter compliance

India	Draft AI Accountability and Ethical Use Bill; targets AI-generated disinformation by recommending mandatory labelling, licensing for content creators, and prosecution mechanisms for deepfake misuse.	Penalties are under development and are expected to include moderation mandates.	Sparked debate on free speech vs. disinformation control; draft law still evolving
Singapore	The Protection from Online Falsehoods and Manipulation Act (POFMA) requires correction notices or takedowns for AI-generated disinformation, including deepfakes, and empowers authorities to act swiftly against misleading content.	Up to **SGD 100,000** for individuals and SGD 1 million for platforms or entities that fail to comply with directives	Strong deterrence; emphasis on public education and platform responsibility
Japan	Act on Promotion of Research and Development and Utilisation of Artificial Intelligence-Related Technologies (March 2025)	Encourages safe AI use and allows government oversight of AI misuse, including disinformation, though it does not impose specific penalties	Focus on innovation balance; critics call for stronger enforcement mechanisms.
Australia	Combatting Misinformation and Disinformation Bill 2024; empowers the Australian Communications and Media Authority (ACMA) to regulate harmful AI-generated content, including deepfakes, but was withdrawn due to concerns over free speech and lack of political support.	None currently	High public concern; legal uncertainty; platforms urged to self-regulate.

3.3.4 Civil Society Readiness

Elections are integral components of democracy, enabling voters to participate in shaping their country's future. 2024 was a year of many opportunities for democracy, with **74 national elections worldwide**. Civil society organisations (CSOs) played an indispensable role in many of these, both in countries where elections were free and fair and in those that were more problematic. In Ghana, for instance, BudgIT Ghana, a CIVICUS Digital Democracy Initiative (DDI) partner, used X/Twitter to educate citizens about the voting process ahead of Election Day. Meanwhile, Fundación Efecto Cocuyo's podcast series in Venezuela brought critical electoral issues into public discourse. They featured the National Electoral Council's head, Aimée del Nogal, in one episode, although the election results were widely seen as fraudulent.

CSOs can play a vital role in promoting informed participation, transparency, and electoral integrity, even in challenging contexts. There is a need for flexible support to scale up such initiatives to promote inclusive and credible electoral processes globally.

As civil society entered the 2024 election cycle, the emphasis shifted not only to observing the polls but also to sustaining civic space beyond the elections. CSOs prepared for post-election scenarios, potential contestation, digital manipulation flare-ups and campaign-era legacy issues.
CIVICUS's DDI supported civil society in tackling election-related challenges by enabling local organisations to counter disinformation, monitor elections with digital tools, and mobilise underrepresented communities for meaningful civic participation. Through strategic resourcing, training, and digital tools, we strengthened electoral integrity and protected democratic spaces.

Building monitoring coalitions and observation capacity

Many CSOs across the globe renewed or expanded election-monitoring networks to respond to the 2024 cycle. Domestic coalitions partnered with electoral commissions, social media platforms, and media organisations to develop rapid incident-reporting systems, scenario planning, and election-day observation protocols.

Case Study: South Africa
Ahead of the 2024 national and provincial elections, South Africa's Indepedent Electoral Commission (IEC) partnered with major social media companies, including Meta (Facebook, Instagram, WhatsApp), Google (YouTube), and TikTok to curb the spread of electoral disinformation. The partnerships aimed to promote credible election information, strengthen content moderation systems, and provide direct reporting channels for false or harmful content related to the elections. It also involved collaboration with fact-checking organisations and the Real411 platform to ensure timely identification and removal of misleading online material. The initiative formed part of the IEC's broader commitment to maintaining electoral integrity and public trust in South Africa's democratic processes.

Digital preparedness and deepfake disinformation mitigation

Given the rising threat of manipulated media and disinformation, CSOs prioritised digital literacy, early warning mechanisms, and public information campaigns. In many election-year contexts, digital platforms, civic tech groups, and media watchers supported civil society efforts to monitor social media trends and coordinate responses.

Case Study: India
India's BOOM Live, BOOM, an independent digital journalism initiative launched a Deepfake tracker. The tracker combined investigative journalism with cutting-edge forensic analysis to expose and explain AI-generated deception. Using a blend of reverse image searches, frame-by-frame video analysis, and metadata forensics, BOOM's team traced the digital fingerprints of manipulated media circulating across Indian social platforms. Each verified deepfake is catalogued with context: who shared it, how it spread, and what narrative it sought to push. Complementing the tracker, BOOM's deepfake literacy campaign turned detection into public education; through workshops, explainers, and interactive videos, it taught citizens how to spot tell-tale visual inconsistencies and question too-perfect "evidence." Together, these efforts transform technical verification into civic empowerment, helping audiences see not just what is fake, but why it matters.

Advocacy, civic education, and inclusive voter engagement

CSOs sought to amplify civic participation and safeguard the electoral process by targeting historically excluded groups, women, and youth. They ran outreach programmes to explain electoral rights, boost voter registration, and help citizens identify misinformation.
In contexts of shrinking civic space, many organisations also developed contingency

plans for legal and advocacy responses, from emergency observation statements to partnerships with international oversight bodies.

Case Study: USA
The League of Women Voters (LWV) in the United States launched a nationwide initiative ahead of the 2024 elections to counter disinformation aimed at women, young voters, and communities of colour—groups historically targeted with voter suppression efforts. Through its "Misinfo 101" campaign and partnerships with grassroots organisations such as Voto Latino and Black Voters Matter, the LWV trained local volunteers to identify and report misleading narratives about voting eligibility, mail-in ballots, and polling stations. The campaign also created multilingual, culturally tailored explainer videos and WhatsApp fact sheets to reach non-English-speaking communities. By combining digital monitoring with community-based voter education, the initiative helped inoculate traditionally excluded groups against coordinated disinformation designed to discourage or confuse them at the ballot box.

3.4. Lessons Learned from the 2024/25 Electoral Cycle

The 2024/25 election cycle offers numerous lessons. The first is simple: prepare, prepare, prepare. It is far better to over than underprepare.

Early warnings can sometimes cause alarm, but new technological threats demand thorough human impact assessments and awareness of both potential and real impacts. In 2024, what could be termed as alarmism led to public awareness, platform intervention, CSO readiness and regulatory action, which, collectively, may have helped reduce the impact of deepfake disinformation.

Another key lesson is that these assessments must happen before new technologies are widely adopted, not after. Reactive responses are always too late. Anticipatory strategies give societies the resilience to absorb shocks.

Lesson Learned	Insight	Policy / Strategic / Advocacy Interventions
Threats Were Overestimated	Only about **251** cases of deepfake disinformation were recorded globally, far fewer than expected.	Avoid alarmist narratives, ground policy in evidence, and prioritise proportional responses; conduct empirical studies to measure potential impact
Quality Over Quantity	Even limited incidents matter due to hyperrealism and emotional resonance.	Focus on detection, rapid response, and public awareness to neutralise "high-impact" deepfakes quickly
Context Shapes Impact	Deepfake effects varied depending on media systems, culture, and digital literacy.	Develop context-sensitive interventions (regional fact-checking, culturally tailored media literacy)
Unknown Actors & Attribution Gaps	**46%** of cases had untraceable origins; foreign/state-linked actors were implicated where identified.	Strengthen cross-border attribution mechanisms, invest in OSINT, and advocate for platform transparency

Constructive Uses Emerged	Parties also used GenAI for civic outreach and accessibility	Encourage positive applications of GenAI in campaigns, while regulating manipulative uses
Resilience Stronger Than Expected	Public awareness, platform measures, and civil society monitoring helped blunt impacts.	Scale up media literacy, strengthen CSO monitoring networks, and embed election-specific preparedness
Regulation Uneven	Laws exist in the EU, Singapore, the USA, and other jurisdictions, but enforcement varies widely.	Push for harmonised global/ regional standards; advocate for transparency obligations on platforms
Broader Harms Overlooked	Focus on elections risks sidelining the harms of deepfakes targeting women/minorities.	Expand definitions of "harmful deepfakes" beyond elections; push for gender-sensitive disinfo frameworks

PART 4: COUNTRY CASE STUDIES

In this section, we provide a brief overview of each of the four country's elections, as part of the case studies, accompanied by carefully and extensively contextualised social, cultural, economic and political data. This data depicts the conditions in which the elections took place and how these may have influenced the spread and prevalence of deepfake disinformation.

The intent is to unravel the vulnerabilities to and resilience factors against disinformation in general, and, where present, deepfakes. This isto encourage a focus on the root causes of disinformation and deepfakes rather than their occurrence.

Alongside the case studies is a proposed Deepfake Risk Matrix. The Matrix allows for more contextually grounded, evidence-based analysis of deepfake disinformation risks. It can facilitate a more nuanced understanding of vulnerabilities across diverse social, political, and technological contexts. The Matrix does this by clarifying vulnerabilities, indicators, and impacts across social, political, economic and regulatory factors as well as malign actors and technological domains. It assigns a risk level to guide appropriate strategic responses.

The Deepfake Risk Matrix can and should be tailored and contextualised.

Deepfake Risk Matrix

Factor	Key Indicators	Impact Metrics	Risk Level	Strategic Response
Social	• Media trust and literacy levels • Polarisation • Digital comms vs. cultural norms (e.g., WhatsApp as the primary method of communication)	• Erosion of political trust • Social fragmentation	Low–High	• Media literacy campaigns • Civic outreach • Community rumour tracking
Economic	• Internet penetration rate • Inequality (e.g., monetisation of disinformation)	• Market manipulation • Resource diversion	Low –High	• Require public labelling of paid-for posts • Platform accountability
Digital Ecosystem	• Internet/mobile access • Platform dominance	• Level of internet and network access • Legal restrictions on content or access to platforms	Low–Medium	• AI detection tools • Platform accountability

Regulatory Context	• Free and fair elections • Freedom of expression • Anti-deepfake legislation	• Overreach vs. underreach • Civic trust in institutions	Low - High	• Deepfake labeling • Platform liability frameworks
Actor Involvement	• Foreign/state actors • Commercial disinfo • Political actors	• Hate speech • Electoral interference	Low - High	• Investigative journalism • OSINT • Real-time debunking: international cooperation

4.1 Namibia

The people of Namibia went to the polls on 27 November 2024 to elect a new President and members of the National Assembly. Namibia operates a proportional hybrid democratic model that blends proportional representation with direct and indirect elections, designed to reflect both national and regional interests.

Despite a relatively high score of **80/100**, the CIVICUS Monitor ranks Namibia as having 'narrowed' civic space. Namibia's 2024 elections, initially scheduled for 27 November but then postponed to 29–30 November due to logistical delays, resulted in a historic victory for the SWAPO movement, as Netumbo Nandi-Ndaitwah became the country's first female President with **58.07%** of the vote.

The pre-election environment was described by Good Governance Africa (2024) as "one of the most fiercely contested in history" as SWAPO, once a dominant liberation movement, faced unprecedented competition. In 2019, President Hage Geingob secured re-election with just **56%** of the vote, a sharp decline from **87%** in 2014. By 2024, the ruling party had to contend with new political forces.

The win for SWAPO was legally contested by the opposition Independent Patriots for Change (IPC) and its allies, which filed a lawsuit on 15 January 2025 alleging voter suppression, irregularities, technical failures and ballot shortages. While the election has become one of the most disputed in Namibia's democratic history, the African Union Election Observation Mission reported that it was largely peaceful and conducted within the national legal framework.

The November 2024 vote took place amidst several labour actions and gender rights-related protests which addressed some of the prevailing socioeconomic issues in the country:

- Contract workers for the City of Windhoek protested for permanent employment and better health benefits, but the City initially refused to recognise their strike or accept their petition.
- The Namibian Economic Freedom Fighters demonstrated against a Rundu service station after a video surfaced showing employees being whipped in exchange for loans.
- Activists and politicians were arrested during a planned Independence Day protest highlighting youth unemployment, with police citing national security concerns over the planned protest.
- Gender activists marched to demand inclusive protections under the Domestic Violence Act, particularly for same-sex couples, and submitted a petition to the authorities.

Overall, Namibia retains a strong democratic culture. Freedom House's Freedom in the World Report (2025) gives Namibia a score of **73/100**, ranking it as "free" and recognising

its multiparty democracy and respect for civil liberties. The Freedom House report notes the following for Namibia:

- **Free and fair elections:** The electoral framework is robust and generally well implemented.
- **Political pluralism:** Opposition parties may freely compete in elections and generally do not encounter intimidation or harassment.
- **Freedom of expression:** Citizens are generally able to express political choices without undue influence from external actors.
- **Media freedom:** The constitution guarantees freedom of the press and freedom of expression; journalists face few legal restrictions and typically work without risk to their personal safety.
- **Women's political participation:** Women are often discouraged from running for office, and few contested the November 2020 regional and local elections.
- **Anti-corruption enforcement:** While Namibia has a sound legal framework, anti-corruption laws are inconsistently enforced.
- **Civil society:** CSOs generally operate without interference, though government leaders occasionally use public platforms to criticise them.

Although a Freedom on the Net report was not compiled for Namibia, the country has consistently been recognised for its strong protection of media freedom and expression, often ranking among the top five countries in Africa. These protections extend to online spaces. According to DataReportal, Namibia's internet penetration rate stands at **62%**, with around **23%** of citizens using social media, primarily Facebook and WhatsApp. However, a digital divide persists. Data costs remain prohibitively high and rural areas still experience limited connectivity.

Namibia currently lacks a specific law addressing disinformation. This area could be developed in the future, provided that any new regulation safeguards free expression and mitigates societal harms and gendered disinformation.

Public attitudes towards Namibia's democracy remain robust, representing strong resilience factors against disinformation. According to Afrobarometer (2024), nine out of ten Namibians (**90%**) feel completely free to vote, and **74%** report high levels of trust in the country's election authorities. These indicators reflect enduring democratic resilience and electoral integrity, both of which serve as buffers against disinformation.

Trust in the media is also high; **75%** of respondents believe journalists can report freely without government interference. This trust acts as another resilience factor against disinformation, as citizens still view the media as a credible source of verified and accurate information. By contrast, in countries such as the United States, where media trust is low, vulnerability to misinformation tends to be much higher as audiences turn to unreliable sources.

Namibia also benefits from a vibrant civil society sector, with numerous CSOs working on issues related to information integrity. These include Namibia Fact Check (fact-checking), the Editors' Forum of Namibia (media freedom), the Namibia Media Trust (digital rights), the Action Namibia Coalition (access to information), and the Women's Leadership Centre (technology-facilitated gender-based violence).

Deepfake Disinformation Report: Namibia

Namibia's 2024 election period saw a marked rise in misinformation, particularly on WhatsApp and Facebook, according to the country's Media Ombudsman. This was especially concerning given that nearly two-thirds of registered voters were young people who rely heavily on social media for political information (Media Ombudsman, 2025).

In partnership with the Institute for Public Policy Research (IPPR), Namibia Fact Check conducted a five-month misinformation monitoring project that revealed the following during the election

period:

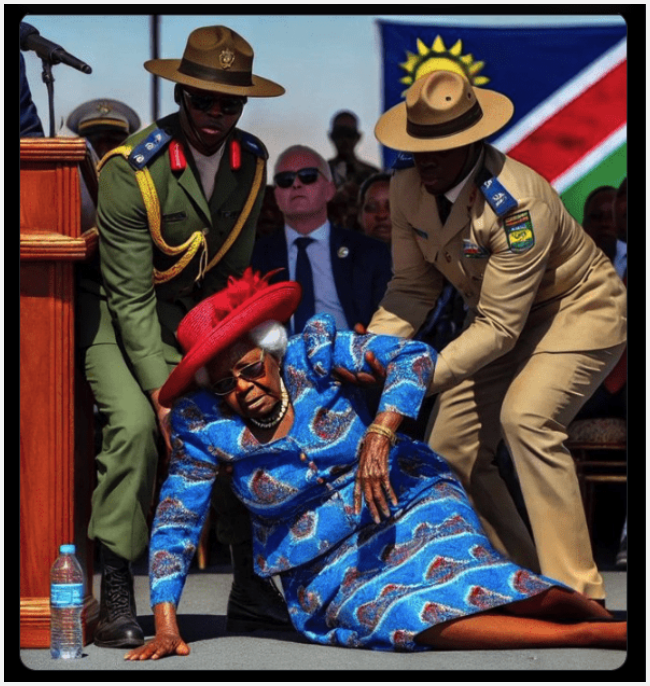
- WhatsApp remained the primary channel for circulating false election-related content, while TikTok, used for the first time in a Namibian election, emerged as a significant new vector of communication.
- Political parties, politicians, and their supporters were all complicit in spreading misinformation, primarily via social media and messaging platforms.
- Coordinated disinformation campaigns were observed across multiple platforms.
- Long-running smear campaigns targeted both ruling and opposition parties, their candidates, electoral authorities, and government institutions.
- These campaigns often involved foreign actors and influencers, with narratives laundered through domestic and international online sources, including manipulated news coverage.
- Namibian news outlets inadvertently amplified false information through superficial event-based reporting and insufficient fact-checking, while mistakes by regional and international media further eroded public trust.
- Electoral authorities appeared ill-equipped to counter the volume and sophistication of online misinformation, including attacks on their credibility.

Reflecting a broader Global South trend, Namibia saw no verified deepfakes during the 2024 elections. However, cheapfakes circulated widely on social media and messaging apps throughout the year. According to Namibia Fact Check and the IPPR, malign actors increasingly experimented with AI-assisted and low-tech manipulations to advance smear campaigns and launder false narratives.

Several notable cheapfakes were detected, though most were low quality and had limited impact such as:

- A video of former US President Joe Biden appearing to endorse Netumbo Nandi-Ndaitwah;
- An audio clip of a candidate making vulgar, tribalist remarks; and
- Various fabricated letters allegedly written by political candidates.

The most prominent cheapfake showed presidential candidate Nandi-Ndaitwah collapsing on stage during a rally. This incident formed part of a broader campaign of gendered disinformation, which fixated on her age and gender to question her fitness for leadership. Such narratives diverted attention from substantive political debate and sought to undermine public perceptions of her candidacy. Despite these efforts, she went on to win the presidency, becoming Namibia's first female head of state.



AI-generated / Deepfake image depicting the 'collapse' of Namibian Presidential Election candidate, Ms. Netumbo Nandi-Ndaitwah started circulating online at the end of October 2024.

Applying the Deepfake Risk Matrix to Namibia's Social Factors

This section applies the Deepfake Risk Matrix to Namibia, focusing on the social dimension of disinformation risks, particularly gendered disinformation, which remains a growing but underaddressed issue in the country's electoral context. Other case studies in this series assess different dimensions of the Matrix, including political, technical, and institutional responses.

Interview: Nambia

CIVICUS discussed the Namibian General Election with Frederico Links, Editor and Project Coordinator for Nambia Fact Check, a CSO and project of the Institute for Public Policy Research (IPPR) started in mid-2019 in response to the rise in political disinformation and propaganda in online political spaces.

We ran two projects concurrently ahead of, during, and after the 27 November 2024 parliamentary and presidential elections.

One was the coordination of a coalition of media and civil society organisations to counter election-related mis- and disinformation, supported by Africa Check with funding from the Google News Initiative. The project sought to help voters critically engage with information and make informed decisions in the voting booth. The coalition collaborated to fact-check politicians and political party claims, provide voters with reliable, nonpartisan information on key issues, and equip the public with the skills they need to identify election misinformation.

The other project we had running was one identifying, monitoring, and tracking election-related mis- and disinformation narratives, actors, and pathways. A report was produced and launched earlier this year capturing what we observed on the electoral information landscape. We saw that social media has become important for political communication and messaging, especially during electoral periods in Namibia. However, political actors (politicians and parties) often used social media to engage in negative campaigning by employing smear campaigns and spreading fake news to intimidate and incite. We also found that the actors associated with or supporting various political parties or causes were behind much of the election-related mis- and disinformation that was circulating across social media and messaging platforms before, during, and after the 27 November 2024 elections.

The emergent use of AI tools to generate content for online dissemination indicates that AI-generated political content, including disinformation content, will probably become a political and electoral headache down the line. While the AI-generated disinformation content observed on the Namibian electoral information landscape was still rather basic or crude and detectable, it seems clear that disinformation actors were experimenting and probably will become better and more sophisticated moving forward. This means that more sophisticated and hard-to-detect AI-generated content, including the use of deep-fakes, will increasingly appear on Namibian political and electoral information landscapes, probably sooner rather than later.

Social Factor	Assessment	Gendered Disinformation Risk
Media Trust & Literacy	Moderate trust in traditional media, but high reliance on social media, with different levels of digital literacy.	Women and rural voters are more vulnerable to manipulated narratives due to limited media literacy.
Social Polarisation	Rising political tensions, especially between SWAPO and emerging opposition parties	Female candidates may be targeted with divisive deepfakes to exploit identity politics.
Cultural Norms & Gender Roles	Patriarchal norms persist, especially in rural areas; women are underrepresented in politics.	Deepfakes often weaponise gender stereotypes to discredit women's leadership and credibility.
Minority Representation	The Ovambo majority dominates; the San and other minorities face systemic exclusion.	Minority women are doubly excluded and vulnerable to invisibilisation and targeted misinformation.
Social Media Behaviour	High virality of unverified content; WhatsApp and Facebook dominate.	Gendered rumours and doctored images spread rapidly in closed networks with little content moderation.
Civil Society Engagement	There are strong watchdogs and women's rights groups, but their reach is limited in remote regions.	CSOs are potential allies in countering gendered disinformation, but they need capacity-building and digital tools.

Strategic interventions targeted at curbing gendered disinformation could include:

- **Targeted Media Literacy:** Prioritise outreach to women, youth, and rural communities with gender-aware digital education.
- **Safe Reporting Channels:** Establish anonymous platforms for female candidates and activists to report deepfake abuse.
- **Narrative Resilience:** Support women-led storytelling initiatives to counteract disinformation with authentic voices.
- **Intersectional Monitoring:** Track disinformation trends that intersect with gender, ethnicity, and geography to enable more innovative interventions.
- **Regulatory Intervention:** A legal framework targeting disinformation in general, including gendered disinformation, with a focus on protecting free speech, while still mitigating harms.

4.2 Ecuador

The Ecuadorian General Election took place on 9 February 2025 to elect a President, the National Assembly, **21** provincial assemblies, and Ecuador's representatives to the Andean Parliament.

Ecuador has a proportional representation electoral system. Voting is compulsory for citizens aged **18 to 65**. For the presidential election, at least **40%** of the valid vote is required, and a difference of at least **10%** from the second in place candidate is needed. If no candidate achieves this, a second round is held

between the two candidates with the most votes. This happened, and a run-off election was held on 13 April 2025, with incumbent President Daniel Noboa re-elected, defeating Luisa González of the Citizen Revolution Movement.

As a country where gang violence has surged, resulting in journalists and media outlets becoming frequent targets of attacks, we ranked Ecuador as having 'obstructed' **(48/100)** civic space in the 2024 CIVICUS Monitor. As an example, following the declaration of a 60-day state of emergency after the escape of a gang leader in January 2024, heavily armed gang members stormed the TC Televisión studio in Guayaquil, interrupting a live broadcast of the El Noticiero news programme. Attackers threatened media workers with firearms, brandished grenades, and fired shots, while it was broadcased live across the country.

Freedom House's Freedom in the World Report (2025) ranks Ecuador as "partly free." While the country holds regular, competitive elections, the influence of organised crime and related violence has increased significantly in recent years, affecting the functioning of state institutions and the security of ordinary citizens. Due process violations, attacks on journalists, human rights abuses and official corruption are ongoing challenges.

A mix of positive and negative factors also contributed to this "partly free" ranking, including:

- Multiple parties compete in Ecuador's political system, but historically they have mainly been personality-based, clientelist, and weak in governance.
- Criminal organisations and related violence increasingly constrain Ecuadorians' political choices. The surge in political violence in 2023 and concerns over illicit campaign financing persisted in 2024. At least three mayors were murdered during the year.
- Citizens enjoy formal political equality, regardless of race, gender, and other such distinctions, though some disparities in access and influence persist in practice. Electoral regulations mandate that women account for 50% of the candidates on party lists in multimember districts.
- Ecuador has long been plagued by corruption, with a weak judiciary and lack of investigative capacity in government oversight agencies, which has contributed to impunity.

Ecuador's Freedom on the Net score is "partly free." According to Datareportal, internet penetration in Ecuador is **83.6%**, with **75%** of the population using social media. This high internet penetration rate, however, masks a digital divide. The Freedom on the Net report states that although Internet access has become more affordable in recent years, it remains prohibitively expensive, with average broadband prices in Ecuador still higher than in many other South American countries. This means that while many are online, the online presence may vary, and this certainly affects citizens' ability to engage in digital activism and participate meaningfully in digital democracy, leaving space for the elite to form their own election narratives relatively unchallenged. This is a significant disinformation vulnerability.

Ecuador's journalists and communicators face significant threats to their safety, especially when reporting on elections and sensitive topics. These threats are sometimes levelled by state actors or organised crime. Some journalists who report on politically sensitive issues have either been forced into exile due to threats against their physical safety or they self-censor through anonymisation. This is perhaps one of the most significant vulnerabilities to disinformation. Without an open and free media space, there is less rigour to fact-check, contextualise, and challenge false claims and information emanating from the government and other actors. In such contexts, disinformation spreads with little resistance. Restricted media environments also drive citizens to informal channels, such as social media or rumours, where content is harder to verify and more easily manipulated. In such contexts, trust in credible information sources declines, accountability weakens, and election integrity is threatened as manipulative

narratives gain greater traction.

Another concern is that while the Ecuadorian government does not employ overt technical censorship, it does so through other means, such as legislative measures, including:

- A provision in the 2015 Organic Law of Telecommunications grants the President unilateral power to take over telecommunications services during a national emergency. CSOs have repeatedly raised concerns about the provision's scope and the possibility of government abuse, given the law's vague standards and lack of independent or impartial oversight. It does not, however, appear to have been invoked, but the existing potential remains a threat.
- Under former President Rafael Correa Delgado, copyright law was frequently used to censor politically sensitive content online. This practice has lessened considerably, but not completely. Journalists have sometimes been pressured to remove content after receiving threats.

On a positive note, digital repression in Ecuador appears to be low, with little evidence of systematic blocking of content. Social media, communications apps, blogs, forums, and circumvention tools are generally accessible. Numerous digital media outlets have emerged over the past decade, and users typically do not need VPNs to access online news. In addition, there are no legal restrictions on digital advocacy or online communities, and social media continued to serve as a tool for social and political mobilisation in Ecuador during the period under review.

The pre-election environment in Ecuador was marked by violence, both in the lead-up to the first vote and run-off, with candidates routinely subjected to death threats. In this digital age, such threats go hand in hand with smear campaigns, harassment, cyberstalking, misinformation, and disinformation.

Indeed, the online space in Ecuador was so dire and disinformation campaigns so pervasive that Freedom House in 2024 noted an online information space "charged with polarisation," including:

- Evidence that inauthentic bot accounts shape online discussions. As an example, following a 2023 presidential debate, social media analysis found that **73%** of posts mentioning Noboa were created by bots or potential bots, compared to **72%** for González.
- Troll accounts seeking to support or discredit specific candidates were seemingly deployed across social networks, including on platforms that had previously been less utilised for these purposes, such as WhatsApp.
- In 2022, Twitter suspended a "botnet" operation composed of **491** inauthentic accounts for engaging in Coordinated Inauthentic Behaviour (CIB) in support of then-President Lasso.
- Online troll accounts were reportedly deployed on behalf of individuals involved in organised crime.
- False or misleading content is often spread through digital platforms and social networks, including information about government officials or political candidates, undermining the reliability of the online information environment
- Political actors used their online platforms to discredit certain journalists.
- Deepfakes were not identified.

The offline environment was equally dire. On 9 August 2023, a presidential candidate was assassinated after a campaign event. In December 2024, President Noboa discovered a car bomb near a place he had planned a campaign stop. Opposition candidate Luisa González reported death threats and increased her security, while Socialist candidate Pedro Granja suspended his campaign due to an attack. Centrist Jimmy Jairala's car was shot at the start of his campaign (Boscán, 2025).

Violence fuelled by the drug trade is pervasive in Ecuador, with no sector of society unaffected, from schools and hospitals to polling stations.

Violence seems to be everywhere in Ecuador, affecting its education, health care, and politics. The candidates vying to lead the South American country—considered one of the most violent in the world in recent years—are well aware of this reality. Insecurity is the primary concern of their voters. **7 in 10** *Ecuadorians fear going out at night, and the country ranks worst on Gallup's Law and Order Index, which annually measures the perception of security in* **140** *countries.*

Digital and media literacy in schools faces challenges in being taught and integrated due to the violence students encounter in public spaces. Since 2022, for example, approximately **90,000 children** have dropped out of school in Ecuador. Homicide is the top cause of death among minors, and criminal gangs extort students for school access; **20%** of children avoid classes out of fear (Boscán, 2025).

CSOs in Ecuador, however, work against disinformation, including organisations such as Fundamedios (press freedom, digital rights), Ecuador Verifica (fact-checking, electoral integrity), Openlab Ecuador (civic tech, disinformation innovation), Observatory of Communication (media analysis, academic research), and Technical University of Loja (media literacy, teacher training).

Deepfake Disinformation Report: Ecuador

Given the above vulnerabilities, it is unsurprising that disinformation was prevalent, including in the post-election period. EU vs Disinfo (2025) reported an attempt by pro-Kremlin operatives to sow doubt about the electoral results in Ecuador. The false claims were rejected by Ecuador's National Electoral Council, whose conclusions were supported by the Organisation of American States (OAS) electoral mission (EOM).

OAS/EOM is a deployment mechanism of a regional body comprising 35 member states across the Americas, focused on promoting democracy, human rights, security and development. It deploys independent experts to monitor elections and assess legality, transparency, media access and voter participation. Their reports are widely regarded as credible and influential, though the OAS has been accused of bias at times. But the election missions are generally seen as rigorous, impartial, and essential safeguards of democratic standards in the region.

As for deepfake disinformation, based on the available information, it is unfortunately unclear whether deepfakes, as technically defined, were disseminated. Reports rather speak to the dissemination of "AI-generated misinformation" or "AI-manipulated content."

In their report published in February 2025, the OAS/EOM warned of a rise in the use of AI; **23%** of viral disinformation included AI-generated material. Verified false information still circulated in journalists and public figures accounts, possibly due to inadequate source verification.

The EU Observer Mission to Ecuador similarly cautioned against a "widespread dissemination of manipulated content on all platforms, frequently amplified through proxy accounts, paid-for content and bot farms," having found the use of AI-generated content increasing throughout the election campaign, and often employed to fuel disinformation and personal attacks against candidates. The content was disseminated across Facebook, Instagram, X and TikTok. The Observer Mission further found:

- **63** cases of "manipulated" videos and **55** such images, and cloned voices or altered audio on six occasions.
- **56** of these items were shared by accounts suspected of being trolls or bots, **48** by influencers or content creators, and **27** by ordinary users.
- While some of the identified AI-generated content was used for satire or genuine campaign promotion, the majority aimed to delegitimise political opponents (**91** cases) and spread disinformation (seven cases).

Applying the Deepfake Risk Matrix to Ecuador's Economic and Digital Ecosystem

For Ecuador, we apply the economic and digital ecosystem variables to understand how the unique conditions in the country could contribute to the proliferation of deepfakes

Interview: Ecuador
CIVICUS discussed Ecuador's presidential election with Jorge Tapia de los Reyes, Coordinator of the Democracy and Politics Department and the Political Funding Observatory of the Citizenship and Development Foundation (FCD). FCD is an Ecuadorian civil society organisation that promotes participation, citizen monitoring, and open government.

The role of organised citizens was crucial to the success of the democratic process. Through various monitoring and observation initiatives, civil society acted as an effective counterweight to potential irregularities.The work of civil society went beyond election observation; we are committed to building an informed and critical citizenry. We understand democracy not only as an act of voting, but as a continuous process of education, information, and participation. With this in mind, we set up a system to monitor and verify fake news on social media to combat disinformation and its harmful effects on the electoral process.

Many disinformation campaigns are specifically designed to create fear and apathy and discourage participation. Our work sought to counter these strategies by providing verified information and reminding people that only the National Electoral Council has the legal authority to issue official results or respond to reports of irregularities.

Factors	Contextual Indicators	Disinformation Risk Assessment
Economic Inequality & Informality	High informality rate; persistent urban–rural divide; economic precarity among youth and rural voters	Economic grievances can be weaponised through populist disinfo and AI-generated narratives.
Digital Penetration & Access	**83.7%** internet penetration; **98.8%** mobile connectivity; rural access still lags behind urban centres.	High connectivity enables rapid disinfo spread; rural gaps hinder verification and response.
Digital Literacy & Education	The country has a low ranking in the global digital skills index and unequal access to digital education.	Vulnerable populations, especially youth and older voters, are more susceptible to media manipulation.
Platform Ecosystem	Dominated by WhatsApp, Facebook, and TikTok (**74%** social media usage)	Encrypted platforms and short-form video apps amplify disinfo with limited moderation.

Strategic interventions targeted at improving the digital ecosystem and other socioeconomic risks could include:

- Advocating for rural digital infrastructure through targeted investment and regulatory reform to lower internet costs, especially in underserved regions;
- Supporting civic tech innovation via initiatives which bring together journalists and developers to build AI-powered fact-checking and transparency tools; and

- Launching national digital literacy campaigns, prioritising youth, rural voters, women and other groups most vulnerable to synthetic media manipulation.

4.3 Germany

The German Federal Election was held on 23 February 2025 to elect the **630** members of the country's Bundestag (Parliament). The members of the Bundestag are directly elected every four years by German citizens through a mixed-member proportional system. Once the Bundestag is sworn in, it elects the Chancellor from its ranks.

Due to suppression of protest movements in the country, particularly related to Palestine, Germany is rated as having 'narrowed' civic space on the CIVICUS Monitor, with a score of **67/100**.

During the pre-election period, videos on social media showed cases of officers pushing, punching, and choking non-resisting protesters in Germany. In one case, a protester was injured to the point of losing consciousness and was reportedly not given any medical assistance for 20 minutes. Palestinian solidarity protests also faced bans and obstruction;journalists were reportedly detained while covering these protests and documenting the police use of excessive force.

In its annual Freedom in the World Report 2025, Freedom House ranks Germany as "free" with a **93/100** score, finding that:

> *Germany is a representative democracy with a vibrant political culture and civil society. Political rights and civil liberties are assured mainly in law and practice. The political system is influenced by the country's totalitarian past, with constitutional safeguards designed to prevent authoritarian rule.*

While Germany has remained firmly committed to liberal democracy and its tenets in the 21st century, the rise of right-wing populism and inflammatory anti-immigrant rhetoric threatens to blemish this record. Such narratives are ripe for disinformation and are at risk of being weaponised to polarise German society.

Freedom House has documented several concerning developments over the past few years in Germany, as follows:

- While freedom of belief is protected by law, eight states ban headscarves for teachers, and Berlin and Hesse prohibit civil servants from wearing them.
- Antisemitism has been on the rise. The Ministry of the Interior recorded an exponential increase in attacks on individuals of the Jewish faith.
- Islamophobia also remains a concern. Attacks against those of the Muslim faith have also increased.
- In late 2023, after the 7 October Hamas attacks and the war in Gaza, pro-Palestinian protests in Germany were heavily restricted. (By early 2024, some restrictions were challenged in court, and some were overturned, but restrictions on slogans or conditions on protest size remained common. Incidents of police brutality against pro-Palestine protesters have also been reported.
- Attacks on refugees and refugee housing declined from approximately **3,500** such cases in 2016, but have nonetheless continued to occur in the country.

Germany is classified as "free" according to Freedom on the Net, with a score of **77 out of 100**, which reflects near-universal internet access (**93%** penetration per Datareportal), as well as a strong online media environment (**77%** of Germans use social media) and a strong and fair judiciary. However, the country faces ongoing challenges from Russian disinformation campaigns and cyberattacks, as well as accusations of censorship. Several related developments are as follows:

- The government occasionally blocks websites or other online content. A large

number of Russian-linked websites and accounts are blocked because of alleged propaganda targeting civil society.

- German online spaces are marked with significant content manipulation, which in some cases has been linked to the far right or to foreign interference campaigns, primarily originating in Russia. Examples include:
  - Far-right actors have spread false and misleading information online. Members of Parliament for the right-wing populist Alternative for Germany (AfD) party have used AI image-generation tools to spread xenophobic messages.
  - Disinformation campaigns from Russian actors also continue to target Germany, and some have involved German politicians. A March 2024 investigation conducted by the Czech Republic's Security Information Service reported on a Russian campaign that had allegedly approached and paid European politicians, including members of the Bundestag, to question the "territorial integrity, sovereignty, and freedom of Ukraine" on the pro-Kremlin news site Voice of Europe.
  - In January 2024, the country's foreign office uncovered **50,000** fake accounts on X that were posted in German and which spread messages critical of Germany's support for Ukraine and the current German government.

Germany has one of the most tightly regulated digital spaces in the world, shaped by both national laws, such as the NetzDG, and EU-wide frameworks, such as the GDPR and Digital Services Act (DSA). There have been concerns of over-regulation from various CSOs and the burden of bureaucratic compliance with laws such as the GDPR.

Ahead of the 2025 election, the EU's AI Act was passed and viewed as a significant step to tackling deepfake disinformation. The AI Act defines a "deep fake" as an "AI-generated or manipulated image, audio or video content that resembles existing persons, objects, places, entities or events and would falsely appear to a person to be authentic or truthful."

It requires the following compliance:

- Deepfake developers and users must disclose AI-generated content to prevent misinformation.
- AI content should be labelled through classification or watermarking.
- Deepfakes that may impact rights or society, such as political manipulation or defamation, are high-risk and face stricter regulations.
- Traceability and accountability require keeping records of deepfake creation for possible origin tracking.
- Malicious uses for social scoring or illegal surveillance are banned as unacceptable risks under the AI Act.

Failure to comply with these requirements can trigger steep penalties, with fines of up to **€35 million** or **7%** of a company's global annual turnover for serious violations. By attaching such weighty sanctions, the AI Act signals the EU's intention to treat deepfake transparency as a matter of democratic integrity.

As Germany headed to the polls in 2025, key issues included: the rising cost of living, an energy crisis, and declining confidence in government, with faith in the German government having dropped to **50%** in 2024 - its lowest point in over a decade and down considerably from a high point of **65%** in 2020 (Gallup, 2025). Low trust in government marks a significant vulnerability to disinformation. When audiences lose trust in government, this can lead them to seek information elsewhere, creating an information vacuum for malign actors to fill.

The rise of the far-right poses another concern. As Election Day approached, Germany entered what was described as an "unusually tense" campaign (ACLED, 2025) due to a political landscape that had shifted from centrism to greater polarisation, as evidenced by widespread protests before the elections. Weeks before the election, hundreds of

thousands of Germans protested nationwide after the centre-right Christian Democratic Union relied on support from the far-right Alternative for Germany (AfD) to push a parliamentary motion for stricter migration laws.

ACLED recorded a **17%** rise to over **4,650** distinct demonstration events, with **1,450** demonstrations opposing right-wing extremism, and which were predominantly led by grassroots movements, CSOs, and centrist parties. This was a welcome pushback, reflected by the AfD's loss of support after the ballots were counted. Concerns over disinformation were high amongst the German public, with a poll finding that **88%** feared manipulation from foreign actors or governments (Shelton, 2025). This is a strong resilience factor, a signal that the population is aware of and inoculated against the threat of disinformation.

Deefake Disinformation Report: Germany

There are no reports thus far that comprehensively cover the amount of disinformation in general and, particularly, the number of deepfakes in Germany. This could be a sign of "deepfake fatigue" or "no news is good news" after the 2024-2025 super election period. Or perhaps that the AI Act, and similar actions, represent a successful deterrent in the EU.

Some data can be collated from two sources, outlined below, but they are somewhat limited in detail.

1. NewsGuard and Correktiv

Newsguard is a private, US-based company that provides clients with tools to evaluate the credibility of online news and information sources. In partnership with Correktiv, a German independent newsroom, Newsguard identified **22 false claims** related to the election in Germany, including disinformation by Russian actors targeting mainstream political parties that support NATO and Ukraine.

Included was a network of **102** AI-generated German-language fake news websites, allegedly linked to US fugitive-turned-Kremlin-propagandist John Mark Dougan. Using AI tools like OpenAI's ChatGPT and DALL-E 3, Dougan is said to have created over 160 fake news sites, disseminating false narratives to millions worldwide, and to Germans, content favourable to the AfD.

Dougan, crowned 2024 Disinformer of the Year by NewsGuard, is a former Florida deputy sheriff, now a "source of information on Russia", who fled the US to Russia while facing a slew of charges, including extortion. He has denied any ownership of the websites.

The websites bear the names of well-known, defunct German media brands and are filled with AI-generated content. The sites look similar and publish articles containing false information about German politicians who are pro-NATO and pro-Ukraine, particularly from the Green Party, which is known for its strong support for Ukraine and the green transition.

There is no reference to the content having been deepfakes; instead, it points to AI-generated content and images, with less emphasis on video and audio, to populate the "news stories."

2. The Institute for Strategic Dialogue (ISD)

The ISD uncovered a coordinated network on X (formerly Twitter) spreading disinformation about German politicians and election-related terror threats. The network of approximately 50 accounts shared traits typical of pro-Russia campaign influence operations, disseminating false claims through videos designed to look like they come from media outlets, law enforcement agencies, and academics. At times, AI-generated deepfakes are used for audio and visual content.

A secondary network comprising more than **6,000 accounts** dedicated to reposting content to ensure reach and virality exhibited the hallmarks of coordinated inauthentic behaviour.

Between 27 and 30 January 2025 alone, the network produced **19 election-related videos**, marking an escalation in activity. These included disinformation about supposed terror threats

tied to the elections and false accusations of corruption and paedophilia targeting prominent political figures, including CDU chancellor candidate Friedrich Merz, Die Linke's Janine Wissler, and Armin Laschet, the former CDU chancellor candidate. To boost credibility, the network branded its AI-manipulated videos with logos from respected outlets such as Deutsche Welle, BBC, and Sky News, while also impersonating government bodies and universities.

The tactics, AI-altered audio, fake captions, QR codes and tagging of journalists, mirror those of the Russia-aligned "Operation Overload" (also known as Matryoshka). Although the campaign's domestic impact appears limited, with most engagement driven by bot networks, the **48 core accounts** nonetheless attracted **2.5 million views**, with engagement tripling in January 2025. Interestingly, the disinformation was disseminated not in German but in English, Spanish, and Arabic, suggesting its true objective was less about swaying German voters and more about undermining confidence in German democracy among international audiences, thereby aligning with broader Kremlin influence strategies across Europe.

Interview: Germany

CIVICUS discussed AI governance challenges with Federica Marconi, researcher in the multilateralism and global governance programme at Italy's Institute of International Affairs, a non-profit think tank that promotes awareness of international politics and contributes to the advancement of European integration and multilateral cooperation. Marconi noted the need for the greater inclusion of civil society in the implementation of AI governance. Part of AI governance includes regulations against the use of deepfakes.

Given AI's impact across sectors, legitimate regulation requires meaningful civil society inclusion. Civil society organisations provide technical expertise, amplify excluded groups' perspectives and advance transparency and accountability. Their participation is crucial to prevent decisions from being dominated by powerful private stakeholders that are driven by economic interests rather than the public good.

Civil society's role is widely acknowledged as essential, but it faces two problems: getting access and having influence.

Access to multilateral forums varies. Some arenas restrict civil society participation entirely; others have established structured channels. But even when these mechanisms exist, access alone isn't enough: having a seat at the table doesn't guarantee that civil society voices will shape decisions.

The solution requires overcoming the notion that state leadership and stakeholder participation are competing legitimacy models. Civil society perspectives can be incorporated through governments via national consultations, advisory bodies or official delegations, while civil society can also engage independently with multilateral institutions through established participation channels.

Applying the Deepfake Risk Matrix to the Regulatory Framework in Germany

Germany's regulatory context illustrates how proactive risk management can both anticipate and mitigate deepfake disinformation threats. Together with the EU, Germany has built a forward-looking legal framework that recognises deepfakes as a systemic risk rather than an afterthought.

That being said, gaps remain; the AI Act still needs German operationalisation, and platform compliance is uneven, leaving spaces where manipulative content can still circulate. The German case thus shows how robust regulation can address critical vulnerabilities before they escalate, while also underscoring the importance of timely, coordinated risk identification.

Regulatory Framework Matrix – Germany (2025)

Dimension	Contextual Indicators	Disinformation Risk Assessment
Legal Framework	Germany applies the EU AI Act, with core provisions already in force since February 2025. Domestic implementation is underway via the draft KI Market Surveillance Act (KIMÜG).	A strong legal foundation is emerging, but enforcement gaps remain until the full rollout in 2026.
Constitutional Safeguards	Rooted in lessons from Germany's totalitarian past, the country has robust protections against authoritarianism and propaganda.	High resilience to state-sponsored disinformation; strong judicial oversight and civil liberties.
Platform Regulation	Relies on the EU Digital Services Act (DSA) and national consumer protection laws for platform accountability.	Platforms are legally bound to moderate harmful content, though enforcement still varies by provider.
AI-Specific Oversight	No standalone German-specific AI law yet; oversight will fall to market surveillance authorities under KIMÜG.	Institutional capacity is growing, but current fragmentation limits proactive disinformation detection.
Civil Society & Media Freedom	Germany scores **93/100** according to Freedom House; the press and watchdog ecosystem remains strong and active.	Strong resilience through media and fact-checking, with civil society playing a central watchdog role.
International Engagement	Germany is a key driver of EU efforts on AI ethics, human rights, and democratic safeguards.	High global influence helps shape international norms around AI and disinformation governance.

4.4 Singapore

On 3 May 2025, Singapore went to the polls to elect 97 Members of Parliament through a mix of Single Member Constituencies or Group Representation Constituencies in a parliamentary system. Once Parliament is constituted, the President (who holds a largely ceremonial role and is elected in a separate election) appoints the head of the majority political party as the Prime Minister.

We rate Singapore's civic space as 'repressed' in the CIVICUS Monitor's People Power Under Attack report. There are ongoing concerns over the use of restrictive laws and the harassment of human rights defenders. The 2019 Protection from Online Falsehoods and Manipulation Act (POFMA) grants excessive powers to the government and has been used to target activists and critics, and block websites. In addition, there are ongoing restrictions on peaceful assembly under the 2009 Public Order Act (POA).

Freedom House's Freedom of the World 2025 ranks Singapore as "partly free," with a score of **48/100**. One party, the ruling People's Action Party (PAP) and the powerful Lee family, have dominated Singaporean politics since the country gained independence in 1959. Due to budget constraints, Freedom House was unable to publish a full report on Singapore; therefore, a similar report, Reporters Without Borders (RSF) World Press Freedom Index (2025), is relied upon to ensure some consistency

across the case studies. While it measures press freedom, this is an indicator critical to assessing whether an information landscape is resilient or vulnerable to disinformation.

RSF ranks Singapore low across its indicators, placing it **126th out of 180** countries worldwide in terms of press freedom. The RSF report states, "While Singapore boasts of being a model for economic development, it is an example of what not to be regarding freedom of the press." Specific concerns include:

- Political Context: The PAP appoints media board members and editors who must enforce the official position of the government. Authorities control foreign media access in Singapore.
- Legal Framework: Singapore's "anti-fake news" law permits the government to correct online content deemed false or that affects public confidence. The Foreign Interference (Countermeasures) Act of 2023 expands state authority over media.
- Economic Context: There is a lack of plurality in the media; the sector is dominated by two groups: MediaCorp (state-owned) and SPH Media Trust (government-funded). Independent outlets often self-censor due to legal and financial pressures.
- Socio-cultural Context: "Out of bounds markers" restrict coverage of sensitive issues, resulting in self-censorship and the predominant dissemination of government-approved views on topics such as labour and human rights.
- Safety and Security: Journalists and bloggers face lawsuits, defamation claims, and smear campaigns from ruling party figures and supporters for their critical reporting.

On internet freedom, Singapore does not fare much better in Freedom House's 2024 Freedom on the Net report. While boasting an internet penetration even higher than Germany's at **95.8%** per Datareportal, (with **88.2%** of the population using social media, Singapore exemplifies that internet access does not equate to the "democratisation" of technology. Several examples are as follows:

- The government routinely orders internet service providers (ISPs) to block access to websites it deems to have published "false information" in violation of the terms of its "anti-fake news law". As an example, in May 2023, Singapore's Ministry of Communications and Information (MCI) ordered the independent news site Asia Sentinel to publish a correction notice after it reported on alleged state harassment of political dissenters. Asia Sentinel placed the notice under an editor's note, affirming it stood by the story; however, authorities deemed it as non-compliant. The following month, the Infocomm Media Development Authority (IMDA) instructed ISPs to block access to the site under Section 11 of POFMA. The website appears to still be blocked at the time of writing.
- The Online Criminal Harms Act (OCHA) was passed in July 2023, giving authorities new powers to block online content, services, and applications.
- The Foreign Interference (Countermeasures) Act (FICA) of 2021 came into effect in December 2023, granting the authorities broad latitude to restrict online activity. It also allows the authorities to designate individuals and organisations as "politically significant persons", which requires them to submit regular reports to the government on their foreign affiliations and donations from foreign citizens.
- The process for restricting online and digital content lacks explicit provisions regarding transparency. Recent legislation intended to address online falsehoods, foreign influence operations, and criminal activity online does not provide independent appeals mechanisms.
- Self-censorship occurs among journalists, commentators, and users who are typically aware that some types of speech or expression may result in civil or criminal consequences.
- While opposition parties can use social media for campaign purposes, their

activities are significantly constrained, however, by police investigations and arrests of those participating in online activism.

There are very few independent civil society organisations in Singapore that investigate or fact-check disinformation. In fact, only one such body, BlackDot Research, appears to be more of a market research firm than an independent CSO.

The true extent of disinformation and deepfake threats is, therefore, difficult to quantify. There are often vague references to "foreign manipulation," with very few empirical studies from CSOs to substantiate these. Most reporting on the subject comes through government channels, and in an environment where media and internet freedoms are tightly restricted, it is often unclear where the truth lies.

Rather than plural independent assessments, Singapore relies heavily on legislation and state-driven interventions, including a raft of laws empowering authorities to determine what constitutes a falsehood. In 2024, this framework was expanded to include a law on deepfakes, further consolidating state oversight of online information.

On 15 October 2024, Singapore's Parliament passed an amendment to the Elections Act, prohibiting the publication, boosting, sharing and reposting of deepfake content depicting election candidates, as well as banning the use of deepfakes during general elections. Outside of elections, deepfakes are regulated by broader laws, such as the "anti-fake news law," or by criminal statutes if they involve defamation, fraud or harm.

The lead up to the 2025 election was termed a "foregone conclusion" for PAP, despite the fact that the Party had not yet faced such a "charismatic politician" as Pritam Singh who leads the opposition Workers' Party (Jey, 2025).

...the traditional pattern of Singaporean elections that features the PAP as its main protagonist looks set to continue, as political leanings are still not defined by ideological commitments but instead by how voters position themselves for or against the ruling party. The PAP's ultimate victory – winning a majority of the 97 electable seats in Parliament – is a basically foregone conclusion. Still, one question is whether the ruling party will be able to retain its two-thirds supermajority. The party looks very likely to do so.

Deepfake Disinformation Report: Singapore

The government-linked outlet CNA reported a single case of deepfake content during the campaign period, noting that **73 TikTok videos** contained "digitally generated or manipulated visuals of prospective candidates," potentially violating the new deepfake law. This appeared more like a compliance check for breaches than a systematic monitoring effort. Beyond this, there are no confirmed reports of deepfake electoral disinformation in Singapore, though whether this reflects an actual absence or simply a lack of independent scrutiny remains unclear.

Applying the Deepfake Risk Matrix to Relevant Actors in Singapore

Singapore is a complex study and masterclass in state legal containment and narrative engineering. Not all is lost, however; there is still room for external actors and civic ingenuity to carve out space for truth.

Actors Matrix – Singapore (2025 Election Context)

Actor Type	Key Entities	Role in Disinformation Ecosystem
State Institutions	Ministry of Communications and Information, IMDA, Ministry of Home Affairs	Enforce POFMA and FICA; issue correction notices; monitor foreign interference; shape narrative control.
Political Actors	People's Action Party (PAP), opposition parties (e.g., PSP, WP)	PAP dominates narrative via state-linked media; opposition faces constraints and risks of censorship.
Media Outlets	Mediacorp (Channel News Asia), The Straits Times, Today Online	State-influenced; high public trust but limited editorial independence; alternative voices suppressed.
Tech Platforms	Meta, TikTok, WhatsApp, YouTube	High penetration; platforms are subject to takedown orders and fines under the deepfake law.
Civil Society & Academia	NUS, SMU, RSIS, DISA, select NGOs	Limited space for dissent; some academic voices raise alarms but operate within tight boundaries.
Diaspora & Expats	Singaporean activists abroad, independent journalists, and OSINT researchers	It is increasingly important to counter state-filtered information and amplify alternative narratives.

Strategic Interventions – Singapore:

- **Invest in OSINT networks:** Support independent researchers and diaspora-led initiatives that monitor disinfo campaigns, especially those targeting opposition figures or civil society.
- **Strengthen expat-led media ecosystems:** Fund and amplify platforms run by Singaporeans abroad that offer alternative coverage and fact-checking outside state control.
- **Build regional disinfo coalitions:** Partner with Southeast Asian watchdogs to track cross-border influence operations and share threat intelligence.
- **Support digital resilience training:** Equip opposition parties, youth activists, and journalists with tools to detect deepfakes, resist manipulation, and navigate legal constraints.
- **Leverage encrypted civic channels:** Use secure messaging and decentralised platforms to distribute verified content and counter disinfo narratives without triggering censorship.

PART 5: ADVOCACY RECOMMENDATIONS

The following recommendations build directly on the evidence presented in this report and on comparative insights from case studies in Namibia, Ecuador, Singapore and Germany. These country-specific contexts illustrate how deepfake and synthetic media disinformation affect the information space differently and depending on the strength of democratic institutions, media ecosystems, and the capacity of civil society. In Namibia, gendered cheapfakes exposed weaknesses in local-language moderation; in Ecuador, AI-generated narratives intersected with violence and criminal networks; in Singapore, tight regulation protected electoral processes but risked restricting free expression; and in Germany, strong regulation and independent fact-checking partnerships proved effective in maintaining public trust.

Together, these case studies demonstrate that the threat of deepfake disinformation is not defined solely by technology but also by the social and institutional systems surrounding it. The recommendations that follow take these lessons to form a set of global advocacy priorities grounded in transparency, equity, collaboration and human rights—principles essential for building democratic resilience in the age of generative AI.

The recommendations are structured around five interconnected action pillars designed to foster proactive, rather than reactive, approaches to information integrity.

3. **Platform Accountability:** Transparency and shared responsibility. Platforms must be transparent about how content circulates and empower independent oversight to ensure fairness across all regions and languages.

4. **AI Ethics and Human Impact Assessments:** Accountability through ethics and inclusion. AI companies must be held to measurable ethical standards that uphold human rights, safety, and inclusivity, especially the inclusion of voices from the Global South.

5. **Regulatory Reform:** Protect rights while preventing harm. Regulatory approaches must safeguard free expression, creativity, and press freedom as well as mitigate the malicious use of synthetic media.

6. **Public Resilience and Literacy:** Empowerment through knowledge. Public resilience is democracy's first line of defence. Strengthening citizens' ability to identify and critically assess synthetic content strengthens immunity against manipulation

7. **Leverage AI to Strengthen Civil Society:** Use technology responsibly for transparency and inclusion. AI should be a tool for civic empowerment, not only a source of risk.

5.1 Platform Accountability:

Recommendation 1: Ensure Global Transparency and Accountability of Social Media Platforms

Rationale:
Social media platforms are the primary vectors for deepfake disinformation. In more than **80%** of countries that held elections in 2024–2025, GenAI content increased and was amplified by opaque algorithms. This uneven moderation, particularly in non-English contexts such as Ecuador and Namibia, undermines electoral integrity and fuels polarisation.

Action:
Establish a CSO-led global campaign, in partnership with multilateral bodies, to advocate for open data access, decentralised moderation, and consistent election safeguards across all countries.

Key Priorities:

- Open data access: Require platforms to provide affordable, research-grade datasets for independent monitoring of algorithmic

performance and content moderation.

- Decentralised moderation: Adopt geofenced systems with local-language reviewers to capture cultural nuance and improve accuracy.
- Election parity: Guarantee that every national election benefits from a dedicated election-integrity team and real-time coordination with civil society.
- After each election, platforms should publish datasets on detections of synthetic media, related enforcement actions, and content moderation timelines to inform evidence-based reform.

Recommendation 2: Guarantee Equitable Data Access and Localised Moderation Standards

Rationale:
Without consistent data access and multilingual coverage, civil society cannot hold platforms accountable. Equity in data and content moderation is essential for fair and transparent digital governance.

Action:
Embed equitable access and localisation requirements in global and national policy frameworks to ensure all regions, particularly the Global South, benefit from the same transparency and protection standards.

Key Priorities:

- Enforce mandatory, affordable data access for CSOs and researchers.
- Mandate localised, language-specific moderation to close linguistic gaps.
- Require post-election transparency reports from platforms be made public in every country.

Recommendation 3: Build a Global Coalition for Platform Reform and Capacity Strengthening

Rationale:
Sustained reform depends on coordinated pressure and shared expertise. Cross-sector coalitions can combine advocacy, monitoring, and capacity-building to drive systemic change.

Action:
Launch a multisectoral "Open Data for Democracy" campaign that brings together civil society, governments, researchers and international organisations to advocate for transparency and accountability in platform governance.

Key Priorities:

- Coalition building: Partner with relevant organisations to ensure Global South leadership.
- Public advocacy: Combine public mobilisation with policy engagement at UN and regional forums, supported by expert webinars and joint reports.
- Decentralised moderation advocacy: Promote in-country moderation teams fluent in local languages; coordinate open letters setting measurable benchmarks; link executive incentives to election-integrity goals; and publish annual public scorecards grading platforms by coverage, speed, and transparency.
- Capacity building: Train CSOs in OSINT and social-media forensics and develop a global dashboard to track platform compliance.

Recommendation 4: Institutionalise International Standards and Monitoring Mechanisms

Rationale:
Long-term accountability requires global norms and regular, transparent performance evaluation. Standard-setting must ensure fairness, transparency, and human rights compliance across digital ecosystems.

Action:
Codify international standards on algorithmic transparency, equitable moderation, and election safeguards through UN and regional bodies, supported by continuous CSO monitoring and independent audits.
Key Priorities:

- Short term: Build coalitions, launch the campaign, and secure pilot data-sharing

agreements with at least two major platforms, including one Global South election.

- Medium term: Embed decentralised moderation and transparency requirements in platform policies and regulatory frameworks.
- Long term: Adopt international norms on data access, algorithmic transparency, and election protection through UN and regional instruments.
- Metrics: Track the number of platforms offering free research-grade application programming interfaces (APIs), post-election datasets, and measurable reductions in language-disparity complaints verified through annual CSO and user surveys.

5.2 AI Ethics and Human Impact Assessments

Demanding ethical standards from tech companies moves governance from reactive to proactive ethics, embedding accountability throughout the AI lifecycle. By coupling civil society advocacy with international cooperation, this approach balances innovation with harm prevention and ensures that the evolution of GenAI aligns with democratic principles and human rights values.

Recommendation 1: Mandate Ethical Standards and Accountability Across the AI Lifecycle

Rationale:
Companies such as OpenAI, Google, and xAI have introduced powerful tools without consistent ethical safeguards. Ethical principles are often reactive and impact assessments typically occur after deployment, when harms such as gendered disinformation or non-consensual imagery have already spread.

Action:
Establish binding ethical requirements across the entire AI lifecycle. This includes mandatory human rights and psychosocial impact assessments (HIAs) at the design, development, and deployment stages, as well as employ continuous post-release monitoring.

Key Priorities:

- Ethics by design: Shift from reactive governance to prevention through early-stage ethical review.
- Lifecycle assessments: Ensure HIAs evaluate not only technical risks but also social and cognitive impacts, including susceptibility to manipulation.
- International alignment: Coordinate with global frameworks such as UNESCO's Recommendation on the Ethics of AI to harmonise standards.

Recommendation 2: Establish Independent Ethics Boards within Technology Companies

Rationale:
Corporate self-regulation has proven inadequate. Ethical oversight must be independent, transparent, and inclusive. Lessons from Meta's Oversight Board show that legitimacy depends on autonomy, diversity, and binding authority. A Global AI Ethics Coalition, including academic and civil society partners, could develop common standards for these boards. The Coalition would define prohibited uses (e.g., electoral manipulation or targeted suppression), establish risk registers and release notes, and oversee regular, regionally disaggregated audits.

Action:
Require major AI firms to establish independent ethics boards with representation from civil society and affected communities. These boards should review high-risk AI systems—especially election-related tools such as voice cloning, targeting systems, and image generation—and publish their findings in the public domain.

Key Priorities:

- Boards must include journalists, human rights lawyers, technical experts and affected communities.
- Clear mandates should require "comply or explain" responses to recommendations.

- Public case summaries and rationales must be published regularly to ensure transparency.

Recommendation 3: Embed Global Ethical Standards in AI Governance Frameworks

Rationale:
Without shared norms, AI governance risks reinforcing Global North dominance and deepening inequities. Global coordination is essential to ensure that ethical standards are inclusive, enforceable, and locally relevant.

Action:
Promote the adoption of international norms through a UN-led AI Ethics Pact that embeds ethical principles into global governance frameworks and national regulations. This will institutionalise standards and make ethics a requirement for access to public or multilateral funding.

Key Priorities:

- Establish clear definitions of harm and accountability mechanisms that prioritise prevention over profit.
- Ensure representation from Global South experts and gender-diverse voices in all norm-setting processes.
- Link compliance with access to public procurement or partnership eligibility.

Recommendation 4: Strengthen Monitoring and Accountability Mechanisms

Rationale:
Ethical standards are only practical when compliance is independently verified. Regular reporting, third-party audits, and public transparency are crucial to maintaining credibility and trust.

Action:
Require all major AI companies to publish annual ethics reports, independently audited and verified by civil society. CSOs should conduct shadow reporting to provide independent performance assessments and expose gaps.

Key Priorities:

- An increased number of companies with functioning ethics boards.
- Evidence of HIA integration across AI product releases.
- Reduction in reported harms and bias incidents (tracked via IPIE and other observatories).

5.3 Building Public Resilience

Building public resilience transforms individuals into informed participants rather than passive recipients of information, creating a cumulative, society-wide shield against the evolving threat of deepfake disinformation.

Recommendation 1: Strengthen Digital Literacy and Critical Thinking

Rationale:
Deepfakes exploit universal cognitive biases and cultural specific norms of trust and sharing. The 2024 evidence shows that awareness and preparation helped limit their impact, yet vulnerabilities persist, especially among excluded groups and in information-saturated environments. Empowering citizens with critical thinking and media literacy skills is the most durable defence against manipulation.

Action:
Design pre-emptive education and awareness campaigns that teach citizens how to identify synthetic media, question sources, and verify information before sharing.

Key Priorities:

- Develop learning materials that explain typical manipulation techniques (voice cloning, spliced video, fabricated context) and simple verification steps: pause, source, date, trace.
- Produce resources in local languages and adapt to communication habits: radio and WhatsApp in rural areas, short videos in urban centres, posters in low-connectivity

zones.

- Partner with teachers' unions, community centres, public broadcasters and newsrooms to co-create workshops and public-service content for schools and youth programmes.

Recommendation 2: Build Collaborative Networks for Trusted Communication

Rationale:
People are more likely to accept corrections and prebunks from sources they already trust. Sustained collaboration between credible local actors strengthens collective immunity against disinformation.

Action:
Establish distributed networks of trusted messengers, such as community radio hosts, diaspora leaders, editors' forums, election authorities and digital rights groups, to issue rapid, credible corrections and prebunks.

Key Priorities:

- Formalise partnerships with clear escalation channels to platforms and election bodies.
- Provide partners with copy-ready prebunk scripts and adaptable broadcast formats.
- Encourage two-way feedback loops so that local observations feed into national and global early-warning systems.

Recommendation 3: Integrate Resilience into Education and Civic Infrastructure

Rationale:
One-off campaigns have a limited effect. Embedding resilience in education systems and civic institutions ensures continuity and long-term impact.

Action:
Mainstream digital-resilience curricula in national education policies and link them to broader civic-engagement initiatives.

Key Priorities:

- Incorporate media-literacy modules into school syllabi and teacher-training programmes.
- Provide sustained funding for public-awareness units within election-related bodies and information-integrity CSOs.
- Align with UNESCO's Media and Information Literacy framework to harmonise standards globally.

Implementation Timeline and Metrics:

- Short term: Launch pilot campaigns in selected countries.
- Medium term: Scale programmes regionally and embed in national curricula.
- Long term: Codify media literacy and resilience standards into global education and governance frameworks.
- Metrics: Measured increases in public awareness and critical-thinking indicators; reduced misinformation-sharing rates in behavioural studies; and number and diversity of partnerships formed.

5.4 Leveraging AI for Civil Society

While much of the policy debate on AI focuses on its risks to democracy, the same technologies hold vast potential to strengthen civil society, expand participation, and enhance resilience.

As the report highlights, AI can be harnessed for social good: from detecting disinformation to improving accessibility, translating civic content across languages, and deepening public engagement. Harnessing these benefits requires proactive leadership from civil society itself, ensuring that AI development and deployment align with human rights, transparency, and inclusion. By investing in ethical adoption, capacity building, and governance frameworks, AI can become a tool that amplifies civic voices, accelerates accountability, and helps safeguard the integrity of democratic processes worldwide.

Recommendation 1: Harness AI for Social Good and Civil Society Capacity

Rationale:
While much of the discourse around AI focuses on its risks, the same technologies can significantly strengthen civil society's reach, responsiveness, and resilience. AI can expand access to information, automate monitoring of disinformation, translate content across languages, and help identify emerging social or political risks. When guided by ethical principles and human rights, AI becomes a tool for empowerment rather than manipulation.

Action:
Encourage and resource civil society to adopt AI tools responsibly across key domains, while ensuring transparency, inclusion, and data protection.

Key Priorities:

- Build awareness and technical literacy within CSOs to safely integrate AI tools.
- Promote open, affordable, and privacy-preserving AI applications tailored to civic needs.
- Establish ethical guidelines and accountability mechanisms to govern use

Below are some positive use cases of AI for civil society:

Application Area	AI Function / Tool	Example / Impact	Ethical & Governance Considerations
Disinformation Monitoring & Analysis	Machine learning for pattern recognition, content clustering, and synthetic media detection	Tools like Reality Defender or Deepware Scanner detect deepfakes early, enabling CSOs to flag harmful content before it spreads virally.	Ensure algorithmic transparency; avoid over-reliance on automated labelling.
Translation & Inclusion	Natural Language Processing (NLP) for real-time multilingual translation	AI-driven translation bridges linguistic gaps in countries like Namibia, improving outreach to rural and indigenous communities.	Prioritise cultural context and local dialect accuracy; address linguistic bias.
Media & Fact-Checking	AI-assisted verification; reverse image search automation	Tools like Truepic and InVID-WeVerify accelerate visual verification for journalists and CSOs.	Combine human review with AI checks; maintain privacy of image metadata.
Data Analysis for Advocacy	Predictive analytics to identify emerging social or political risks	AI-driven dashboards visualise trends in online hate speech or electoral discourse, helping with early intervention.	Require data anonymisation and informed consent for datasets.

Accessibility & Inclusion	Speech-to-text and text-to-speech tools; summarisation and adaptive interfaces	AI enhances accessibility for people with disabilities (e.g., voice interfaces for visually impaired voters).	Ensure compliance with data privacy and disability-access standards.
Crisis Response & Humanitarian Aid	AI forecasting for disaster response and spikes in misinformation	Predictive models assist humanitarian CSOs to anticipate floods or surges in misinformation.	Avoid surveillance misuse; ensure community consent in data collection.
Civic Engagement & Dialogue	Chatbots and conversational AI for public information campaigns	Civic bots provide verified election information or counter disinformation narratives in real time.	Ensure transparency so users know they are interacting with AI, not humans.

Recommendation 2: Build Ethical and Technical Capacity in the Civil Society Sector

Rationale:
For AI to be a force for good, civil society must not only have access to the tools but also understand and shape how they are built. The current asymmetry between AI developers and civic actors risks replicating existing digital inequities. By embedding ethical AI use into civil society practice, the sector can transform from a reactive actor to an innovative co-architect of the digital future, using the same technologies that spread disinformation to build trust, transparency, and resilience.

Action:
Develop regional AI Capacity Hubs for Civil Society, linking universities, technical experts, and grassroots organisations to co-design open-source civic AI tools.

Key Priorities:

- Provide training on AI ethics, data governance, and risk assessment for CSOs.
- Support the creation of shared repositories of open, local-language civic datasets.
- Partner with philanthropic donors and public institutions to fund civic-tech incubators.

Recommendation 3: Establish Governance Frameworks for Responsible AI Use

Rationale:
Civil society's adoption of AI must reflect the same accountability standards demanded of governments and corporations. Ethical guardrails ensure that civic applications of AI respect privacy, autonomy, and human dignity.

Action:
Develop a Civic AI Ethics Charter outlining clear principles for transparency, data protection, inclusivity, and public accountability.

Key Priorities:

- Mandate transparent disclosure of AI-assisted outputs and decision-making processes.
- Incorporate community feedback loops to assess impact and unintended consequences.
- Align governance standards with frameworks such as UNESCO's Recommendation on the Ethics of AI and the OECD's Principles on AI.

Implementation Timeline and Metrics:

Short term: Identify existing AI use cases and launch regional training pilots in 3–5 countries.

Medium term: Establish AI Capacity Hubs and release the Civic AI Ethics Charter.
Long term: Institutionalise AI literacy and ethical governance within civil society networks globally.

Metrics: Number of CSOs adopting AI tools ethically; documented improvements in digital inclusion; number of civic datasets or open tools developed; and independent audits confirming adherence to the ethics charter.

GLOSSARY

Algorithms
A fixed series of steps that a computer performs to solve a problem or complete a task. On social media platforms, algorithms compile and present content based on users' engagement history and predicted interests, often influencing the spread of disinformation.

Algorithmic Bias
Systematic errors in algorithms that lead to unfair or skewed outcomes, such as prioritising certain content or demographics. Disinformation, or algorithmic bias, can amplify misleading narratives and marginalise accurate information, often reinforcing stereotypes or polarising content.

Algorithmic Transparency
The degree to which the operations, criteria, and decision-making processes of algorithms, particularly recommender algorithms, are openly disclosed and understandable to users and researchers. Lack of transparency on social media platforms fuels disinformation by obscuring how content is prioritised or amplified, hindering efforts to detect manipulation or bias.

Amplification
The process of increasing the reach or visibility of content, either organically (through shares, likes, and comments) or artificially (via bots, sock puppets, or astroturfing). Amplification can also occur independently of algorithms or through coordinated efforts to manipulate platform rankings.

Artificial Intelligence (AI)
Computer systems performing tasks that typically require human intelligence, such as learning or pattern recognition. In disinformation, AI generates convincing fake content (e.g., deepfakes, text, images) or helps detect manipulation campaigns.

Automation
Software tools designed to complete tasks with minimal human direction. In disinformation, automation amplifies misleading narratives through bots or coordinated campaigns.

Cognitive Biases
Unconscious thinking patterns that influence how people interpret information. Disinformation exploits these unconscious thinking patterns to increase the likelihood that individuals will accept or share false narratives.

Content Moderation
The process of detecting and addressing content that violates the terms of use utilises both automation and human review. Actions include demonetisation, downgrading or removal. Disinformation persists due to inconsistent moderation, particularly in non-English languages.

Content Removal
A moderation decision to delete content violating a platform's Terms of Service. Enforcement varies across languages and regions, raising concerns about transparency and consistency.

Coordinated Inauthentic Behaviour (CIB)
Networks of accounts secretly work together to sway online narratives by employing strategies such as identical posts and coordinated timing., which is central to many influence operations.

Data Access
The ability to retrieve digital information from platforms, often via APIs or scraping, is crucial for disinformation research. Platform policies are increasingly restricting this tool for disinformation research.

Data Mining
The process of discovering patterns in large social media datasets to detect coordinated behaviours, influential accounts or the spread of narratives that allows data collection (e.g., scraping) to transform raw data into insights.

Debunking
Exposing and correcting false claims through fact-checking, investigations or exposés, and which aims to counter disinformation and misinformation.

Deepfakes
Synthetic multimedia content that convincingly mimics real people or events, typically created to deceive. It is now increasingly produced using accessible, user-friendly Generative AI platforms such as ChatGPT, Grok, and others. Deepfakes enable malicious actors to craft realistic fake videos, audio, or images for disinformation campaigns, such as impersonating public figures or spreading false narratives.

Deep Learning
A subfield of AI that enables computers to learn from data and improve performance without being explicitly programmed. Machine learning systems identify patterns in existing data and apply that knowledge to make predictions, classifications or generate outputs when exposed to new information.

Digital Democracy
The use of digital technologies to support democratic participation, transparency, and accountability. It encompasses online civic engagement, access to information, and the ability to express political views in safe and inclusive digital spaces without fear of retribution. Digital democracy depends on the integrity of the online information environment. When polluted by disinformation, misinformation, hate speech or manipulation, these spaces can become tools for exclusion, polarisation, democratic erosion and undermine electoral integrity.

Digital Literacy
This is the capacity to evaluate and interact with digital content critically, identify information disorders, and safeguard online privacy.

Digital Resilience
The capacity of individuals or societies to withstand and adapt to digital threats, including disinformation, surveillance, censorship and others.

Digital Rights
Online freedoms and protections include privacy, freedom of expression, and access to information. It encompasses safeguards against censorship, surveillance, and other online harms.

Disinformation
False information deliberately created or spread to cause harm, often for political, financial or social motives.

Electoral Disinformation
False or misleading information deliberately spread to influence elections, undermine electoral integrity or manipulate voter behaviour, thereby threatening digital democracy. It includes tactics like voter suppression, disinformation, election denialism, microtargeted disinformation, fake news, deepfakes or narrative hijacking to sow distrust, polarise voters or discredit candidates.

Electoral Integrity
The degree to which electoral processes are free, fair, and credible, as well as supported by transparent systems, impartial institutions, and an informed electorate. In digital spaces, electoral integrity relies on the integrity of digital democracy in the online information environment. Voters must be able to access accurate, trustworthy information without being misled by disinformation, misinformation, hate speech or foreign and domestic manipulation. A compromised digital environment can distort public perception, suppress participation, and undermine trust in electoral outcomes.

Fact-Checking
Verifying the accuracy of public statements or reports.

Fake News
Disinformation formatted to resemble authentic news, often as falsified articles or websites.

Foreign Information Manipulation and Interference (FIMI)
The deliberate actions undertaken by a foreign government or entity to exert influence over another country's decision-making, policies or public opinion, often in ways that benefit the foreign actor's interests. These operations can employ covert and overt methods, including

disinformation campaigns, cyberattacks, and financial inducements, and are often designed to undermine democratic institutions, manipulate public discourse or advance a foreign government's strategic objectives. (also known as Foreign Influence Operations).

Freedom of Speech
The right to express opinions without censorship or penalty. It can be misused to resist content moderation and disinformation can be used to suppress targeted groups' freedom of speech.

Gendered Disinformation
False or misleading content specifically designed to target women and gender non-conforming individuals, particularly those in public, political or activist roles. It blends traditional disinformation tactics with gender-based abuse to silence, discredit or intimidate its targets. Typical techniques include the spread of misogynistic narratives; the sexualisation and manipulation of images or videos (including deepfakes); the reinforcement of harmful gender stereotypes; and coordinated harassment campaigns involving threats of violence, doxxing or cyberattacks. The goal is not only reputational harm but also to deter political participation and limit visibility in public discourse.

Generative AI
AI that creates new content (e.g., text, images, audio) used in disinformation to produce deepfakes or tailored propaganda. Examples include ChatGPT, DALL-E and Grok.

Hate Speech
Communication attacking individuals or groups based on protected attributes (e.g., race, gender). It often overlaps with disinformation to incite violence or silence voices.

Inferred Identities
A set of personal or social characteristics, such as race, gender, religion, location or political affiliation, deduced by algorithms or platforms based on a user's online behaviour, interactions, and data patterns rather than voluntarily disclosed information. These algorithmically derived profiles are often used in microtargeting, including disinformation campaigns, to craft tailored content that exploits perceived vulnerabilities or group affiliations.

Information Disorders
An umbrella term for various forms of harmful or misleading content that distort truth and undermine public discourse. It includes disinformation, misinformation, malinformation, propaganda, conspiracy theories, clickbait, satire or parody shared as fact, hoaxes, trolling, imposter content, synthetic media and others.

Information Integrity
An ecosystem where accurate, reliable information is consistent and accessible and where freedom of expression is protected.

Information Overload
A state where excessive information volume overwhelms critical processing, enabling disinformation to spread unnoticed, especially during crises.

Information Vacuum
A lack of timely, accurate information, allowing disinformation, rumours or conspiracies to fill the gap, a common occurrence during crises or elections.

Influence Operations
Coordinated efforts to manipulate public opinion or behaviour using deceptive tactics such as disinformation or fake accounts.

Influencers
Individuals with large social media followings who shape opinions or behaviour.

Influencer-for-Hire
An influencer paid to amplify specific narratives or discredit opponents, often covertly.

Inoculation Theory
The strategy, grounded in behavioural science, aims to foster psychological resilience against misinformation by presenting individuals with a toned-down version of deceptive content, accompanied by refutations or

counterarguments. This "prebunking" approach equips individuals to recognise and reject future attempts at manipulation.

Internet Shutdowns
Intentional interruptions of internet connectivity intended to regulate the flow of information, often intensifying misinformation by restricting access to accurate information.

Labelling
Content moderation practice of applying informational labels to posts or accounts to provide context (e.g., marking disinformation or sensitive content).

Large Language Models (LLMs)
AI trained on vast text data to generate human-like language. Examples include ChatGPT, Grok, and LLaMA.

Linguistic Disparity in Moderation
The inconsistent or inadequate moderation of harmful content, such as disinformation or hate speech, in non-English languages due to limited automated systems or human oversight. Malicious actors exploit this gap by using native languages or word camouflage to bypass detection, particularly in electoral disinformation campaigns targeting diverse linguistic communities.

Malign Actors
Individuals, groups or commercial entities intentionally spreading disinformation or manipulating information ecosystems.

Malinformation
Truthful information shared to cause harm, often by revealing private data or using facts out of context (e.g., doxxing).

Machine Learning (ML)
A subfield of artificial AI that enables computers to learn from data and improve performance without being explicitly programmed. Machine learning systems identify patterns in existing data and apply that knowledge to make predictions, classifications or generate outputs when exposed to new information.

Manufactured Amplification
Deliberate boosting of content visibility through deceptive means (e.g., bots, sockpuppets) to distort perceived popularity or credibility.

Media Literacy
Competencies to critically engage with media, as well as assess source credibility and truthfulness.

Microtargeting
The practice of sending highly tailored content or ads to small, specific groups based on personal characteristics and beliefs.

Misinformation
False information spread without intent to mislead and is believed to be true by sharers.

Natural Language Processing (NLP)
A field of AI that enables computers to understand, interpret, and generate human language. In disinformation research, NLP is used to analyse large volumes of social media content, detect harmful narratives, identify emotional tones and automate the recognition of patterns of misinformation and disinformation.

Online Violent Extremism
Using digital platforms to promote or incite ideologically motivated violence, often through echo chambers and algorithmic reinforcement.

Open-Source Intelligence (OSINT)
The practice of collecting, analysing, and interpreting publicly available information from digital, print, and broadcast sources to generate actionable insights. In the context of disinformation, OSINT leverages social media platforms, news outlets, websites, forums and multimedia content to detect coordinated manipulation, trace the origins of false narratives, and identify threat actors. OSINT is a foundational method in digital investigations, election monitoring, and media forensics, valued for its transparency, verifiability, and ethical alignment when conducted responsibly.

Prebunking
Anticipating and countering disinformation before it spreads, using past fact-checks to prepare responses.

Propaganda
A form of strategic political communication aimed at influencing public opinion or behaviour in support of a political, ideological or institutional agenda. Propaganda typically involves the selective use of facts, emotional appeals, repetition and symbolic messaging to persuade and mobilise audiences. While not always false or harmful, propaganda can become problematic when it distorts reality, suppresses dissent, or legitimises authoritarianism. In electoral contexts, it is distinct from disinformation, although the two may intersect.

Psychographic Profiling Data
Information that categorises individuals based on psychological attributes such as values, beliefs, interests, attitudes, lifestyles and personality traits. This data is often derived from online behaviour, including social media activity, likes, shares, and browsing habits, and is used to predict and influence decision-making, especially in targeted advertising and political microtargeting campaigns.

Recommender Algorithm
An automated system used by social media platforms to select, rank, and present content based on user behaviour, interests, and engagement signals. Powered by machine learning, this algorithm prioritises attention-grabbing content, often amplifying disinformation by creating feedback loops that reinforce cognitive biases, such as confirmation bias, and entrench user beliefs.

Regulatory Responses
Government or institutional policies and laws aimed at combating disinformation, hate speech, or platform manipulation. Examples include content moderation mandates, transparency requirements for algorithms, or penalties for spreading false information. These efforts aim to enhance information integrity but face challenges in enforcement and striking a balance with free speech.

Social Media Data
Publicly available content and metadata from platforms are used to detect disinformation trends or campaigns.

Social Media Digital Forensics
The specialised process of collecting, preserving, and analysing social media data to uncover evidence of harmful activities, such as disinformation, cyberbullying or hate speech, often perpetrated by anonymous accounts. Techniques include metadata analysis, linguistic profiling, network mapping and reverse image searches to trace origins, identify hidden networks or attribute content to malicious actors despite anonymity. This field is critical for exposing coordinated manipulation and ensuring admissible evidence for legal or public accountability.

Social Media Metrics
The analysis of social media data to provide a quantitative measurement of a topic; for example, analysing the conversation volume on a specific topic and comparing that against other topics.

Social Media Monitoring
The real-time tracking and recording of social media activity, such as mentions, hashtags, or keywords, to observe engagement, flag incidents, and identify disinformation as it spreads.

Social Media Listening
The process of tracking and analysing online conversations to understand public sentiment, detect emerging trends, and uncover disinformation patterns. Unlike social media monitoring, which focuses on observing and recording activity, social listening interprets meaning and context.

Synthetic Media
AI-generated or manipulated content used to create convincing disinformation.

Targeted Harassment
Coordinated online attacks to threaten or silence individuals, often overlapping with disinformation or hate speech.

Technology-Facilitated Gender-Based Violence (TFGBV)
Gender-based harm via digital platforms, including harassment, doxxing or gendered disinformation targeting women or gender-diverse individuals.

Trolling
Inflammatory online behaviour to provoke negative reactions, often used in disinformation to distract or polarise.

Troll Farm
A group engaging in coordinated trolling or bot-like narrative promotion, also called a troll army.

User-Generated Content (UGC)
Any form of content created and voluntarily shared by individual users on digital platforms, rather than by the platforms themselves, professional media or paid content producers. It stands in contrast to coordinated content produced by content farms, bot farms or commercial disinformation operators.

Web Scraping
Extracting data from websites without APIs; used in disinformation research but may violate platform terms of service.

BIBLIOGRAPHY

1. Afrobarometer. (2024). Namibians Express Confidence In Elections, But Weakening Trust In Electoral Commission. https://www.afrobarometer.org/publication/ad876-namibians-express-confidence-in-elections-but-weakening-trust-in-electoral-commission/
2. Alanazi, S.,& Asif, S. (2024). Exploring Deepfake Technology: Creation, Consequences And Countermeasures. Hum.-Intell. Syst. Integr. 6, 49–60 https://doi.org/10.1007/s42454-024-00054-8
3. Alanazi, S., Asif, S., Caird-daley, A., & Moulitsas, I. (2025). Unmasking Deepfakes: A Multidisciplinary Examination Of Social Impacts And Regulatory Responses. Hum.-Intell. Syst. Integr. (2025). https://doi.org/10.1007/s42454-025-00060-4
4. Boscán, A. (2025, April 7). Violence shadows Ecuador's presidential election. Think Global Health. https://www.thinkglobalhealth.org/article/violence-shadows-ecuadors-presidential-election
5. Chesney, R., & Citron, D. K. (2019). Deep Fakes: A Looming Challenge For Privacy, Democracy, And National Security. California Law Review, 107(6), 1753–1819. https://dx.doi.org/10.2139/ssrn.3213954
6. Diya, S. R. (2024). Political cheap fakes are a blind spot for platforms in the Global South. Tech Global Institute. https://www.context.news/ai/opinion/cheap-fakes-are-a-blind-spot-for-platforms-in-the-global-south
7. Freedom House. (2023) The Repressive Power Of Artificial Intelligence. https://freedomhouse.org/report/freedom-net/2023/repressive-power-artificial-intelligence#generative-ai-supercharges-disinformation
8. Freedom House. (2024). Freedom on the Net: 2024. https://freedomhouse.org/report/freedom-net
9. Freedom House. (2025). Freedom in the World: 2025. https://freedomhouse.org/report/freedom-world/2025
10. Gallup. (2025, February). Germany's Election: 5 Key Issues Facing The Next Chancellor. . https://news.gallup.com/poll/656579/germany-election-key-issues-facing-next-chancellor.aspx
11. Hagendorff, T. (2024). Mapping the Ethics of Generative AI: A Comprehensive Scoping Review. Minds & Machines 34, 39. https://doi.org/10.1007/s11023-024-09694-w
12. Hameleers, M. (2024). Cheap Versus Deep Manipulation: The Effects Of Cheapfakes Versus Deepfakes In A Political Setting. International Journal of Public Opinion Research, 36(1), Article edae004. https://doi.org/10.1093/ijpor/edae004
13. International Panel on the Information Environment (IPIE). (2025). The Role Of Generative AI Use In 2024 Elections Worldwide. https://www.ipie.info/research/tp2025-2
14. Ipsos. (2023). Global Views On AI And Disinformation. https://www.ipsos.com/en-nz/global-views-ai-and-disinformation
15. Kahn, J. (2024). How Generative AI Is Helping Fact-Checkers Flag Election Disinformation, But Is Less Useful In The Global South. Reuters Institute. https://gijn.org/stories/how-generative-ai-helps-fact-checkers/
16. Kapoor, S., & Narayanan, A. (2024). We Looked At 78 Election Deepfakes. Political Misinformation Is Not An AI Problem. Knight First Amendment Institute at Columbia University. https://knightcolumbia.org/blog/we-looked-at-78-election-deepfakes-political-misinformation-is-not-an-ai-problem
17. Kuo, R., & Marwick, A. E. (2021). Critical Disinformation Studies: History, Power, And Politics. Harvard Kennedy School Misinformation Review, 2(6). https://misinforeview.hks.harvard.edu/article/critical-disinformation-studies-history-power-and-politics/
18. Madrid-Morales, D., Wasserman, H., Gondwe, G., Ndlovu, K., Sikanku, E., Tully, M., Umejei, E., & Uzuegbunam, C. (2021). Comparative Approaches to Mis/Disinformation|

Motivations for Sharing Misinformation: A Comparative Study in Six Sub-Saharan African Countries. International Journal of Communication, 15, 20. https://ijoc.org/index.php/ijoc/article/view/14801

19. McKay, C., & Trauthig, I. (2025). Then And Now: How Does AI Electoral Interference Compare In 2025?. Centre for International Governance Innovation. https://www.cigionline.org/articles/then-and-now-how-does-ai-electoral-interference-compare-in-2025/
20. Pataranutaporn, P., Lee, T. H., Lee, J., Kim, J., Wong, A., & Maes, P. (2025). Synthetic Human Memories: AI-Edited Images and Videos Can Implant False Memories and Distort Recollection. In Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems (CHI '25). Association for Computing Machinery, New York, NY, USA, Article 538, 1–20. https://doi.org/10.1145/3706598.3713697
21. Quah, S. J. (2025). Singapore Election: All About The Lived Economy. The Lowy Institute. https://www.lowyinstitute.org/the-interpreter/singapore-election-all-about-lived-economy
22. Regan, J. (2024, October 2). A Brief History Of Deepfakes. Reality Defender. https://www.realitydefender.com/insights/history-of-deepfakes
23. Rebelo, J. (2024, May 16). India's Generative AI Election Pilot Shows Artificial Intelligence In Campaigns Is Here To Stay. University of Texas at Austin Center for Media Engagement. https://mediaengagement.org/wp-content/uploads/2024/10/Indias-Generative-AI-Election-Pilot-Shows-Artificial-Intelligence-In-Campaigns-Is-Here-To-Stay.pdf
24. Schiller, V. & Harbath, K. (2025, January 17). The first A.I. elections: The dog that didn't bark? (Not exactly). Aspen Digital. https://www.aspendigital.org/blog/first-ai-elections/
25. Schneier, B., & Sanders, N. (2024). The apocalypse that wasn't: AI was everywhere in 2024's elections, but deepfakes and misinformation were only part of the picture. Harvard Kennedy School Ash Centre for Democratic Governance and Innovation. https://ash.harvard.edu/articles/the-apocalypse-that-wasnt-ai-was-everywhere-in-2024s-elections-but-deepfakes-and-misinformation-were-only-part-of-the-picture/
26. Seger, E., Ovadya, A., Siddarth, D., Garfinkel, B., & Dafoe, A. (2023). Democratising AI: Multiple meanings, goals, and methods. In Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society (AIES '23) (pp. 715–722). Association for Computing Machinery. https://doi.org/10.1145/3600211.3604693
27. Shelton, J. (2025, February). Germany: Nearly 90% Of Voters Fear Manipulation. DW. https://www.dw.com/en/germany-nearly-90-of-voters-fear-foreign-manipulation/a-71528481
28. Simchon, H., Wittenberg, C., Finkel, E., & Druckman, J. N. (2024). The Persuasive Effects Of Political Microtargeting In The Age Of Generative Artificial Intelligence. PNAS Nexus, 3(2). https://doi.org/10.1093/pnasnexus/pgae035
29. Simon, F. M., Altay, S., & Mercier, H. (2023). Misinformation Reloaded? Fears About The Impact Of Generative AI On Misinformation Are Overblown. Harvard Kennedy School Misinformation Review, 4(4). https://doi.org/10.37016/mr-2020-127
30. Simon, F. M., & Altay, S. (2025, July 7). Don't Panic (Yet): Assessing The Evidence And Discourse Around Generative AI And Elections. Knight First Amendment Institute at Columbia University. https://knightcolumbia.org/content/dont-panic-yet-assessing-the-evidence-and-discourse-around-generative-ai-and-elections
31. Singh, S., & Dhumane, A. (2025). Unmasking digital deceptions: An integrative review of deepfake detection, multimedia forensics, and cybersecurity challenges. MethodsX, 15, 103632. https://doi.org/10.1016/j.mex.2025.103632

32. Tai, Y. C., Patni, K. N., Hemauer, N., Desmarais, B., & Lin, Y. R. (2025). GenAI Vs. Human Fact-Checkers: Accurate Ratings, Flawed Rationales. In Proceedings of the 17th ACM Web Science Conference (WebSci '25) (pp. 516–521). ACM. https://doi.org/10.1145/3717867.3717896

33. Walker, C. P., Schiff, D. S., & Schiff, K. J. (2024). Merging AI Incidents Research with Political Misinformation Research: Introducing the Political Deepfakes Incidents Database. Proceedings of the AAAI Conference on Artificial Intelligence, 38(21), 23053-23058. https://doi.org/10.1609/aaai.v38i21.30349

34. Wardle, C., & Derakhshan, H. (2017). Information Disorder: Toward An Interdisciplinary Framework For Research And Policymaking. Council of Europe. https://edoc.coe.int/en/media/7495-information-disorder-toward-an-interdisciplinary-framework-for-research-and-policy-making.html

35. Weikmann, T., Greber, T., & Nikolaou, A. (2024). After Deception: How Falling For A Deepfake Affects The Way We See, Hear, And Experience Media. The International Journal of Press/Politics, 29(3), 450–469. https://doi.org/10.1177/19401612241233539

36. Winner, L. (1980). Do Artifacts Have Politics? Daedalus, 109(1), 121–136. http://www.jstor.org/stable/20024652

37. Yilmaz, I., Morieson, N., & Demir, M. (2022). Religious Populism, Cyberspace And Digital Authoritarianism In Asia. European Centre for Populism Studies. https://www.populismstudies.org/religious-populism-cyberspace-and-digital-authoritarianism-in-asia-the-cases-of-india-indonesia-pakistan-malaysia-turkey/

38. Zhou, X., & Shen, C. (2024). Processing Of Misinformation As Motivational And Cognitive Biases. Frontiers in Psychology, 15, 13321. https://pmc.ncbi.nlm.nih.gov/articles/PMC11393549/

Digital Democracy Initiative

The Digital Democracy Initiative (DDI) is a programme to safeguard inclusive democracy and human rights in the digital age. It focuses on supporting local civil society in the Global South, particularly in countries undergoing democratic regression and where civic space is under pressure.

For more information visit:
digitaldemocracyinitiative.net

CIVICUS

CIVICUS is a global alliance of civil society organisations and activists working to strengthen citizen action and civil society throughout the world.

For more information visit:
civicus.org