

AI-DRIVEN DISINFORMATION, DEEPPAKES, AND ELECTORAL INTEGRITY IN EMERGING DEMOCRACIES

Azzallea Ega Riesta Christsanda¹, Muhammad Fitrah², Nofriadi Kurnia Putra³

¹ Universitas Gadjah Mada

² Universitas Gadjah Mada

³ Universitas Gadjah Mada

azzallea.ega1102@mail.ugm.ac.id

Article Info

Article history:

Received Apr 12th, 20xx

Revised May 20th, 20xx

Accepted May 10th, 20xx

Keyword:

AI-driven disinformation;
deepfakes; electoral integrity;
emerging democracies

ABSTRACT

The rapid diffusion of generative artificial intelligence has transformed electoral disinformation from conventional falsehoods into scalable forms of synthetic manipulation that challenge trust, accountability, and electoral legitimacy. This literature review examines how AI-driven disinformation, particularly deepfakes, synthetic media, AI-assisted political messaging, bot-supported amplification, and liar's dividend dynamics, affects electoral integrity in emerging democracies. Drawing on 17 reviewed studies spanning experimental, computational, conceptual, policy, and mixed-method designs, the article synthesises the principal forms of AI-driven disinformation, the causal mechanisms through which they operate, and the governance responses proposed to address them. The review finds that the most consistent effects do not lie in direct vote conversion, but in authenticity confusion, epistemic instability, false recognition, declining trust in information environments, and delegitimation of electoral procedures and outcomes. Evidence on candidate evaluation and voting intention exists, but it is more mixed and conditional than evidence on distrust, uncertainty, and institutional vulnerability. The review also shows that platform regulation, detection systems, fact-checking, digital literacy, and institutional responses remain partial and uneven in their effectiveness, particularly where regulatory capacity and public trust are weak. Overall, the article argues that AI-driven disinformation threatens electoral integrity by degrading the epistemic and institutional foundations on which democratic legitimacy depends.



© 2026 The Authors. Published by PT Akademi Pemilu dan Demokrasi (APD).
This is an open access article under the CC BY license
(<https://creativecommons.org/licenses/by/4.0/>)

INTRODUCTION

A profoundly transformed digital information ecology increasingly dictates the contemporary electoral arena. Within this paradigm, political communication is no longer monopolized by traditional political parties, mainstream media conglomerates, or formal state institutions. Rapidly expanding digital infrastructures have exponentially accelerated the circulation of information, dissolved geographic boundaries, and drastically reduced the barriers to producing and distributing political content. Concurrently, algorithmic curation across platforms such as Instagram, TikTok, and Facebook has rendered political exposure highly personalized, engagement-driven, and structurally vulnerable to sensationalized or emotionally manipulative content. This transformation has precipitously shifted discursive authority toward digital influencers, anonymous accounts, and non-traditional actors, thereby fragmenting the public sphere and eroding shared epistemic standards of verification. Within this volatile environment, the advent of generative artificial intelligence further exacerbates these existing vulnerabilities. Generative AI facilitates the rapid, large-scale, and highly economical production of synthetic political texts, images, and videos. Consequently, it engenders a paradoxical communication ecosystem wherein unprecedented democratic openness and severe manipulative vulnerabilities expand simultaneously (Amin, 2025; Siahaan

& Ivanna, 2024; Simbolon & Ivanna, 2024; Utami et al., 2023; Vandito et al., 2025; Wildan, 2025).

Quantitative indicators vividly illustrate the sheer magnitude of this structural transformation. Global internet penetration expanded exponentially between 1996 and 2017, while mobile internet access in the United States surged from 6.9% in 2008 to 65.4% in 2017, signifying a definitive shift toward a mobile-first information architecture. Electoral communication has concurrently become hyper-platform-intensive. This is evidenced by the generation of over 240 million election-related tweets during the 2020 United States presidential election and nearly 17 million tweets surrounding the 2017 MacronLeaks phenomenon in France. Empirically verified electoral hoaxes predominantly circulate via social networking platforms, while the proliferation of emergent synthetic content has accelerated at an unprecedented pace. For instance, a recent analysis identified 556 synthetic tweets on the X platform that collectively garnered over 1.5 billion views. The rapid mainstreaming of generative AI across diverse sectors further indicates that technologies capable of scaling persuasive and highly localized content production are no longer marginal innovations; rather, they have solidified into highly accessible, foundational communicative infrastructures (Alain et al., 2025; Casero-Ripollés et al., 2025; Corsi et al., 2024; Ferrara, 2017; Ferrara et al., 2020; Greenberg-Worisek et al., 2019; Henke, 2024; Lava-Santos et al., 2023; Pedriza, 2021; Russo, 2023).

Contextualized against this backdrop, the structural nature of electoral disinformation is undergoing a profound mutation, transitioning from conventional text-based or crudely edited visual falsehoods toward sophisticated AI-driven disinformation campaigns. These novel threat vectors encompass deepfakes, synthetic media, Large Language Model (LLM)-generated propaganda, AI-enhanced bot amplification, and the exploitation of the "liar's dividend." The distinctiveness of these modalities lies not merely in their technological novelty, but in their structural capacity to seamlessly integrate hyper-realism, unprecedented velocity, programmatic automation, and plausible deniability. Deepfakes and highly advanced synthetic formats effectively obliterate the epistemological boundary between authentic documentary evidence and fabricated media. Simultaneously, LLM-generated content enables malicious actors to scale and geographically localize disinformation campaigns with unparalleled precision.

The ensuing democratic consequences are severe: these campaigns systematically distort voter information environments, amplify public confusion and misperception, actively erode trust in legacy media and electoral institutions, legitimize baseless narratives of electoral fraud, and ultimately catalyze either civic demobilization or intensely polarized mobilization. Consequently, AI-driven disinformation threatens not only factual accuracy but also the foundational epistemic and institutional prerequisites for electoral integrity and democratic legitimacy (Bär et al., 2023; Bennett et al., 2018; Bontridder & Pouillet, 2021; Mauk & Grömping, 2023; Momeni, 2024; Prochaska et al., 2023; Romănishyn et al., 2025; Vaccari & Chadwick, 2020; Williams et al., 2024). These systemic risks acquire acute significance within emerging democracies. In these highly volatile contexts, institutional fragility, uneven regulatory capacity, heavy reliance on mobile-first platforms, pervasive digital literacy deficits,

and entrenched identity-based polarization render electoral systems disproportionately vulnerable to AI-assisted manipulation.

Extant literature demonstrates that political communication in these regions frequently depends upon closed or end-to-end encrypted platforms, notably WhatsApp, TikTok, and Instagram. Crucially, these jurisdictions often lack the robust regulatory apparatuses and oversight mechanisms available to more consolidated democracies. Nevertheless, current scholarship remains disproportionately dominated by empirical evidence derived from the Global North. Furthermore, the existing literature is highly fragmented across disparate disciplines, primarily siloed within misinformation studies, electoral integrity research, and platform or AI governance. Consequently, a severe deficit persists regarding the comparative theoretical understanding of how AI-driven disinformation operates specifically within emerging democracies, and how information disorder, epistemic integrity, and democratic resilience can be conceptually integrated. This review directly addresses this critical scholarly gap by synthesizing these disparate literatures into a unified analytical framework, focusing explicitly on the causal mechanisms, democratic risks, and governance responses pertinent to electorally vulnerable contexts (Bhattacharya et al., 2024; Bontridder & Pouillet, 2021; Broda & Strömbäck, 2024; Evangelista & Bruno, 2019; Fisher et al., 2023; Humprecht et al., 2020; Lorenz-Spreen et al., 2022; Schipper, 2025).

Building on the transformation of digital information ecosystems, this review addresses the widening gap between the rapid development of AI-driven disinformation and the still fragmented understanding of its electoral consequences in emerging democracies. Deepfakes, synthetic media, and other AI-assisted manipulations pose a more complex challenge than conventional digital disinformation because they threaten electoral integrity not only through false content, but also through their effects on trust in electoral institutions, voting behaviour, and the legitimacy of election outcomes. Yet existing scholarship remains divided across misinformation studies, electoral integrity research, and platform or AI governance, limiting a coherent explanation of how these dynamics interact within democratic processes. This gap is especially important in emerging democracies, where weaker institutional capacity, uneven digital literacy, and sharper political polarization may intensify both exposure and vulnerability.

In response, this literature review examines how AI-driven disinformation affects electoral integrity in emerging democracies by identifying its main forms, tracing its causal mechanisms, and assessing its implications for democratic vulnerability and resilience. More specifically, it explains how such content may generate distrust, confusion, demobilization, delegitimation, or polarizing mobilization, while also evaluating how platform regulation, AI governance, fact-checking, digital literacy, and institutional responses shape the capacity of democratic systems to withstand these risks. In this way, the review treats electoral integrity not merely as a procedural matter, but as a broader issue of institutional trust, democratic accountability, and the credibility of election outcomes under conditions of intensified digital manipulation.

Guided by these overarching objectives, this review operationalizes its analytical focus through two primary research questions:

1. How does the extant literature explicate the forms and causal mechanisms through which AI-driven disinformation, specifically encompassing deepfakes, synthetic media, and the liar's dividend, impacts electoral integrity, institutional trust, voting behavior, and the ultimate legitimacy of election outcomes within emerging democracies?
2. How does current scholarship evaluate the efficacy of platform regulation, AI governance frameworks, fact-checking initiatives, digital literacy programs, and broader institutional responses in either mitigating or inadvertently amplifying the democratic risks posed by AI-driven disinformation in these vulnerable electoral contexts?

The principal novelty of this review resides in synthesizing three traditionally siloed fields: AI-driven disinformation, electoral integrity, and platform governance. Rather than reductively conceptualizing synthetic media as mere technical disruptions, this study firmly situates them within broader debates on democratic accountability and electoral legitimacy. Crucially, it foregrounds the acute vulnerabilities endemic to emerging democracies. In these complex settings, extreme platform dependence, regulatory deficits, and asymmetric civic resilience generate profound political effects inadequately captured by Global North paradigms. By constructing a conceptual bridge between information disorder and democratic resilience, this synthesis enhances both theoretical coherence and pragmatic policy relevance. Finally, Indonesia serves as a vital heuristic reference point, empirically demonstrating how disinformation proliferating across networks like WhatsApp and TikTok fundamentally dictates electoral trust and shapes future subnational democratic contests

RESEARCH METHODS

This study employs a structured qualitative literature review to examine how AI-driven disinformation affects electoral integrity in emerging democracies. The methodological choice is appropriate because the field is still conceptually fragmented across misinformation studies, electoral integrity research, and platform or AI governance, while the article's two research questions require an integrative synthesis of forms, causal mechanisms, democratic risks, and governance responses. Rather than measuring a single causal effect, the review is designed to map and interpret how the existing literature explains the relationship between deepfakes, synthetic media, automated amplification, liar's dividend dynamics, trust in electoral institutions, voting behaviour, and the legitimacy of election outcomes. In line with the manuscript's analytical scope, the review therefore treats electoral integrity as both a procedural and an epistemic issue, and approaches the literature as a body of evidence requiring comparative and interpretive synthesis rather than simple aggregation.

The evidentiary base for the review consists of the studies included in Table 1 of the manuscript, which were selected on the basis of their substantive relevance to one or both research questions. The corpus covers a wide methodological range, including online experiments, survey-based studies, computational analyses, mixed-method and panel studies, conceptual articles, and policy-oriented governance analyses. This breadth was retained because the topic cannot be adequately understood through one design alone: experimental studies help clarify micro-level mechanisms such as authenticity confusion and false

recognition, while computational, conceptual, and policy studies are more useful for explaining platform dynamics, regulatory responses, and broader democratic implications. For each study, the review extracted and organized key information on country or political context, study design and data source, principal analytical focus, central findings, and direct relevance to the review's concerns with AI-driven disinformation, electoral integrity, and governance responses.

The analytical procedure relied on thematic comparative synthesis. After close reading of the selected studies, the evidence was grouped into three main dimensions: first, the principal forms of AI-driven disinformation identified across the literature; second, the causal mechanisms through which these forms affect trust, perception, voting-related attitudes, and electoral legitimacy; and third, the platform, regulatory, and civic responses proposed or evaluated in relation to these risks. The analysis then compared convergences, divergences, and limitations across the corpus, with particular attention to the uneven geographical distribution of evidence. As the reviewed literature is concentrated largely in advanced democracies, while direct evidence from emerging democracies remains more limited, the review adopts a cautious interpretive stance when drawing implications for electorally vulnerable contexts. This approach allows the article to identify robust cross-study patterns while also preserving the contextual limits of the current evidence base.

RESULTS AND DISCUSSION

1. Characteristics of the Reviewed Studies

This subsection provides an overview of the evidentiary base used in this review before moving to the thematic synthesis of findings. It maps the 17 reviewed studies in terms of their empirical settings, research designs, data sources, and substantive relevance to AI-driven disinformation, electoral integrity, and governance responses.

Theme	Study	Context & Method	Key Findings
Causal Mechanisms	Vaccari & Chadwick (2020)	UK; experiment	Deepfakes increase uncertainty and reduce trust in news.
	Dobber et al. (2021)	Netherlands; experiment	Exposure to deepfakes lowers evaluations of political candidates.
	Ahmed (2023)	US; survey	Deepfake exposure increases skepticism, especially among low cognitive ability individuals.
	Köbis et al. (2021)	UK; experiment	Low detection accuracy and high overconfidence (<i>authenticity bias</i>).
	Farooq & de Vreese (2025)	UK; experiment	Higher realism increases perceived authenticity of AI content.

	Schiff et al. (2025)	US; experiment	<i>Liar's dividend</i> : politicians can dismiss real evidence as misinformation.
Democratic Risks	Pawelec (2022)	Cross-national; conceptual	Deepfakes erode trust, increase polarization, and disrupt epistemic order.
	Obioha-Val et al. (2025)	Cross-national; quantitative	Deepfake incidents correlate with declining public sentiment.
	Murphy et al. (2025)	Ireland; experiment	Deepfakes induce false memories and affect voting intention.
	Ferrara (2017)	France; big data	Bots amplify electoral disinformation, though with limited penetration.
Electoral Integrity & Behavior	Mont'Alverne et al. (2024)	Brazil; mixed methods	Trust in news reduces misinformation belief, but misinformation also erodes trust.
	Hwang et al. (2021)	South Korea; experiment	Deepfakes increase persuasiveness and sharing intention.
	Murphy et al. (2025)	Ireland; experiment	Demonstrates direct effects on voting intention.
Governance Responses	Diakopoulos & Johnson (2021)	US; ethical analysis	Emphasizes literacy, verification, and platform moderation.
	Romanishyn et al. (2025)	Policy review	Highlights regulatory and civic resilience strategies.
	Obioha-Val et al. (2025)	Cross-national; quantitative	Enforcement capacity is key to effective regulation.
	Mansur (2025)	Comparative qualitative	Proposes layered democratic resilience strategies.
	Hwang et al. (2021)	Experiment	Media literacy reduces deepfake effects.

As shown in Table 1, the reviewed literature is methodologically diverse, covering online experiments, survey-based studies, computational analyses, panel designs, conceptual and policy-oriented studies, and comparative governance analyses. However, the corpus remains unevenly distributed across contexts, with a strong concentration on advanced democracies such as the United Kingdom, the United States, the Netherlands, France, Ireland, and South Korea, while evidence directly grounded in emerging democracies is still more limited, with Brazil and a small number of cross-national discussions providing the main points of reference. Table 1 also shows that the literature is more developed in examining deepfakes, synthetic media, trust erosion, and epistemic uncertainty than in systematically evaluating institutional and regulatory responses in emerging electoral settings. These characteristics are important

because they shape both the strengths and the limits of the review's subsequent discussion of causal mechanisms, democratic risks, and governance responses

2. Forms of AI-Driven Disinformation and Causal Mechanisms Affecting Electoral Integrity in Emerging Democracies

Across the reviewed studies, AI-driven disinformation appears in several interrelated forms rather than as a single technological artifact. The most visible form is the political deepfake video, examined in experimental and conceptual work as a vehicle for fabricated candidate behavior, reputational attack, or false endorsement (Vaccari & Chadwick, 2020; Dobber et al., 2021; Diakopoulos & Johnson, 2021; Murphy et al., 2025). This is complemented by synthetic audio and broader audio-visual manipulation, including voice cloning, deepfake clips, and cheap but convincing hybrid manipulations that blur distinctions between high-end synthetic media and simpler falsifications (Pawelec, 2022; Murphy et al., 2025). A second form is AI-generated visual content, especially realistic disinformation images whose persuasive force lies less in explicit argument than in perceived authenticity (Farooq & de Vreese, 2025). A third form is AI-generated political messaging and synthetic persona construction, discussed particularly in the governance-oriented literature as part of a broader generative propaganda environment in which fabricated avatars, fake identities, and scalable persuasive content can be integrated into political communication ecosystems (Romanishyn et al., 2025; Mansur, 2025). A fourth form concerns automated amplification, where bot-supported dissemination, network coordination, and platform-scale recirculation expand the reach and apparent salience of manipulated content, even when the originating artifact is not itself a deepfake (Ferrara, 2017; Romanishyn et al., 2025). Finally, the literature increasingly treats liar's dividend dynamics as a distinct form of informational harm, because the political consequences do not depend only on fake content being believed, but also on authentic evidence becoming deniable once publics know that convincing falsification is possible (Pawelec, 2022; Schiff et al., 2025).

The causal mechanisms identified across the corpus suggest that the main danger is not always direct persuasion in the narrow sense of changing preferences after exposure. Instead, the literature points first to authenticity confusion and evidentiary uncertainty. Vaccari and Chadwick (2020) show that synthetic political video may increase uncertainty even when it does not strongly increase outright deception, while Köbis et al. (2021) demonstrate that many viewers are poor at detecting deepfakes and simultaneously overestimate their own detection ability. Farooq and de Vreese (2026) extend this logic by showing that visual realism itself can push viewers toward judging AI-generated images as authentic, often with reduced confidence, while dependence on external detection tools becomes part of the interpretive process. Together, these studies suggest that AI-driven disinformation works by destabilizing the evidentiary status of what citizens see and hear. This matters electorally because campaign communication, candidate conduct, and allegations about electoral malpractice all rely on media artifacts being treated as credible evidence. Once that evidentiary baseline is weakened,

voters are more likely to misrecognize fabricated claims, suspend judgment, or treat contested information as impossible to verify.

A second mechanism is false recognition and misperception, which links synthetic media to candidate evaluation and voting-related judgments. Dobber et al. (2021) find that political deepfake exposure can lower evaluations of the depicted politician, even if party-level effects are narrower and conditioned by predispositions. Murphy et al. (2025) similarly show that homemade political deepfakes can generate false memories for fabricated scandals and, in some conditions, reduce reported voting intention, although the effects are not always stronger than text or synthetic audio. This is important because it indicates that AI-driven disinformation does not need to be uniquely irresistible to be politically consequential. Even mixed or modest persuasive effects can matter when they operate in high-salience electoral moments, target candidate credibility, or reinforce pre-existing affective and partisan biases. Ahmed (2023) adds that deepfake exposure and concern are associated with greater news skepticism, particularly among individuals with lower cognitive resources, which implies that the harms of synthetic media are socially differentiated rather than evenly distributed. The broader implication is that electoral distortion may occur through cumulative misperception, reputational contamination, and heightened suspicion, not simply through mass belief conversion.

A third mechanism, and arguably the most consequential for electoral integrity, is erosion of trust in both information environments and democratic institutions. The reviewed literature repeatedly shows that AI-driven disinformation can corrode confidence in the media, in the reliability of political evidence, and in procedures that depend on shared factual baselines (Vaccari & Chadwick, 2020; Pawelec, 2022; Obioha-Val et al., 2025). Mont'Alverne et al. (2024) provide the clearest electoral evidence here. In Brazil's 2022 presidential election, belief in electoral misinformation about voting procedures and institutions was tied to a broader nexus among news consumption, trust in news, and electoral trust. Their findings are especially important because they connect misinformation belief not just to mistaken factual judgments, but to confidence in how elections are administered and interpreted. In this sense, AI-driven disinformation threatens electoral integrity not only by persuading voters about candidates, but by delegitimizing the fairness, transparency, and credibility of the electoral process itself. Diakopoulos and Johnson (2021) anticipate similar harms through scenarios involving deception, intimidation, reputational harm, and undermining trust in electoral processes, while Pawelec (2022) argues that the most serious democratic damage may lie in epistemic disruption, weakened deliberation, and declining legitimacy of collective decisions. These mechanisms help explain why the effects may manifest as voter confusion, demobilization, or cynical withdrawal in some settings, and as polarizing mobilization in others.

The literature also makes clear that liar's dividend dynamics intensify these risks by converting the existence of synthetic media into a generalized political defense strategy. Schiff et al. (2025) show that politicians can sometimes preserve support by falsely claiming that incriminating information is misinformation, especially when the underlying evidence is text-based rather than audiovisual. Pawelec (2022) similarly treats the liar's dividend as a downstream consequence of informational trust decay, while Romanishyn et al. (2025) emphasize that AI-driven bot networks and synthetic content can produce a wider environment in which citizens struggle to distinguish genuine from manufactured consensus. Ferrara's (2017) analysis of MacronLeaks, although focused on bot amplification rather than generative AI, is useful here because it demonstrates how coordinated dissemination can magnify disinformation visibility and create transnational cascades that threaten electoral discourse even when direct vote effects remain limited. Overall, the reviewed scholarship is increasingly strong in identifying the principal forms and mechanisms through which AI-driven disinformation can affect electoral integrity, trust in institutions, voting behaviour, and the legitimacy of outcomes. Yet the evidence genuinely rooted in emerging democracies remains limited. Brazil provides the clearest direct electoral case, South Korea offers adjacent experimental evidence, and cross-national policy studies discuss cases such as India, Taiwan, and other non-Western settings, but much of the causal evidence still comes from Europe and North America. The literature therefore explains the problem more convincingly than it documents it in emerging democratic contexts, leaving a clear imbalance between the scale of the risk and the geographical distribution of robust empirical evidence.

3.3. Platform Regulation, AI Governance, and Institutional Responses to Democratic Risks in Emerging Electoral Contexts

The reviewed literature assesses responses to AI-driven disinformation as a multi-level governance problem in which platform regulation, public regulation, and civic resilience interact, often unevenly, rather than as a problem solvable by any single intervention. At the platform level, the core responses discussed across the studies include content moderation, labeling, detection systems, provenance verification, and rules governing the circulation of manipulated political media (Diakopoulos & Johnson, 2021; Romanishyn et al., 2025; Obioha-Val et al., 2025; Mansur, 2025). Yet the literature is cautious about treating these measures as sufficient. Platform interventions may slow exposure, reduce recirculation, or increase friction around suspicious content, but their effectiveness depends on timing, consistency, and the degree to which users trust both the warning systems and the institutions behind them. Farooq and de Vreese (2025) show that AI detection tools do not operate in a neutral vacuum, because users' judgments remain shaped by visual realism and their own confidence. Likewise, Köbis et al. (2021) suggest that user overconfidence in deepfake detection limits the simple assumption that awareness or prompts will meaningfully improve resistance. In this sense,

platform regulation can mitigate risk, but it cannot fully resolve the deeper problem of authenticity confusion and evidentiary instability already identified in the broader literature.

The platform-level evidence also shows that governance failure can itself amplify vulnerability. Ferrara (2017) demonstrates how automated amplification and network coordination can increase the visibility and reach of manipulative content during election periods, even when direct persuasion effects remain uncertain. Romanishyn et al. (2025) and Mansur (2025) similarly argue that bots, synthetic identities, and scalable AI-generated content allow disinformation to circulate as part of a broader information environment rather than as isolated false artifacts. This matters because democratic harm is intensified when platforms do not merely host manipulated content but structurally facilitate its diffusion through recommendation systems, virality logics, and inconsistent moderation. Diakopoulos and Johnson (2020) therefore frame platform responsibility not only in terms of removing falsehoods, but also in terms of verification, moderation standards, and anticipatory harm reduction in election-sensitive contexts. Across the studies, then, platform accountability is treated as necessary but uneven, especially where commercial incentives, moderation opacity, and transnational information flows undermine consistent enforcement.

At the state and regulatory level, the literature places growing emphasis on AI governance, legal frameworks, and institutional enforcement capacity. Obioha-Val et al. (2025) provide the strongest direct argument that regulatory capacity is decisive, finding that enforcement strength matters more than formal regulatory breadth alone in shaping the effectiveness of responses to deepfakes and related harms. Romanishyn et al. (2025) reach a similar conclusion in policy terms, arguing that fragmented or purely declarative regulation is insufficient unless backed by adaptive institutional coordination, monitoring systems, and cross-border cooperation. Mansur (2025) likewise underscores that governance responses to generative AI and algorithmic manipulation require layered institutional design rather than isolated legal prohibition. The implication across these studies is that regulation matters less as symbolic commitment than as operational capacity: states need not only rules, but also agencies, technical expertise, inter-institutional coordination, and credible enforcement mechanisms. Where these are weak, regulatory responses may exist on paper while democratic vulnerability persists in practice.

This point is especially important for emerging electoral contexts, where the literature repeatedly implies structural asymmetries between risk exposure and governance capacity. Although direct empirical evidence from emerging democracies remains limited, the reviewed corpus suggests that such contexts may face sharper vulnerability because they often combine high platform dependence, weaker enforcement capacity, unequal digital literacy, and already fragile trust in electoral institutions (Mont'Alverne et al., 2024; Romanishyn et al., 2025; Mansur, 2025). The Brazilian case analysed by Mont'Alverne et al. (2024) is particularly instructive because it shows that electoral misinformation interacts with trust in news and

confidence in election-related information environments. This implies that institutional response cannot be reduced to technical content management alone. Where trust in electoral administration is already contested, even accurate corrective measures may be interpreted through partisan suspicion. In such environments, governance failure is not only the absence of regulation, but also the inability of institutions to command credibility when interventions occur.

The societal and civic resilience literature adds an important but qualified layer to this assessment. Hwang et al. (2021) provide the strongest experimental support for digital literacy, showing that media literacy interventions can reduce the persuasive and sharing effects of deepfake-enhanced disinformation. Diakopoulos and Johnson (2021) also identify education and public awareness as part of the necessary response toolkit, while Romanishyn et al. (2025) place civic education and cognitive resilience within a broader democratic defense strategy. At the same time, the reviewed studies do not support a strong claim that literacy alone can solve the problem. Köbis et al. (2021) show that simple awareness does not reliably improve deepfake detection, and Vaccari and Chadwick (2020) suggest that synthetic media may generate uncertainty even without convincing large numbers of citizens that false content is true. This means that fact-checking and literacy may help, but often by moderating susceptibility at the margins rather than by restoring a stable epistemic environment. In highly polarized or low-trust settings, corrections may arrive too late, may fail to travel as far as the original content, or may themselves become politicized.

The literature is also increasingly attentive to the interaction between governance instruments and public trust. Schiff et al. (2025) show that liar's dividend dynamics complicate corrective strategies because politicians can sometimes exploit public awareness of misinformation to deny authentic evidence and preserve support. Pawelec (2022) similarly argues that the democratic threat of deepfakes lies not only in deception but in the broader corrosion of shared standards of proof and accountability. This is why detection tools, verification labels, and fact-checking mechanisms are not automatically effective. Their impact depends on whether audiences regard them as credible, whether platforms apply them consistently, and whether electoral institutions themselves are trusted enough to anchor corrective communication. In low-trust environments, technological interventions may even intensify suspicion by confirming to citizens that nothing can be taken at face value.

Overall, the reviewed studies suggest that responses to AI-driven disinformation can mitigate democratic risks, but usually only partially and conditionally. Platform governance remains uneven, detection systems are helpful but not self-sufficient, fact-checking and digital literacy have real but limited effects, and state regulation matters most where enforcement capacity is credible and sustained (Hwang et al., 2021; Obioha-Val et al., 2025; Farooq & de Vreese, 2025; Romanishyn et al., 2025). For emerging electoral contexts, the evidence base is still thinner

than the scale of the problem requires. Brazil offers the clearest directly relevant case, while other discussions are often comparative, anticipatory, or extrapolated from non-emerging democratic settings. Even so, the literature converges on one central conclusion: democratic resilience against AI-driven disinformation depends not only on technological intervention, but on the interaction among institutional credibility, regulatory capacity, and platform governance. Where those three elements are weak or misaligned, governance responses are likely to remain reactive, fragmented, and insufficient to protect electoral integrity.

3.4. Comparative Synthesis, Research Gaps, and Future Directions

Taken together, the 17 reviewed studies suggest a consistent cross-study pattern: AI-driven disinformation affects electoral politics less through straightforward vote conversion than through uncertainty, distrust, and wider epistemic instability. Across experimental, conceptual, computational, and policy-oriented studies, the most recurrent finding is that synthetic media, bot-supported dissemination, and liar's dividend dynamics destabilize citizens' confidence in what counts as credible political evidence (Vaccari & Chadwick, 2020; Köbis et al., 2021; Pawelec, 2022; Schiff et al., 2025). This destabilization matters because it links manipulated content not only to misperception, but also to reduced trust in media, greater skepticism toward political information, and broader doubts about democratic procedures and public accountability (Ahmed, 2023; Obioha-Val et al., 2025; Mont'Alverne et al., 2024). The literature therefore indicates that the most consistent democratic effect is not direct proof of changed vote choice, but rather erosion of the informational conditions under which electoral legitimacy is sustained.

A second comparative pattern is that evidence on electoral legitimacy and institutional trust appears more robust than evidence on concrete behavioural change. Some studies do find effects on candidate evaluation, false memory formation, or reported voting intention (Dobber et al., 2021; Murphy et al., 2025), but these effects are often conditional, mixed, or smaller than public debate sometimes assumes. By contrast, uncertainty, credibility distortion, and declining trust recur much more consistently across the corpus (Vaccari & Chadwick, 2020; Farooq & de Vreese, 2025; Mont'Alverne et al., 2024). This suggests that AI-driven disinformation is especially consequential where electoral integrity depends on shared confidence in evidence, procedures, and outcome acceptance. In that respect, the reviewed studies collectively shift attention from persuasion alone toward a broader account of democratic vulnerability shaped by epistemic disruption, cynical withdrawal, and polarizing suspicion (Pawelec, 2022; Diakopoulos & Johnson, 2021; Romanishyn et al., 2025).

A third pattern concerns governance responses. Across the literature, platform regulation, fact-checking, digital literacy, and regulatory interventions are treated as necessary but largely reactive, partial, and unevenly tested. Media literacy can reduce the persuasive effects of deepfake-enhanced disinformation, but its protective power is neither automatic nor universal (Hwang et al., 2021; Köbis et al., 2021). Detection tools may assist judgment, yet their usefulness depends heavily on public trust, interpretive context, and user confidence rather than on technical performance alone (Farooq & de Vreese, 2025). At the regulatory level, the strongest recurring conclusion is that institutional capacity matters more than formal policy aspiration: enforcement strength, coordination, and credibility shape whether governance responses meaningfully mitigate democratic risks or remain largely symbolic (Obioha-Val et al., 2025; Romanishyn et al., 2025; Mansur, 2025). Platform accountability, meanwhile, remains inconsistent, particularly where amplification logics and moderation opacity continue to structure exposure pathways (Ferrara, 2017; Diakopoulos & Johnson, 2021).

These patterns also reveal several clear research gaps. First, the evidence base remains dominated by studies from advanced democracies, especially the United States and Europe, with only limited direct evidence from emerging democracies. Brazil provides the strongest electoral case in this corpus, while South Korea appears in experimental work adjacent to electoral concerns, but the broader geography of emerging democratic vulnerability remains underdeveloped (Mont'Alverne et al., 2024; Hwang et al., 2021). Second, the literature still focuses disproportionately on deepfakes, often at the expense of other forms of AI-driven disinformation such as synthetic political messaging, hybrid manipulation, and AI-assisted amplification infrastructures (Pawelec, 2022; Romanishyn et al., 2025). Third, empirical evaluation of governance responses remains thin. Fact-checking, platform moderation, literacy interventions, and electoral preparedness are frequently recommended, but much less often tested under realistic electoral conditions. Fourth, there is still weak integration between misinformation studies, electoral integrity scholarship, and platform or AI governance research, which often proceed in parallel rather than as a shared analytical field.

Future research should therefore move in several directions. Comparative work across emerging democracies is especially needed to clarify how AI-driven disinformation interacts with weaker enforcement environments, uneven digital literacy, and fragile electoral trust. Greater attention should also be given to closed or semi-private platforms such as WhatsApp and similar messaging spaces, which are strikingly underexamined in this corpus despite their likely importance in real electoral communication flows. In addition, future studies should more directly evaluate whether fact-checking, literacy interventions, platform safeguards, and electoral preparedness measures actually reduce democratic harm under campaign conditions rather than only in abstract or controlled settings. Finally, stronger theoretical integration is needed among information disorder, epistemic integrity, and democratic resilience so that

synthetic media are analyzed not as isolated technological novelties, but as mechanisms embedded in broader struggles over trust, proof, accountability, and outcome legitimacy.

The contribution of this review lies in bringing these strands together within a more integrated framework. Rather than treating deepfakes, electoral misinformation, and governance responses as separate conversations, the review connects the main forms of AI-driven disinformation, the causal mechanisms through which they operate, the democratic risks they generate, and the institutional responses proposed to address them. In doing so, it shows that the central problem is not only whether AI-driven disinformation changes minds directly, but whether it degrades the epistemic and institutional foundations on which electoral integrity in emerging democracies ultimately depends.

4. Conclusion

This review shows that AI-driven disinformation poses a democratic threat that extends beyond false content itself. Across the 17 reviewed studies, the most consistent pattern is that deepfakes, synthetic media, automated amplification, and liar's dividend dynamics disrupt the evidentiary foundations of electoral politics by generating uncertainty, confusion, misperception, and distrust. While some studies indicate effects on candidate evaluation, false memory formation, and reported voting intention, the broader and more robust consequence is the erosion of trust in information environments, electoral procedures, and the legitimacy of outcomes. In this sense, the main risk is not simply persuasion, but the weakening of the epistemic conditions that allow citizens to judge evidence, hold political actors accountable, and accept election results as credible.

The review also finds that governance responses remain necessary but incomplete. Platform moderation, detection tools, fact-checking, digital literacy, and regulatory interventions can mitigate risk, yet their effectiveness is conditional on enforcement capacity, institutional credibility, and public trust. This is especially significant for emerging democracies, where high platform dependence, uneven digital literacy, and fragile confidence in electoral institutions may intensify vulnerability while limiting response capacity. The central contribution of this article is therefore to connect the forms of AI-driven disinformation, their causal mechanisms, their democratic risks, and the governance responses they provoke within a single analytical framework. At the same time, the review underscores a major scholarly imbalance: the problem is increasingly well theorised, but empirical evidence rooted directly in emerging democratic contexts remains too limited relative to the scale and urgency of the threat.

REFERENCES

- Ahmed, S., & Ahmed, Saifuddin. (2023). Navigating the maze: Deepfakes, cognitive ability, and social media news skepticism. *New Media & Society*, 25(5), 1108–1129. <https://doi.org/10.1177/14614448211019198>
- Alain, G., Crick, J., Snead, E., Quatman-Yates, C., & Quatman, C. (2025). Evaluating User Interactions and Adoption Patterns of Generative AI in Health Care Occupations Using Claude: Cross-Sectional Study. *Journal of Medical Internet Research*, 27. <https://doi.org/10.2196/73918>
- Amin, M. (2025). Pengaruh Algoritma Media Sosial Terhadap Narasi keislaman di Ruang Digital. *El-Wasathy: Journal of Islamic Studies*. <https://doi.org/10.61693/elwasathy.vol31.2025.148-158>
- Bär, D., Pierri, F., De Francisci Morales, G., & Feuerriegel, S. (2023). Systematic discrepancies in the delivery of political ads on Facebook and Instagram. *PNAS Nexus*, 3. <https://doi.org/10.1093/pnasnexus/pgae247>
- Bennett, W., Livingston, S., Bennett, L., & Carroll, L. (2018). The disinformation order: Disruptive communication and the decline of democratic institutions. *European Journal of Communication*, 33, 122–139. <https://doi.org/10.1177/0267323118760317>
- Bhattacharya, S., Agarwal, N., & Poudel, D. (2024). Analyzing the impact of symbols in Taiwan's election-related anti-disinformation campaign on TikTok. *Social Network Analysis and Mining*, 14. <https://doi.org/10.1007/s13278-024-01385-9>
- Bontridder, N., & Poulet, Y. (2021). The role of artificial intelligence in disinformation. *Data & Policy*, 3. <https://doi.org/10.1017/dap.2021.20>
- Broda, E., & Strömbäck, J. (2024). Misinformation, disinformation, and fake news: lessons from an interdisciplinary, systematic literature review. *Annals of the International Communication Association*, 48, 139–166. <https://doi.org/10.1080/23808985.2024.2323736>
- Casero-Ripollés, A., Alonso-Muñoz, L., & Moret-Soler, D. (2025). Spreading False Content in Political Campaigns: Disinformation in the 2024 European Parliament Elections. *Media and Communication*. <https://doi.org/10.17645/mac.9525>
- Corsi, G., Marino, B., & Wong, W. (2024). The spread of synthetic media on X. *Harvard Kennedy School Misinformation Review*. <https://doi.org/10.37016/mr-2020-140>
- Diakopoulos, N., & Johnson, D. (2021). Anticipating and addressing the ethical implications of deepfakes in the context of elections. *New Media & Society*, 23(7), 2072–2098. <https://doi.org/10.1177/1461444820925811>
- Dobber, T., Metoui, N., Trilling, D., Helberger, N., & De Vreese, C. (2021). Do (Microtargeted) Deepfakes Have Real Effects on Political Attitudes? *The International Journal of Press/Politics*, 26, 69–91. <https://doi.org/10.1177/1940161220944364>
- Evangelista, R., & Bruno, F. (2019). WhatsApp and political instability in Brazil: targeted messages and political radicalisation. *Internet Policy Rev.*, 8. <https://doi.org/10.14763/2019.4.1434>

- Farooq, A., & De Vreese, C. (2025). Deciphering authenticity in the age of AI: how AI-generated disinformation images and AI detection tools influence judgements of authenticity. *AI & SOCIETY*. <https://doi.org/10.1007/s00146-025-02416-5>
- Ferrara, E. (2017). Disinformation and Social Bot Operations in the Run Up to the 2017 French Presidential Election. *Interorganizational Networks & Organizational Behavior EJournal*. <https://doi.org/10.5210/fm.v22i8.8005>
- Ferrara, E., Chang, H., Chen, E., Murić, G., & Patel, J. (2020). Characterizing social media manipulation in the 2020 U.S. presidential election. *First Monday*, 25. <https://doi.org/10.5210/fm.v25i11.11431>
- Fisher, J., Gadjanova, E., & Hitchen, J. (2023). WhatsApp and political communication in West Africa: Accounting for differences in parties' organization and message discipline online. *Party Politics*, 30, 934–948. <https://doi.org/10.1177/13540688231188690>
- Greenberg-Worisek, A., Kurani, S., Rutten, L. F., Blake, K., Moser, R., & Hesse, B. (2019). Tracking Healthy People 2020 Internet, Broadband, and Mobile Device Access Goals: An Update Using Data From the Health Information National Trends Survey. *Journal of Medical Internet Research*, 21. <https://doi.org/10.2196/13300>
- Henke, J. (2024). Navigating the AI era: university communication strategies and perspectives on generative AI tools. *Journal of Science Communication*. <https://doi.org/10.22323/2.23030205>
- Humprecht, E., Esser, F., & Van Aelst, P. (2020). Resilience to Online Disinformation: A Framework for Cross-National Comparative Research. *The International Journal of Press/Politics*, 25, 493–516. <https://doi.org/10.1177/1940161219900126>
- Hwang, Y., Ryu, J. Y., & Jeong, S.-H. (2021). Effects of Disinformation Using Deepfake: The Protective Effect of Media Literacy Education. *Cyberpsychology, Behavior, and Social Networking*, 24(3), 188–193. <https://doi.org/10.1089/cyber.2020.0174>
- Köbis, N. C., Doležalová, B., & Soraperra, I. (2021). Fooled twice: People cannot detect deepfakes but think they can. *IScience*, 24(11), 103364. <https://doi.org/https://doi.org/10.1016/j.isci.2021.103364>
- Lava-Santos, D., José, G.-R., & Llorca-Abad, G. (2023). Crude, anonymous, partisan, sectoral and anti-elitist. Electoral disinformation in Spain (2019-2022). *El Profesional de La Información*. <https://doi.org/10.3145/epi.2023.sep.06>
- Lorenz-Spreen, P., Oswald, L., Lewandowsky, S., & Hertwig, R. (2022). A systematic review of worldwide causal and correlational evidence on digital media and democracy. *Nature Human Behaviour*, 7, 74–101. <https://doi.org/10.1038/s41562-022-01460-1>
- Mansur, M. A. (2025). AI and Cyber-Enabled Threats to Democracy through Algorithmic Manipulation and Generative AI in Undermining Democratic Integrity. *European Scientific Journal ESJ*. <https://doi.org/10.19044/esipreprint.8.2025.p658>
- Mauk, M., & Grömping, M. (2023). Online Disinformation Predicts Inaccurate Beliefs About Election Fairness Among Both Winners and Losers. *Comparative Political Studies*, 57,

965–998. <https://doi.org/10.1177/00104140231193008>

- Momeni, M. (2024). Artificial Intelligence and Political Deepfakes: Shaping Citizen Perceptions Through Misinformation. *Journal of Creative Communications*, 20, 41–56. <https://doi.org/10.1177/09732586241277335>
- Mont’Alverne, C., Arguedas, A. R., Banerjee, S., Toff, B., Fletcher, R., & Nielsen, R. K. (2024). The Electoral Misinformation Nexus: How News Consumption, Platform Use, and Trust in News Influence Belief in Electoral Misinformation. *Public Opinion Quarterly*, 88, 681–707. <https://doi.org/10.1093/poq/nfae019>
- Murphy, G., Ching, D., Meehan, E., Twomey, J., Bolger, A., & Linehan, C. (2025). An Average Joe, a Laptop, and a Dream: Assessing the Potency of Homemade Political Deepfakes. *Applied Cognitive Psychology*. <https://doi.org/10.1002/acp.70061>
- Obioha-Val, O., Gbadebo, M., Olaniyi, O., Chinye, N. C., & Balogun, A. (2025). Innovative Regulation of Open Source Intelligence and Deepfakes AI in Managing Public Trust. *Journal of Engineering Research and Reports*. <https://doi.org/10.9734/jerr/2025/v27i21400>
- Pawelec, M. (2022). Deepfakes and Democracy (Theory): How Synthetic Audio-Visual Media for Disinformation and Hate Speech Threaten Core Democratic Functions. *Digital Society*, 1. <https://doi.org/10.1007/s44206-022-00010-6>
- Pedriz, S. B. (2021). Sources, Channels and Strategies of Disinformation in the 2020 US Election: Social Networks, Traditional Media and Political Candidates. *Journalism and Media*. <https://doi.org/10.3390/journalmedia2040036>
- Prochaska, S., Duskin, K., Kharazian, Z., Minow, C., Blucker, S., Venuto, S., West, J., & Starbird, K. (2023). Mobilizing Manufactured Reality: How Participatory Disinformation Shaped Deep Stories to Catalyze Action during the 2020 U.S. Presidential Election. *Proceedings of the ACM on Human-Computer Interaction*, 7, 1–39. <https://doi.org/10.1145/3579616>
- Romănishyn, A., Malytska, O., & Goncharuk, V. (2025). AI-driven disinformation: policy recommendations for democratic resilience. *Frontiers in Artificial Intelligence*, 8. <https://doi.org/10.3389/frai.2025.1569115>
- Russo, D. (2023). Navigating the Complexity of Generative AI Adoption in Software Engineering. *ACM Transactions on Software Engineering and Methodology*, 33, 1–50. <https://doi.org/10.1145/3652154>
- SCHIFF, K. J., SCHIFF, D. S., & BUENO, N. S. (2025). The Liar’s Dividend: Can Politicians Claim Misinformation to Evade Accountability? *American Political Science Review*, 119(1), 71–90. <https://doi.org/DOI: 10.1017/S0003055423001454>
- Schipper, T. (2025). Disinformation by design: leveraging solutions to combat misinformation in the Philippines’ 2025 election. *Data & Policy*, 7. <https://doi.org/10.1017/dap.2025.18>
- Siahaan, H. S., & Ivanna, J. (2024). Transformasi Partisipasi Politik Masyarakat Desa melalui E-Voting dalam Pemilihan Kepala Desa di Desa Sibolangit, Kecamatan Merek. *Journal*

- on Education*. <https://doi.org/10.31004/joe.v6i4.6143>
- Simbolon, P. A. L., & Ivanna, J. (2024). Komunikasi Politik Pemerintahan Desa dalam Meningkatkan Kinerja Pelayanan Masyarakat Raya Palka Sindangsari. *Journal on Education*. <https://doi.org/10.31004/joe.v6i4.6341>
- Utami, O. N., Setiawati, K. E., Dina, E. S., Hafida, M., & Salsabila, U. H. (2023). Perkembangan Teknologi Informasi Dan Komunikasi Dalam Pembelajaran Pendidikan Agama Islam. *Jurnal Pendidikan Islam*. <https://doi.org/10.37286/ojs.v9i1.218>
- Vaccari, C., & Chadwick, A. (2020). Deepfakes and Disinformation: Exploring the Impact of Synthetic Political Video on Deception, Uncertainty, and Trust in News. *Social Media + Society*, 6. <https://doi.org/10.1177/2056305120903408>
- Vandito, R. L., Paramesti, A., Aminudin, M. A. V., & Nurdin, N. (2025). KONSEP E-DEMOCRACY PADA TATA KELOLA PEMERINTAHAN: STUDI KASUS PENERAPAN APLIKASI TANGERANG LIVE DALAM MEMBUKA PARTISIPASI PUBLIK. *Governance*. <https://doi.org/10.33558/governance.v13i1.10502>
- Wildan, W. (2025). Dari Manuskrip ke Digital: Transformasi Studi Tafsir di Era Digital (Studi Analisis Website Altafsir.com). *El-Wasathy: Journal of Islamic Studies*. <https://doi.org/10.61693/elwasathy.vol31.2025.99-113>
- Williams, A., Burke-Moore, L., Chan, R. S.-Y., Enock, F., Nanni, F., Sippy, T., Chung, Y.-L., Gabasova, E., Hackenburg, K., & Bright, J. (2024). Large language models can consistently generate high-quality content for election disinformation operations. *PLOS One*, 20. <https://doi.org/10.48550/arxiv.2408.06731>
- Technology Australian Geothermal Energy Conference*, Canberra, Australia: Geoscience Australia. Diakses dari http://www.ga.gov.au/image_cache/GA11825.pdf