Article

The International Journal of Press/Politics
1–20

DOI: 10.1177/19401612251395434
journals.sagepub.com/home/hij
Sage

AI-Driven Disinformation and Political Influence on WhatsApp in South Africa's 2024 Elections

Gregory Gondwe[1,2]

Abstract

This study investigates the role of encrypted messaging apps, specifically WhatsApp, in disseminating political disinformation during South Africa's 2024 general elections. Drawing on a dataset of 22,384 messages from 47 politically active WhatsApp groups and 6,283 unique users, the study identifies how AI-generated deepfakes, emotional manipulation, and ideological group structures facilitated the spread of false political narratives. Using MyFactChecker, an AI-powered sentiment and verification tool, we reveal that disinformation most frequently relied on fear-based appeals (41%), identity-driven rhetoric (32%), and content mimicking credible journalism (27%). The findings show that disinformation gained traction not through factual accuracy but through emotional resonance and relational trust within ideologically cohesive groups. Even when flagged as false, corrective content was often dismissed as propaganda or foreign interference, thus underscoring how truth contests unfold within closed, affective networks.

Keywords

encrypted messaging apps, disinformation, AI-generated deepfakes, electoral misinformation, WhatsApp politics

Introduction

In recent years, encrypted messaging apps (EMAs), such as WhatsApp have become powerful symbols for political communication and influence, especially across the

[1]California State University, San Bernardino, CA, USA
[2]Harvard Faculty Associate, Berkman Klein Center for Internet & Society, Cambridge, MA, USA

Corresponding Author:
Gregory Gondwe, California State University, 5500 University Pkwy, San Bernardino, CA 92407, USA.
Emails: Gregory.gondwe@csusb.edu; ggondwe@cyber.harvard.edu

Global South. Unlike public platforms such as Instagram, Facebook, or X (formerly Twitter), EMAs like WhatsApp offer end-to-end encryption, while shielding messages from public scrutiny and making these channels difficult to monitor for researchers, regulators, and fact-checkers (Donovan and Boyd 2021). While this enhances privacy for users, it also creates fertile ground for the unchecked spread of disinformation, particularly in societies where electoral trust is fragile.

In recent years, EMAs have become central to digital disinformation campaigns. Their private, trusted networks of family, friends, and political supporters amplify the credibility of false information (Donovan and Boyd 2021). This pattern has become especially pronounced during elections across the Global South. For example, in Brazil's 2018 presidential elections, misinformation shared via WhatsApp significantly influenced electoral outcomes (Nemer and Marks 2024). Tardáguila et al. (2018) found that, among a set of widely shared political images analyzed from public WhatsApp groups during the Brazilian election, more than half were classified as misleading or false. Despite interventions such as collaborative fact-checking projects like Comprova, the encrypted nature of these platforms has limited the ability of watchdogs to trace or contain disinformation.

Although these developments have drawn increasing attention in Latin America, the role of EMAs in electoral disinformation within African contexts remains critically underexplored. Yet, there is growing evidence that the mechanisms of disinformation on EMAs are dynamic and adaptable. In Zambia, WhatsApp has served both as a platform for grassroots mobilization and as a channel for state-aligned disinformation campaigns (Gondwe 2024a). Kenya's 2022 general elections saw the strategic use of manipulated audio clips and deepfakes shared via WhatsApp to smear political candidates (Nyabola 2023). These cases point to a global trend: EMAs operate in a regulatory vacuum, where encryption obscures the origin, intention, and velocity of content dissemination (Munger et al. 2024; Rossini 2023). As EMAs become increasingly central to political discourse in regions with limited digital oversight, they also become more attractive for covert influence operations (Resende et al. 2019).

South Africa presents a particularly important context for advancing theory on digital disinformation and political influence. The country's recent political history is marked by high levels of party competition, persistent economic and racial inequality, and ongoing debates over democratic legitimacy. Its post-apartheid landscape features both vibrant multiparty politics and deep social divisions, making it an ideal setting for examining how disinformation interacts with trust, identity, and mobilization in complex democracies. The widespread use of WhatsApp, combined with contested narratives and historical legacies of media distrust, allows researchers to explore not only the transmission of disinformation but also its entanglement with local struggles over authority, belonging, and collective memory.

The 2024 general elections in South Africa marked a watershed moment for misinformation research on the continent. For the first time, AI-generated deepfakes became a major feature of electoral disinformation on WhatsApp. Viral videos falsely depicting U.S. President Joe Biden threatening sanctions, Donald Trump endorsing fringe South African political parties, and rapper Eminem attacking the ruling party gained

significant traction. According to group-level forwarding statistics and visible share counts, these videos were widely circulated and appeared in multiple large WhatsApp groups. However, due to WhatsApp's privacy protections, it is not possible to determine precise viewership numbers. These developments, however, unfolded largely beneath the radar of regulators, journalists, and researchers, owing to the very architecture that makes WhatsApp both ubiquitous and inscrutable.

Against this backdrop, this study addresses the gap by examining WhatsApp disinformation during South Africa's 2024 general elections. Specifically, it investigates how political narratives (both organic and AI-generated) spread within closed messaging environments. Special attention is given to the role of AI-generated content, the architecture of group-based influence, and the limitations of current verification tools. In this study, "political influence" is understood not as direct changes in voter behavior but as the ways AI-generated disinformation shapes group narratives, reinforces partisan identities, and amplifies political messaging within WhatsApp groups. In situating the analysis in the South African context, this research seeks to deepen understanding of the patterns of misinformation in EMAs, while also contributing to theory on digital propaganda, group-based communication, and the challenges of safeguarding electoral integrity in underrepresented regions such as sub-Saharan Africa.

Literature and Theoretical Framework

Disinformation, Echo Chambers, and the Challenge of EMAs

Over the past ten years, EMAs such as WhatsApp, Signal, and Telegram have changed how people communicate in many parts of the world, especially in the Global South. These apps are now important for both daily conversations and political organizing (Nemer 2022; Martín et al. 2021). Unlike public platforms such as Facebook or X (formerly Twitter), EMAs use end-to-end encryption. Only senders and recipients can see the messages. This setup not only protects privacy but also makes it hard for researchers, regulators, and fact-checkers to monitor what is happening (Donovan and Boyd 2021; Gondwe 2025; Rossini 2023).

EMAs can be helpful for civic action in places where media freedom is limited or trust in institutions is low. At the same time, they make it easier for political disinformation to spread out of sight. For example, during Brazil's 2018 presidential election, more than half of the most shared political images in WhatsApp groups were misleading or false, spreading quickly through networked sharing (Tardáguila et al. 2018). Projects like Comprova tried to fact-check these messages, but found it difficult because encryption made tracing and stopping false information almost impossible.

The problem of disinformation extends far beyond the mere transmission of falsehoods. Bennett and Livingston (2020) define disinformation as intentionally false or misleading information shared for political, ideological, or financial gain. Importantly, recent research emphasizes that the potency of disinformation comes not just from its content, but from its emotional resonance, its exploitation of network dynamics, and the technological affordances that enable amplification (Bennett and Livingston 2018;

Tucker et al. 2018). In this view, disinformation is best conceptualized as a socio-technical phenomenon: its spread depends on strategic narrative construction, technological amplification, and the emotional investments of participants.

One foundational concept in this literature is the "echo chamber," defined as a digital environment where individuals primarily encounter perspectives that affirm their existing beliefs (Pariser 2011; Sunstein 2017). Social media's algorithms and user behaviors can foster such insularity, but EMAs intensify this effect. EMAs can generate what might be called "double echo chambers" not only are messages and ideas continuously reinforced, but group membership itself is selective, often comprising of tightly knit, ideologically similar individuals (Cinelli et al. 2021). This dual insulation strengthens social trust and emotional validation within the group, making disinformation more likely to be accepted, shared, and defended against external critique (Tucker et al. 2018; Zhang 2022). The impact of such "double echo chambers" is especially significant in regions where EMAs have rapidly become central to political life. This is evident in recent African elections, where these platforms have played a growing, and sometimes controversial, role in the political systems of most African countries.

EMAs and Electoral Disinformation in Africa. The role of EMAs in African democracies has received comparatively less scholarly attention, despite their growing influence in shaping political discourse. Studies from Zambia and Kenya illustrate how encrypted platforms facilitate both grassroots mobilization and the covert spread of politically motivated disinformation (Gondwe 2024a; Nyabola 2023). In Kenya's 2022 general elections, deepfake videos and doctored audio clips circulated widely via WhatsApp, smearing opposition figures and reinforcing ethno-political divisions. Similarly, in Zambia, WhatsApp was used simultaneously by civil society to organize protests and by state actors to disseminate counter-narratives aimed at suppressing dissent (Gondwe 2024b). Notably, these platforms operate in regulatory vacuums. As Rossini (2023) and Munger et al. (2024) point out, traditional methods of content moderation and platform governance are largely ineffective in encrypted spaces. The combination of technical opacity and the persuasive power of interpersonal networks creates a disinformation environment that is both high-impact and low-visibility.

A new dimension of disinformation has emerged through the use of generative artificial intelligence, particularly in the creation of deepfakes. Deepfakes, defined as synthetic media generated with AI, are increasingly used to fabricate politically incendiary content, ranging from videos and audio clips to images and textual messages (Chesney and Citron 2019). Their realism and virality have raised alarms about their potential to distort electoral outcomes and erode public trust in democratic processes (Vaccari and Chadwick 2020). The 2024 South African general elections marked a watershed moment in this regard. Unlike traditional broadcast media or even open social platforms, EMAs create "intimate publics" (Zhang 2022; Zhu et al. 2022) where trust in the sender (often a family member, friend, or community leader) imbues messages with a veneer of credibility.

Research has shown that misinformation received via private messages is more likely to be believed and shared, particularly when it resonates with preexisting

political or ideological beliefs (Guess et al. 2019; Tandoc et al. 2020). The architecture of closed messaging groups further exacerbates confirmation bias, creating echo chambers where falsehoods are rarely challenged. In addition, studies in Latin America (Martín et al. 2021; Resende et al. 2019) and Asia (Munger et al. 2024) demonstrate that forwarding limits and message labeling can reduce virality, but these technical fixes often lag behind the creativity of disinformation actors, who adapt quickly to platform constraints. The African context, with its lower digital literacy and weaker regulatory oversight, may be even more vulnerable to the weaponization of these platforms.

The Spiral of Silence and Networked Propaganda. To understand the dynamics of disinformation dissemination within WhatsApp groups, this study draws on two complementary theories in political communication: Elisabeth Noelle-Neumann's Spiral of Silence and the more recent framework of Networked Propaganda (Benkler et al. 2018).

The Spiral of Silence posits that individuals are less likely to express minority opinions in public for fear of social isolation. This silence allows dominant narratives (whether truthful or false) to gain momentum unchallenged (Noelle-Neumann 1974). In WhatsApp groups, where social relationships and perceived in-group loyalty are strong, users are often reluctant to challenge misinformation, especially when it aligns with group consensus or emotional resonance (Madrid-Morales 2021; Tully et al. 2022). This facilitates the uncritical spread of disinformation, reinforcing dominant views and marginalizing dissenting voices. This theory is particularly useful in explaining the *passivity* of recipients who suspect content might be false but choose not to contradict it within politically charged or familial groups. It also helps explain how fringe narratives gain traction when repeated unchallenged in ideologically homogeneous groups.

On the other hand, Benkler et al.'s (2018) theory of networked propaganda provides a structural complement to the spiral of silence. It emphasizes how decentralized networks, such as WhatsApp groups, can be systematically leveraged for coordinated disinformation campaigns. In this model, elite actors inject false content into the network, which is then organically amplified through user-driven redistribution. The process is iterative and resilient, relying not on centralized control but on distributed trust and peer validation. This framework is important for understanding how deepfakes and emotionally charged disinformation videos in this study were propagated across heterogeneous yet ideologically aligned WhatsApp networks. It also underscores the limitations of traditional gatekeeping models and highlights the need for new paradigms of verification and accountability.

Drawing on a synthesis of the spiral of silence and networked propaganda frameworks, this study proposes a multilayered approach to understanding the proliferation of disinformation within EMAs. Rather than viewing disinformation solely as a technological or content-based issue, we argue that its diffusion in EMAs is best understood as a sociopolitical process shaped by intertwined dynamics at the individual, group, and systemic levels. At the microlevel, users in WhatsApp groups are

embedded in tightly knit, often ideologically homogenous social environments where interpersonal trust is high and dissent can be socially costly. In such settings, individuals may choose silence over contradiction when confronted with dominant (even inaccurate) narratives, a phenomenon that aligns with Elisabeth Noelle-Neumann's *Spiral of Silence* theory. The fear of social isolation or disruption of group harmony leads to self-censorship, which in turn allows disinformation to circulate unchecked and become normalized within these intimate digital spheres.

At the meso-level, disinformation gains momentum through group dynamics and emotional engagement. Content that evokes fear, anger, indignation, or humor travels faster and is more likely to be shared than neutral or fact-based information. This mirrors findings from affective politics literature, which suggests that emotionally resonant messages bypass critical evaluation and instead foster immediate action and in-group solidarity. Within WhatsApp groups, the virality of such emotionally charged content accelerates the spread of falsehoods, often reinforcing existing biases and sharpening polarization. At the macrolevel, we consider how networked propaganda (as theorized by Benkler et al. (2018)) operates within the structural affordances of EMAs. Politically motivated actors, including campaign operatives, state-affiliated influencers, and partisan activists, exploit the decentralized nature and encryption of platforms like WhatsApp to introduce false content that cascades through networks with little oversight.

The opaque infrastructure of EMAs hinders platform-level moderation and regulatory intervention, enabling the fabrication and circulation of coordinated disinformation campaigns under the guise of organic user behavior. Therefore, we argue that understanding disinformation within EMAs necessitates a departure from purely technological or informational perspectives. Taken together, the hybrid theoretical lenses assert that the spread of disinformation in EMAs is not accidental, nor merely the result of uninformed users. Instead, it is a consequence of systemic vulnerabilities and deep social processes. The interplay of microlevel conformity, mesolevel emotional contagion, and macrolevel strategic exploitation reveals that disinformation in EMAs is as much about *who* shares and *why*, as it is about *what* is shared.

Despite these theoretical advancements in understanding disinformation on EMAs, significant research gaps persist, particularly concerning empirical data from African electoral contexts. Large-scale analyses of encrypted disinformation during national elections in the region remain scarce. This gap becomes even more urgent in light of the rise of AI-generated content, such as deepfakes, which add new layers to how disinformation is produced, perceived, and propagated. While such content has garnered attention in the context of open social media, that is, X, Instagram, Facebook, etc., its role within encrypted platforms (where moderation is nearly impossible, and messages circulate among familiar, credible sources) remains largely unexplored. Moreover, methodological challenges posed by encryption have left researchers with limited tools for capturing and analyzing disinformation flows within these private spaces, often resulting in data that is either anecdotal or incomplete. Given these limitations, this study seeks to address the following research question:

RQ: *How did political narratives circulate within WhatsApp groups during South Africa's 2024 general elections, and what observable group behaviors and platform features were associated with their diffusion, credibility, and resistance to verification?*

Methods

This study used qualitative content analysis to examine political disinformation disseminated via WhatsApp during South Africa's 2024 general elections. Data were collected over five months, from February 26 to June 30, 2024, covering both preelection build-up and postelection reactions. The protocol received ethical clearance from the institutional IRB on January 28, 2024.

To build a diverse sample, we identified 47 politically active WhatsApp groups through systematic searches of public invitation links posted on election-related forums, public social media, and through referrals from our research assistants' networks. These groups varied in party affiliation, geographic region, and group size (ranging from 24 to over 1,000 members), and included both open and closed groups. Open groups were defined as those accessible to anyone via a public invitation link. Research assistants joined these groups using the invite link. For closed groups, we contacted group administrators to explain the study's purpose and request permission for research access. When access was granted, administrators were asked to notify group members about the presence of a researcher. In closed groups, informed consent was obtained by informing members about the observers, the research aims, and the anonymization of all data.

For open groups, individual consent was not typically required; strict anonymization was enforced, no private or one-on-one messages were collected, and all user identities were removed before analysis. In some cases, research assistants joined open groups where individual members may not have been directly informed. In line with leading digital ethnography scholarship (Markham and Buchanan 2015; Williams et al. 2017), such observation was considered ethically permissible due to rigorous anonymization and privacy safeguards. All protocols were reviewed and approved by the IRB.

Data Collection and Analysis

Research assistants used WhatsApp Web's "Export chat" feature to collect messages from groups, with some content manually archived as needed. The final dataset consisted of 22,384 unique messages shared by 6,283 users. All messages were anonymized prior to analysis, and private/direct messages were excluded. Data analysis proceeded in several stages. First, messages were archived and logged in a centralized spreadsheet. Second, each item was analyzed using MyFactChecker,[1] an open-source AI-powered tool that cross-references content against verified databases, identifies sentiment patterns, flags disinformation typologies, and provides corrective feedback.

We note that MyFactChecker is not yet peer-reviewed for African languages; therefore, all sentiment and typology classifications, particularly those relating to fear-based or identity appeals, were cross-checked and, if needed, corrected by bilingual human coders. Where possible, we compared the date of fact-checking publications with the timestamps of group messages to assess whether circulation occurred before or after verification. In some cases, disinformation persisted despite fact-checks already being available, while in others, fact-checks were introduced later. Because group activity and external verification timelines were not always fully aligned, we describe circulation patterns conservatively, avoiding claims of direct causal effect.

Initially, messages were coded as "factual," "misleading," or "false" based on manual review and fact-checking. To provide a more nuanced understanding, we further categorized messages using taxonomies from Diaz Ruiz (2025) and Kapantai et al. (2020), including misinformation, disinformation, mal-information, satire/parody, imposter content, fabricated content, manipulated content, and false connection/context. We also coded for content type, intent, and dissemination strategy. Coding was performed by two independent coders, with disagreements resolved by consensus.

To determine which messages were "most widely shared," we tracked "forwarded" and "forwarded many times" labels in WhatsApp, counted repeated postings across groups, and recorded how often each message or media item appeared in multiple groups. This allowed us to estimate relative reach and identify key narratives or deepfakes circulating widely in our sample. For each prominent disinformation narrative, we compared WhatsApp content to contemporaneous coverage in mainstream South African media (e.g., News24, Daily Maverick, SABC News) and to leading fact-checking databases (e.g., Africa Check, AFP Fact Check). Research assistants searched for keywords and topics matching group narratives, coded the coverage as confirmatory, contradictory, or neutral, and documented discrepancies. Disagreements were resolved through group discussion.

To ensure coding reliability among the 15 research assistants, we double-coded a random sample of 10 percent of messages in each coding wave. Intercoder agreement was assessed using Cohen's kappa, which averaged .92 ($p < .001$, 95% CI) across key categories. Discrepancies were discussed during weekly consensus meetings, leading to iterative refinement of the codebook (Guest et al. 2012). For highly subjective dimensions, codes were finalized by group consensus and reviewed by the principal investigator. Throughout, we recognized the ethical distinctions between observation with explicit disclosure and participant consent, and observation in open groups, maximizing transparency and minimizing risk. Rigorous anonymization procedures were enforced in all cases. These procedures ensured that the study adhered to the highest ethical standards while contributing to a deeper understanding of digital disinformation in African electoral contexts.

Findings

Drawing on a dataset of 22,384 messages from 47 WhatsApp groups involving 6,283 unique users between February and June 2024, this study identified three major

patterns in the circulation of political narratives: (1) the widespread presence and acceptance of AI-generated disinformation, (2) the central role of group-based trust in message credibility and diffusion, and (3) the limits of verification tools within the encrypted messaging environment. While these findings reflect dynamics specific to WhatsApp groups in this electoral context, the study does not assess whether these patterns are unique to WhatsApp or differ from other social media group structures.

AI-Generated Narratives as Instruments of Electoral Strategy

In the months leading up to South Africa's 2024 general elections, there was a noticeable shift in the way political narratives were shaped and circulated. WhatsApp groups, already central to political conversation, became saturated with highly convincing AI-generated content, particularly deepfake videos and synthetic audio clips that featured well-known international figures. These instances did not appear isolated or marginal. Rather, they were widely circulated and seemed to serve functions consistent with influencing political perception, reinforcing partisan identities, and mobilizing emotion in an already polarized media environment. Among the most widely shared pieces of AI-generated disinformation was a deepfake video that appeared to show U.S. President Joe Biden issuing a stern warning: that the United States would impose sanctions if the African National Congress (ANC) were to win the election. The video was rendered with high visual fidelity, using voice-cloning software and lip-syncing algorithms that mimicked Biden's delivery style and rhetorical cadence with uncanny accuracy. Though entirely fabricated, the clip was quickly absorbed into anti-ANC WhatsApp groups, where it was interpreted as validation of longstanding suspicions about the ANC's global reputation. The clip's persuasive impact appeared less tied to any factual basis and more to the affective confirmation it offered to users predisposed to view the ANC as politically toxic.

Another deepfake that gained similar traction featured Donald Trump, who was shown offering a dramatic endorsement of the uMkhonto weSizwe (MK) party. The video's appeal was both visual and symbolic in that Trump's global populist persona was mobilized to add weight and perceived legitimacy to a relatively new political formation. With over 158,000[2] recorded views, the video spread rapidly through MK-aligned groups and was treated not as satire or fabrication but as a diplomatic breakthrough worth celebrating and redistributing. Much like the Biden clip, it was rarely questioned within sympathetic circles, even when it was flagged as false by external fact-checkers. A third example took a different but equally potent approach, blending pop culture with political provocation. This video depicted rapper Eminem (again rendered through advanced AI tools) voicing support for the Economic Freedom Fighters (EFF) while criticizing the ANC.

The video's design leaned heavily on emotional symbolism. It fused the authority of a global celebrity with the rebellious, anti-establishment energy that has long defined the EFF's public image. It reached over 173,000[3] viewers, many of whom responded with admiration and urgency. In EFF-associated groups, the deepfake was not only circulated but also actively championed, contributing to a shared sense of

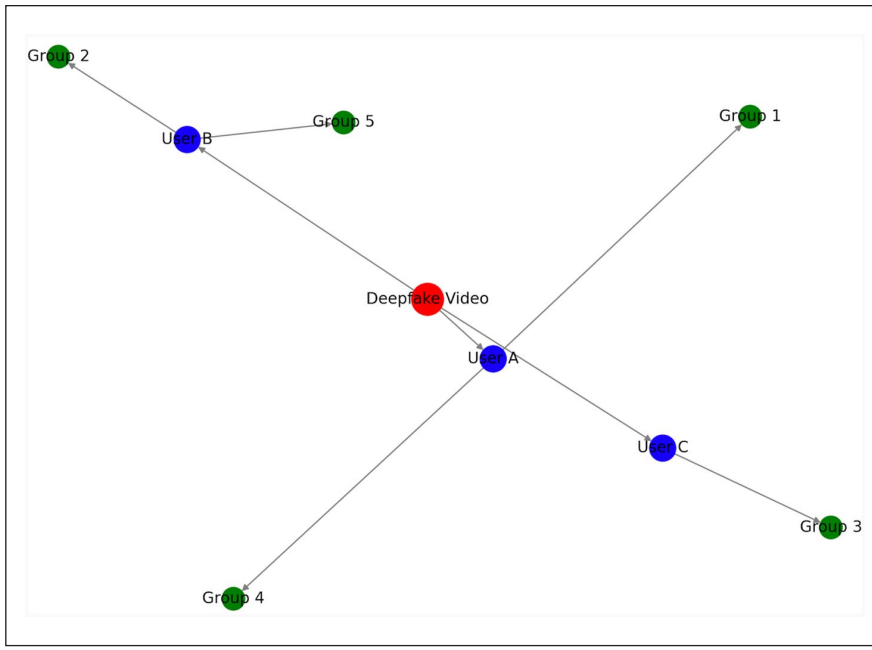

Figure 1. Disinformation flow across WhatsApp groups.
This network diagram illustrates how a single deepfake video (Donald Trump endorsing MK) spread from one large group (>500 users) into five additional groups, facilitated by three bridge participants. The flow highlights horizontal, not top-down, message spread—echoing the decentralized logic of networked propaganda.

momentum and outsider solidarity. An analysis of sharing patterns for one AI-generated deepfake (purportedly featuring Donald Trump endorsing the MK party) shows how it spread horizontally across interconnected WhatsApp groups (see Figure 1). While the figure shows how a message can move horizontally through interconnected groups, it reflects only a subset of the observed diffusion and should not be taken as representative of all message flows in the dataset.

Across all three cases, the common thread was not just technical sophistication but also narrative timing and emotional design. These AI-generated clips did not circulate in a vacuum. They emerged at critical junctures in the campaign cycle, often timed around political rallies, key debates, or electoral announcements. They succeeded not simply because they were realistic but also because they resonated emotionally with targeted audiences and echoed sentiments already circulating within group conversations. Even when flagged as false by tools such as MyFactChecker, these videos continued to travel across platforms.[4] Fact-checking interventions appeared to have limited impact, particularly in ideologically homogenous groups where such efforts were viewed with suspicion or outright hostility. Many users dismissed correction attempts as "mainstream manipulation" or foreign interference, reinforcing a broader

skepticism toward institutional verification. As a result, the AI-generated content retained its affective power, serving less as contested information and more as political ammunition in group-based narrative battles.

Group Structure and the Virality of Misinformation

This study found that the spread of political disinformation on WhatsApp during South Africa's 2024 general elections was rooted in the underlying architecture of WhatsApp group design, the ideological composition of these groups, and the relational dynamics among users. As a result, disinformation circulated through a set of structural and affective logics that made synthetic content both viral and socially validated. As a result, group size and the presence of authority figures emerged as key variables that influenced the velocity of message forwarding. Larger groups (especially those exceeding 300 members) functioned as high-capacity channels where information moved swiftly and with minimal resistance. These groups were also socially organized in ways that encouraged deference to certain individuals.[5]

Messages originating from group administrators or respected community members were treated as credible, regardless of their accuracy. This was particularly important for AI-generated narratives, which often gained traction not because of their technical sophistication but because of the trust conferred by the sender.[6] In this way, group architecture became a conduit through which false information could rapidly scale. While group size enabled speed, it was ideological alignment that created stickiness. Politically oriented WhatsApp groups were often ideologically homogenous, especially those that expressed strong anti-ANC or pro-opposition sentiments. Within these echo chambers, narratives (whether AI-generated or organically constructed) were rarely interrogated for accuracy. Instead, they were embraced, reframed, and emotionally reinforced. Content that aligned with group beliefs was often interpreted as confirmation of long-held grievances, with users appending their own commentary or anecdotes to intensify its resonance.

Rather than questioning misleading content, participants deepened its emotional charge, transforming disinformation into what might be called *shared emotional truth*. Even when corrective information was introduced, that is, through fact-checking links or screenshots of news articles, it was often met with suspicion, dismissed as propaganda, or reframed as a foreign attempt to meddle in domestic politics. This rejection of verification points to a deeper phenomenon of the emergence of epistemic enclaves. In other words, these groups were not just communities of shared belief but of shared distrust. Within such enclaves, the boundaries of truth were defined internally, by the group itself, rather than through engagement with external sources. The social dynamics in these spaces allowed falsehoods to flourish not in spite of efforts to correct them, but because those corrections came from outside the group's moral or political universe. As a result, misinformation was not merely tolerated, but it was metabolized, emotionally integrated, and redeployed in ways that reinforced group identity.

The viral power of disinformation, however, did not remain confined to single groups. Many users were simultaneously members of multiple WhatsApp and

Facebook groups that span different ideological, geographic, and social communities. These individuals acted as bridges, with the responsibility of enabling messages to travel horizontally across networks. This intergroup cascading effect gave AI-generated content the appearance of spontaneous, widespread circulation. When the same narrative appeared in multiple contexts (sometimes altered slightly in wording or format) it gained perceived legitimacy. Repetition across groups functioned as a proxy for authenticity, creating a false sense of consensus. Users encountering the same message in five different groups were more likely to believe it was true.[7] In this way, the spread of disinformation was not driven by isolated bad actors or passive consumption but by a deeply social process. WhatsApp's closed-group architecture, coupled with ideological alignment and the presence of cross-group participants, created a self-reinforcing ecosystem. Within it, AI-generated narratives moved rapidly, were emotionally reinterpreted to align with local grievances, and gained the illusion of broad consensus through patterns repetition, which seems to raise suspicions of coordinated or strategic amplification. The combination of technical affordances and human behavior thus rendered WhatsApp a powerful platform for the amplification of false political narratives during the election period.

Emotional Resonance and the Affective Life of Disinformation

The analysis of emotional appeals within disinformation content circulated on WhatsApp during South Africa's 2024 general elections reveals a clear reliance on affective manipulation to drive engagement and diffusion. As illustrated in Figure 2, fear-based appeals constituted the largest share of disinformation messages (41%), often invoking economic collapse, foreign sanctions, or civil unrest to provoke panic and urgency. Identity-based rhetoric followed closely at 32 percent, employing race, religion, and cultural threat narratives to deepen group polarization and reinforce ingroup solidarity. Notably, 27 percent of the content adopted the visual and structural features of legitimate journalism, such as datelines, press-style headlines, and mimicked news formatting to fabricate credibility and confuse recipients.

Essentially, the analysis of message sentiment, supported by the AI tool *MyFactChecker*, revealed the centrality of emotional appeal in determining the virality of disinformation. In other words, our data revealed that disinformation was not only a matter of content but of feeling. Political falsehoods did not spread merely because they existed but because they stirred something. Using *MyFactChecker's* sentiment analysis, we identified a clear trend in how disinformation spreads. Messages that evoked strong emotions (especially fear, identity, or outrage) were shared far more frequently than those with neutral or purely factual content. Among all the messages flagged as false or misleading, 41 percent deployed fear-based appeals, tapping into deep-seated anxieties around economic collapse, looming foreign sanctions, and the specter of civil unrest. These messages often framed the election as a life-or-death scenario, using language saturated with existential dread and urgency.

One widely shared post warned that a particular election outcome would "trigger international sanctions overnight," while another falsely claimed that a currency crash

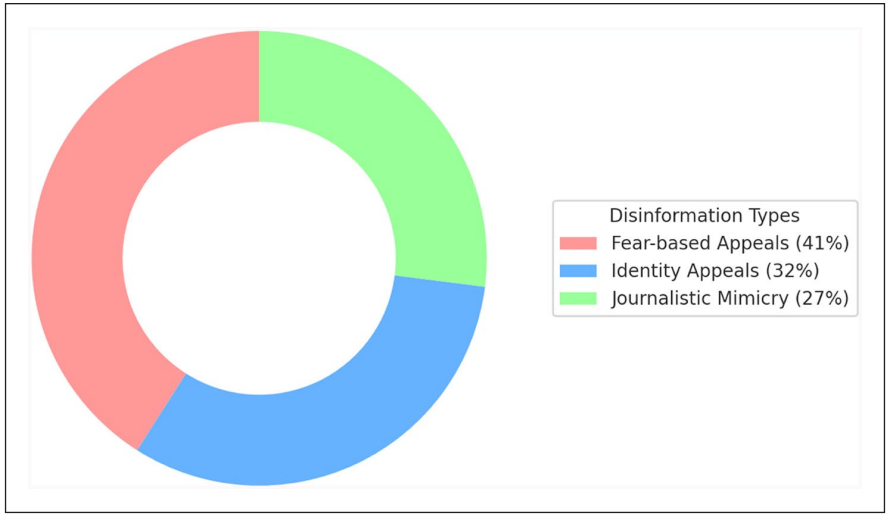

Figure 2. Distribution of emotional appeals in disinformation content.
Figure 2 shows that among the flagged disinformation messages, 41 percent were fear-based, 32 percent identity-based, and 27 percent mimicked journalistic aesthetics. These emotional typologies correlated strongly with virality and user engagement.

was imminent if the ruling party remained in power. Such kind of emotions were further substantiated with factual news reported from legitimate sources like Reuters, which reported on the currency, stocks, and bonds dropping amid political uncertainty (Anders and Strohecker 2024). Such emotionally loaded narratives were often punctuated by urgent calls to action like, "Forward this now before it's too late!." This created a feedback loop of panic and amplification.

In another category, 32 percent of the messages relied on identity-based appeals, especially those rooted in race and religion. These messages evoked fear and anger through populist rhetoric, suggesting that certain political parties would privilege one ethnic or religious group over another, or that the election represented a final stand for cultural survival. Deepfake videos and audio clips in this category often exaggerated or fabricated threats from "the other side," framing political opponents as existential enemies. For example, the rhetoric associated with Julius Malema and the EFF frequently employed identity-based appeals, often targeting racial and ethnic divisions.

Specifically, messages circulating within WhatsApp groups included deepfake audio clips purportedly featuring Malema making inflammatory statements about the need to "reclaim land" from white South Africans, framing it as a necessary act of historical justice. While most of these clips were developed from true events, many of them were often accompanied by manipulated images or fabricated statements. This amplified existing anxieties about land redistribution and racial inequality, tapping into deep-seated historical grievances. Furthermore, messages attributed to EFF supporters often depicted political opponents, particularly those from predominantly

white or minority parties, as threats to the "survival" of black South Africans. The result was a portrayal of elections as a zero-sum game with existential consequences. In groups already primed by shared grievances or historical injustices, such narratives quickly escalated into collective indignation and hardened group solidarity.

A further 27 percent of the disinformation content was styled to mimic legitimate journalism, borrowing the aesthetic and structural conventions of credible news outlets to fabricate legitimacy. These pieces often looked like press releases or breaking news updates, which blurred the lines between fact and fiction. Users encountering these posts were less likely to question their validity, particularly when the design matched that of trusted national broadcasters, international agencies, or embassies. The journalistic formatting of these posts may have contributed to their credibility. This blending of form (professional presentation with deceptive content) added a layer of credibility that allowed these falsehoods to pass unchallenged, especially in fast-moving group chats.

Resistance to Verification and the Limits of Corrective Tools

Despite the implementation of the AI-powered *MyFactChecker* tool, the study observed a widespread failure of verification mechanisms to counteract or neutralize false narratives. In groups which seemingly exhibited some strong political homogeneity, corrective content either failed to penetrate or was outright rejected. One of the most significant barriers was the architectural opacity of WhatsApp itself. The platform's end-to-end encryption and private group structures, while vital for user privacy, simultaneously create a hostile terrain for fact-checkers. Corrections issued by external organizations or embedded in public information systems rarely permeated the sealed boundaries of politically oriented groups. Unlike public-facing platforms where verified information can be algorithmically boosted or inserted into users' feeds, WhatsApp's design disallowed any centralized intervention. Consequently, factual rebuttals and corrective updates often remained external to the very audiences most influenced by the falsehoods. In the absence of direct access, these fact-checking efforts operated in parallel rather than in response to misinformation, thus rendering their presence largely symbolic.

Equally troubling was the deep-seated mistrust toward verification itself. Within many of the ideologically cohesive groups studied, fact-checking content was not simply ignored but actively rejected. Users frequently interpreted verification efforts as forms of ideological interference, dismissing them as propaganda by mainstream media, opposition campaigns, or foreign actors. In numerous instances, corrective messages were met with derision, framed as elite manipulation or attempts to stifle "alternative truths." This resistance was often more than rhetorical; it was affective and performative. In some groups, members responded to fact-checks with emojis, sarcastic commentary, or counter-memes that ridiculed the credibility of verification sources. This mockery reinforced in-group solidarity and delegitimized external authority, further embedding the original falsehoods into the group's narrative ecosystem.

Discussion

This study set out to investigate the mechanisms and dynamics of political disinformation dissemination through WhatsApp during South Africa's 2024 general elections. It aimed to address both empirical and conceptual gaps in the study of EMAs by analyzing 22,384 messages from 47 politically active groups. Specifically, it examined how AI-generated disinformation, group-based trust, and verification failures facilitated the spread of false narratives. Anchored in the Spiral of Silence (Noelle-Neumann 1974) and Networked Propaganda (Benkler et al. 2018), the study examined not just the content of disinformation but the sociotechnical infrastructures that sustain it. Using MyFactChecker, an AI-driven verification tool, the study generated granular insights into how emotional, architectural, and ideological forces shaped the misinformation ecosystem within WhatsApp's closed and encrypted environments.

The findings extend existing literature on the opacity and potency of EMAs in political disinformation campaigns (Donovan and Boyd 2021; Rossini 2023). While prior research on Brazil's 2018 elections (Nemer 2022; Tardáguila et al. 2018) discuss the use of hierarchical forwarding structures, this study reveals how AI-generated content in South Africa's 2024 elections added layers of emotional intensity and audiovisual manipulation to such strategies. Deepfakes featuring figures such as Joe Biden, Donald Trump, and Eminem show that disinformation has moved beyond text to immersive synthetic media, echoing the alarm raised by Chesney and Citron (2019) on the growing threat of deepfakes. These findings also support Vaccari and Chadwick's (2020) conclusion that emotionally charged content tends to bypass critical scrutiny, spreading rapidly within ideological communities. More crucially, this study finds that such content derives persuasive power not only from its form or content but also from the relational trust between sender and recipient. In ideologically aligned WhatsApp groups, messages were judged by affective credibility rather than factual accuracy—an observation that aligns with Zhang's (2022) concept of "intimate publics" and Tandoc et al.'s (2020) analysis of interpersonal trust in disinformation networks. These group dynamics validate the Spiral of Silence, as internal social pressure discouraged dissent, and support the Networked Propaganda model by showing how decentralized trust structures enable sustained disinformation flows (Benkler et al. 2018).

A key implication of this study is the inadequacy of traditional verification approaches in encrypted environments. Despite the introduction of MyFactChecker, attempts at verification failed to slow or reverse the spread of disinformation in politically homogeneous groups.[8] In many cases, fact-checks were met with skepticism and reframed as elitist, foreign, or politically biased interventions. Rather than correcting falsehoods, fact-checking often triggered defensive responses and reinforced group cohesion around misinformation. This reflects a broader structural issue: WhatsApp's design (its end-to-end encryption, intimate group architecture, and invisibility to external monitors) renders conventional fact-checking efforts socially and structurally illegible. This raises urgent concerns about the efficacy of current content moderation and platform governance strategies, which are primarily designed for open, surveillable platforms (Donovan and Boyd 2021).

These findings have wider implications for other Global South contexts, where EMAs are central to political communication, often in environments marked by institutional distrust and low media credibility. In Kenya, ethnically aligned WhatsApp groups show similar patterns of disinformation, especially during elections. In Brazil, emotionally resonant and nationalist rhetoric spread through WhatsApp mirrors the South African experience (Nemer 2022; Tardáguila et al. 2018). In the Philippines, encrypted messaging platforms like Viber and Facebook Messenger are used to circulate polarizing narratives, often beyond the reach of moderation (Lanuza et al. 2021). The interplay of emotional appeal, ideological homophily, and platform affordances seen in South Africa likely applies across these contexts.

However, South Africa presents unique characteristics rooted in its post-apartheid political environment. The legacy of racialized governance, economic inequality, and the symbolic power of liberation movements such as the ANC and the emergent MK Party creates a distinctive information ecology. The MK Party's use of EMAs to mobilize identity-based grievances demonstrates how disinformation exploits historical memory and factional loyalties—dynamics that may not be as potent elsewhere. Future research could test these dynamics comparatively by tracing the virality and emotional framing of political messages in WhatsApp groups in Brazil's favelas, Kenya's ethnically polarized constituencies, or Filipino diaspora networks. Techniques such as ethnographic content tracing, network analysis, and sentiment classification could help assess whether emotional misinformation consistently outperforms fact-based content, and whether ideological cohesion accelerates its spread.

These patterns create significant challenges for democratic processes in the Global South, especially in countries with limited regulatory oversight and lower levels of digital literacy. Our findings show that disinformation often works by exploiting emotions, taking advantage of trusted social connections, and avoiding detection by conventional verification methods. Content moderation alone may not be enough to protect electoral integrity in these environments. Based on our analysis, we suggest that interventions should include proactive educational efforts (such as pre-bunking common falsehoods), strengthening local networks of trust, and developing digital literacy initiatives that are rooted in the specific needs of affected communities. Approaches tailored to local contexts may be more effective in reducing the impact of disinformation than external or generic fact-checking efforts.

While the study identified consistent patterns of disinformation diffusion, it does not claim direct causality between message attributes and behavioral outcomes. Instead, it interprets these associations as indicators of broader sociotechnical dynamics. Although emotionally resonant content showed a higher likelihood of circulation, the study does not assert that it directly changed political opinions or voting behavior. These patterns suggest correlations and not definitive effects. Going forward, research should continue to explore affective dimensions of misinformation in EMAs, particularly through ethically sound methods for studying encrypted networks. There is also a need for interdisciplinary collaboration among technologists, social scientists, and local stakeholders to develop context-sensitive tools for misinformation mitigation. Cross-national comparative studies would help uncover recurring vulnerabilities, effective resistance strategies, and policy innovations that transcend regional borders. In sum, this study charts the

disinformation systems of South Africa's 2024 elections and underscores the urgent need to rethink our theoretical and practical responses to misinformation in the age of AI, affective polarization, and encrypted communication.

ORCID iD

Gregory Gondwe https://orcid.org/0000-0001-7444-2731

Funding

The authors disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: The study did not receive direct funding other than professional development funds from The Department of Communication and Media at California State University - San Bernardino.

Declaration of Conflicting Interests

The author declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Notes

1. It is important to note that neither the research team nor anyone associated with the MyFactChecker project actively introduced or injected any content, including fact-checks, into the observed WhatsApp groups. We emphasize that all corrective content we analyzed was introduced by regular group users not by the research team. Our role was strictly observational.
2. View counts for widely circulated videos were estimated by combining the total membership of WhatsApp groups in which the videos appeared with publicly available view metrics from external platforms (such as TikTok or YouTube), where the videos were originally shared. These figures should be interpreted as upper-bound estimates of potential reach, as audience overlap and cross-posting between platforms likely result in some double-counting.
3. The video was circulated in several large WhatsApp groups associated with the EFF, with a combined group membership totaling up to 173,000 users. Due to WhatsApp's privacy protections and lack of view analytics, this figure should be understood as an upper-bound estimate based on group sizes and observed repost frequency, rather than a precise count of unique viewers. Many group members responded to the deepfake with admiration and urgency, and within EFF-associated groups, the video was not only widely shared but also actively championed, contributing to a sense of momentum and outsider solidarity.
4. Even videos flagged as false by tools such as MyFactChecker continued to circulate across WhatsApp groups. In several cases, we observed fact-checking posts (e.g., Africa Check links or MyFactChecker outputs) within the dataset, yet the same videos appeared again in subsequent group conversations. However, we note that not all sequences allowed for strict before/after temporal tracing, so we present these as concurrent patterns rather than direct causal effects.
5. It is important to note that, as our analysis does not include data from public or alternative social media group structures, we cannot assert that these dynamics are unique to WhatsApp or are causally distinct from those on other platforms. Our observations are therefore specific to the closed group context and should be interpreted as exploratory.

6. We note that content often gained traction when posted by group administrators or highly active members. While prior literature suggests that interpersonal trust within EMAs can enhance credibility (e.g., Tandoc et al. 2020; Zhang 2022), our data do not allow us to directly measure or confirm trust. We therefore interpret these patterns cautiously as consistent with (but not proof of) the role of relational dynamics in message diffusion.
7. The repeated appearance of the same message across multiple groups could plausibly function as a proxy for truthfulness, as repetition may create an illusion of verification. However, our dataset does not allow us to infer the internal beliefs of users directly.
8. We emphasize that while disinformation often circulated alongside or even after fact-checking interventions, our data cannot always establish strict before-and-after causality. Thus, we interpret these patterns as evidence of persistence rather than definitive proof of fact-checking failure.

References

Anders, T., and K. Strohecker. 2024, May 30. "South African Assets Slip on Uncertain Election Outcome." *Reuters*. https://www.reuters.com/markets/emerging/south-african-assets-slip-uncertain-election-outcome-2024-05-30/

Benkler, Y., R. Faris, and H. Roberts. 2018. *Network Propaganda: Manipulation, Disinformation, and Radicalization in American Politics*. Oxford University Press.

Bennett, W., and S. Livingston. 2020. *The Disinformation Age*. Cambridge University Press.

Bennett, W. L., and S. Livingston. 2018. The disinformation order: Disruptive communication and the decline of democratic institutions. *European journal of communication*, 33(2), 122–139.

Chesney, B., and D. Citron. 2019. "Deep Fakes: A Looming Challenge for Privacy, Democracy, and National Security." *Calif. L. Rev*. 107: 1753.

Cinelli, M., G. De Francisci Morales, A. Galeazzi, W. Quattrociocchi, and M. Starnini. 2021. "The Echo Chamber Effect on Social Media." *Proceedings of the National Academy of Sciences* 118, no. (9): e2023301118.

Diaz Ruiz, C. 2023. "Disinformation on Digital Media Platforms: A Market-Shaping Approach." *New Media and Society* 27, no. (4): 2188–211. https://doi.org/10.1177/14614448231207644 (Original work published 2025)

Donovan, J., and D. Boyd. 2021. "Stop the Presses? Moving from Strategic Silence to Strategic Amplification in a Networked Media Ecosystem." *American Behavioral Scientist* 65, no. (2): 333–50.

Gondwe, G. 2024a. "Digital Natives, Digital Activists in Non-Digital Environments: How the Youth in Zambia use Mundane Technology to Circumvent Government Surveillance and Censorship." *Technology in Society* 79: 102741.

Gondwe, G. 2024b. "Audience Perception of Fake News in Zambia: Examining the Relationship Between Media Literacy and News Believability." *International Communication Research Journal* 47, no. (1): 1–16.

Gondwe, G. 2025. "Investigative Journalism and AI-Driven Subterfuge in Countries with Limited Press Freedom in East Africa." *Journalism Practice* 19, no. (5), 1–20.

Guess, A., K. Munger, J. Nagler, and J. Tucker. 2019. "How Accurate are Survey Responses on Social Media and Politics?." *Political Communication* 36, no. (2): 241–58.

Guest, G., K. M. MacQueen, and E. E. Namey. 2012. Validity and reliability (credibility and dependability) in qualitative research and data analysis. In *Applied thematic analysis* edited by, G. Guest, K. M. MacQueen and E. E. Namey, pp. 79–106. SAGE Publications.

Kapantai, E., A. Christopoulou, C. Berberidis, and V. Peristeras. 2020. "A Systematic Literature Review On Disinformation: Toward a Unified Taxonomical Framework." *New Media and Society* 23, no. (5): 1301–26. https://doi.org/10.1177/1461444820959296 (Original work published 2021).

Lanuza, J. M. H., R. Fallorina, and S. Cabbuag. 2021. "Understudied Digital Platforms in the Philippines." *Internews*, December.

Madrid-Morales, D., H. Wasserman, G. Gondwe, et al. 2021. "Motivations for Sharing Misinformation: A Comparative Study in Six Sub-Saharan African Countries." *International Journal of Communication*, 15, no. (2021), 1200–19.

Markham, A. N., & E. Buchanan. 2015. Ethical concerns in internet research. In *International encyclopedia of the social & behavioral sciences*, edited by J. D. Wright, 2nd ed., Vol. 10, pp. 606–13. Elsevier.

Martín, I. G., F. O. Mohedano, and M. E. P. Peláez. 2021. "Communication and Cultural Spaces in Times of COVID-19." *Vivat Academia. Revista de Comunicación* 154: 21–43.

Munger, K., A. Villegas-Cruz, J. Gallego, and M. Vásquez-Cortés. 2024. "Reenviado Muchas Veces': How Platform Warnings Affect WhatsApp Users in Mexico and Colombia." *Political Communication* 41, no. (5): 719–42. https://doi.org/10.1080/10584609.2024.2326130.

MyAIFactChecker: Young Brain Builders. https://www.myaifactchecker.org/

Nemer, D., & W. Marks. 2024. *The human infrastructures of misinformation: A case study of Brazil's heteromated labor* [Conference paper]. Cambridge Studies on Governing Knowledge Commons Workshop, University of Illinois Urbana-Champaign. Cambridge University Press. https://hdl.handle.net/2142/117214

Nemer, D. 2022. *Technology of the Oppressed: Inequity and the Digital Mundane in Favelas of Brazil*. MIT Press.

Noelle-Neumann, E. 1974. "The Spiral of Silence a Theory of Public Opinion." *Journal of communication*, 24, no. (2), 43–51.

Nyabola, N. 2023. Africa's digital public sphere. In *Routledge handbook of African political philosophy*, edited by U. Okeja, 1st ed., pp. 15–29. Routledge.

Pariser, E. 2011. *The Filter Bubble: How the New Personalized Web is Changing What We Read and How We Think*. Penguin.

Resende, G., P. Melo, H. Sousa, J. Messias, M. Vasconcelos, J. M. Almeida, and F. Benevenuto. 2019. (Mis)Information Dissemination in WhatsApp: Gathering, Analyzing and Countermeasures. In *Proceedings of The World Wide Web Conference*, edited by J. McAuley, pp. 818–828. Association for Computing Machinery.

Rossini, P. 2023. "Farewell to Big Data? Studying Misinformation in Mobile Messaging Applications." *Political Communication* 40, no. (3): 361–66. https://doi.org/10.1080/10584609.2023.2193563.

Sunstein, C. R. 2017. "Forcing People To Choose Is Paternalistic." *Mo. L. Rev*. 82: 643.

Tandoc, E. C. Jr, D. Lim, and R. Ling. 2020. "Diffusion of Disinformation: How Social Media Users Respond to Fake News And Why." *Journalism* 21, no. (3): 381–98.

Tucker, J. A., A. Guess, P. Barberá, C. Vaccari, A. Siegel, S. Sanovich, D. Stukal, and B. Nyhan. 2018. *Social media, political polarization, and political disinformation: A review of the scientific literature* [Working paper]. SSRN Electronic Journal. https://doi.org/10.2139/ssrn.3144139

Tully, M., D. Madrid-Morales, H. Wasserman, G. Gondwe, and K. Ireri. 2022. "Who is Responsible for Stopping the Spread of Misinformation? Examining Audience Perceptions

of Responsibilities and Responses in Six Sub-Saharan African Countries." *Digital Journalism* 10, no. (5): 679–97.

Tardáguila, C., F. Benevenuto, and P. Ortellado. 2018. "Fake News Is Poisoning Brazilian Politics. WhatsApp Can Stop It." *International New York Times*.

Vaccari, C., and A. Chadwick. 2020. "Deepfakes and Disinformation: Exploring the Impact of Synthetic Political Video On Deception, Uncertainty, and Trust in News." *Social Media + Society* 6, no. (1): 2056305120903408.

Williams, M. L., P. Burnap, and L. Sloan. 2017. "Towards an Ethical Framework for Publishing Twitter Data in Social Research: Taking into Account Users' Views, Online Context, and Algorithmic Estimation." *Sociology* 51, no. (6): 1149–68.

Zhang, C. Y. 2022. *Dreadful Desires: The Uses of Love in Neoliberal China*. Duke University Press.

Zhu, Q., M. Esteve-Del-Valle, and J. K. Meyer. 2022. "Safe Spaces? Grounding Political Talk in WhatsApp Groups." *New Media & Society* 26, no. (9): 5423–44. https://doi.org/10.1177/14614448221136080.

Author Biography

Gregory Gondwe is an Assistant Professor of Journalism and New Media Technologies at California State University, San Bernardino, and a Faculty Associate at Harvard's Institute for Rebooting Social Media. His research focuses on AI, journalism, and disinformation in the Global South.