Influence of AI-Generated Deepfakes and Synthetic Media on Scientific and Public Consensus: A Case Study of Taraba State, Nigeria

Prof. Gladys Uzezi Jack & Waksum, Jerry

Department of Science Education, Faculty of Education, Taraba State University, Jalingo, Nigeria

DOI: https://dx.doi.org/10.47772/IJRISS.2026.100700337

Received: 13 July 2026; Accepted: 18 July 2026; Published: 31 July 2026

ABSTRACT

Artificial-intelligence tools capable of producing photorealistic deepfakes and manipulated scientific or health imagery have advanced faster than the institutional, educational, and regulatory mechanisms meant to detect them, raising the question of how such synthetic content shapes scientific and public consensus. This paper offers a structured narrative synthesis of scholarship spanning 2014 to 2026, drawing particular attention to sources from 2020 through 2026, examining how deepfakes and synthetic media influence trust, consensus acceptance, and unaided human detection ability, and situating global findings within Nigeria's information ecosystem. Sources were identified through academic search platforms (including ScienceDirect, Wiley Online Library, PubMed/PMC, and arXiv) and credible Nigerian fact-checking and legal-analysis outlets, screened for relevance to detection accuracy, consensus, or trust outcomes, and synthesized narratively. Across the reviewed literature, unaided human detection of deepfakes consistently approximates chance level and is frequently accompanied by overconfidence, technical labellings and awareness campaigns alone show limited effect, while structured, feedback-based media-literacy training shows the most consistent improvement. Nigerian evidence indicates that fact-checking organizations rely on forensic software rather than unaided judgement to confirm manipulation, that synthetic content has already been used to inflame ethnic and political tension, and that community-anchored trust structures may matter as much as technical verification capacity. The review concludes that addressing deepfake-driven threats to scientific and public consensus in contexts such as Taraba State, Nigeria, requires coordinated institutional, educational, and community-based responses rather than technological solutions alone, and it offers recommendations for tertiary institutions, health authorities, broadcast media, policymakers, and community structures.

Keywords: deepfakes, synthetic media, scientific consensus, misinformation, digital literacy, science communication, Nigeria

INTRODUCTION

Advances in generative artificial intelligence have reshaped how visual information circulates, enabling the creation of highly realistic synthetic images, audio, and video widely known as deepfakes. Earlier forms of image manipulation, from darkroom retouching to conventional photo-editing software, typically left identifiable traces such as lighting inconsistencies or anatomical implausibility. Contemporary deepfakes generated through generative adversarial networks and diffusion models instead approximate the perceptual thresholds of authentic recordings, making them substantially harder for untrained observers to detect (Goodfellow et al., 2014; Ho et al., 2020). This technical trajectory has outpaced the verification habits, media-literacy training, and regulatory frameworks that societies rely on to sustain shared scientific and public understanding, creating what several reviewers now describe as a crisis of institutional and epistemic trust (UNESCO, 2025).

Nigeria has not been insulated from this development. Fact-checking organizations have documented a growing catalogued of AI-manipulated videos targeting Nigerian public figures, including fabricated clips attributed to the President and other national leaders, and synthetic content designed to inflame ethnic and regional tension ahead of Nigeria's 2027 general elections (Dubawa, 2026; Dubawa/PRNigeria, 2025). Analysts note that

Nigerian fact-checkers themselves require specialized forensic software, rather than unaided observation, to confirm that circulating content is synthetic (Dubawa, 2026), and legal commentators point to weak enforcement mechanisms, low levels of public digital literacy, and a social-media environment that remains largely unregulated as compounding vulnerabilities (Olisa Agbakoba Legal, 2025/2026). Nigeria's regulatory response remains nascent: the National Information Technology Development Agency's draft White Paper on Online Harm Protection and Content Moderation acknowledges the threat of manipulated media but had, at the time of this review, not been converted into a binding legal framework.

Purpose of the Study

This review aims to bring together existing empirical and grey literature examining how deepfakes and synthetic media affect scientific and public consensus, human detection ability, and institutional trust, and to interpret that synthesis for a state-level Nigerian context, specifically Taraba State, a predominantly agrarian state in Nigeria's North-East geopolitical zone where public-health messaging, agricultural-extension advice, and civic communication depend heavily on radio, community gatherings, and increasingly, smartphone-based social media.

Statement of the Problem

While a substantial body of international literature documents that unaided humans struggle to detect deepfakes and that such content erodes trust in political, health, and scientific information generally, this literature is concentrated in Western, urban, or nationally aggregated settings. Comparatively little synthesis exists on how these dynamics apply to sub-national, resource-constrained African contexts where radio, community leadership structures, and low-bandwidth social media dominate information flow, and reviewers in the deepfake-detection field have themselves flagged this as a research gap requiring dedicated attention (Diel et al., 2024; Schlippe et al., 2025). No identified synthesis has yet connected the international detection-accuracy and consensus literature to Nigeria's documented, active deepfake circulation within a sub-national, largely rural context such as Taraba State, and this review is designed to address that gap.

Research Question and Hypothesis

The following questions guided this review:

RQ1: What does the empirical literature indicate about the effect of deepfake exposure on trust and consensus acceptance relative to authentic or conventionally manipulated content?

RQ2: How accurate is unaided human detection of deepfakes, and what factors improve or fail to improve that accuracy?

RQ3: What does the Nigerian and African-context literature specifically indicate about deepfake circulation, institutional verification capacity, and community-level response?

RQ4: What theoretical frameworks best explain why synthetic media proves persuasive and difficult to detect?

In place of statistically tested hypotheses, which would be inappropriate for a literature-synthesis design, this review evaluated the following directional propositions against the weight of the reviewed evidence:

H1: Exposure to deepfakes reduces trust and consensus acceptance relative to authentic content.

H2: Unaided human detection of deepfakes is substantially less accurate than detection of conventionally manipulated content and frequently approximates chance level.

H3: Passive awareness alone is insufficient to meaningfully improve detection accuracy, whereas structured, feedback-based training shows measurable benefit.

H4: Institutional and community-level verification capacity in resource-constrained settings such as Taraba State lags behind the sophistication of circulating synthetic content.

LITERATURE REVIEW

Deepfakes trace their origin to generative adversarial networks, where two competing components — a generator and a discriminator — work together to progressively sharpen synthetic realism (Goodfellow et al., 2014), a process further extended by diffusion-based models that narrow the perceptual gap between authentic and synthetic content (Ho et al., 2020). Empirical detection studies converge on a consistent pattern: unaided observers cannot reliably distinguish deepfakes from authentic material, frequently perform close to chance level, and tend to overestimate their own accuracy even when offered incentives (Köbis et al., 2021; Nightingale & Farid, 2022). Large-scale systematic reviews reinforce this inconsistency across the field, demonstrating that structured training built around feedback — as opposed to passive awareness campaigns or content labelling — yields the most dependable gains in detection accuracy (Diel et al., 2024; Somoray et al., 2025; Groh et al., 2022).

Cross-cultural evidence indicates that contextual familiarity improves recognition of authentic content without necessarily improving detection of fabricated material, and highlights a persistent scarcity of research on African populations (Schlippe et al., 2025). Within Nigeria specifically, misinformation exposure has been found to be nearly universal and linked more to distrust of government than to information gaps (Wonodi et al., 2022), while a community-fellow intervention in neighbouring Niger State demonstrated that training trusted local messengers can measurably improve misinformation detection and response at community level (Erim et al., 2025). Nigerian fact-checking and legal-analysis sources further document active deepfake circulation, limited forensic verification capacity, and an under-regulated information environment (Dubawa, 2026; Dubawa/PRNigeria, 2025; Olisa Agbakoba Legal, 2025/2026).

Three theoretical lenses frame the interpretation of this literature. Media Richness Theory (Daft & Lengel, 1986) explains why visually rich synthetic content is especially persuasive relative to text-based rumour. Information Manipulation Theory (McCornack, 1992) explains how such content succeeds by exploiting the ordinarily reasonable assumption that video or image content is authentic. Berger and Luckmann's (1966) Social Construction of Reality helps explain why, in under-resourced settings, credibility tends to be established through community practice and trusted messengers rather than technical inspection alone. Together, these strands establish that Nigeria's documented deepfake activity has not yet been specialized together with the international detection-accuracy and consensus literature for a Nigerian sub-national context, which is the gap this review addresses.

METHODOLOGY

This review adopted a systematic narrative-synthesis approach guided by the PRISMA 2020 reporting standard (Page et al., 2021), adapted for a non-meta-analytic synthesis given the methodological heterogeneity of the source literature. Searches were conducted across ScienceDirect, Wiley Online Library, PubMed/PMC, arXiv, and Google Scholar using combinations of the terms "deepfake," "synthetic media," "detection accuracy," "misinformation," "scientific consensus," "public trust," and "Nigeria" or "Africa," supplemented by targeted searches of credible Nigerian fact-checking and legal-analysis outlets, including Dubawa, PRNigeria, TheCable Fact Check, and Olisa Agbakoba Legal, to ground the review in the local information ecosystem.

The search prioritized English-language sources published between 2014 and 2026, with emphasis on 2020–2026 given the acceleration of both generative-AI capability and Nigerian misinformation scholarship over this period. Sources were included if they empirically examined human detection of, trust in, or consensus responses to deepfakes or synthetic media, or if they provided credible investigative or legal-analytic evidence of deepfake circulation and institutional response within Nigeria; purely algorithmic detection-method papers with no human or social outcome, and sources lacking verifiable authorship or publication details, were excluded. Screening for relevance, duplication, and credibility yielded twenty-one sources retained for narrative synthesis, fifteen of which are summarized in Table 1; findings were organized thematically around detection accuracy,

persuasiveness and consensus effects, and institutional/community readiness, and interpreted through the theoretical frameworks described above.

RESULTS

Table 1 summarizes the design and key findings of the principal studies retained for synthesis. The reviewed literature clusters around three consistent findings. First, studies employing controlled detection tasks report that unaided human accuracy in spotting deepfakes varies considerably between studies and often hovers near chance level, with participants repeatedly inclined to judge synthetic material as genuine and to overrate their own detection skill (Diel et al., 2024; Köbis et al., 2021; Nightingale & Farid, 2022). Second, studies testing interventions generally agree that passive approaches — basic awareness messaging or content labelling — yield only modest gains in detection accuracy, while structured, interactive training built on feedback produces measurable improvement (Groh et al., 2022; Somoray et al., 2025). Third, the Nigeria-specific and African-context evidence, while smaller in volume, consistently indicates active circulation of synthetic and manipulated content within civic and health information ecosystems, limited institutional verification capacity, and a comparative reliance on community-anchored trust structures over technical verification (Dubawa, 2026; Dubawa/PR Nigeria, 2025; Erim et al., 2025; Olisa Agbakoba Legal, 2025/2026; Wonodi et al., 2022).

Table 1.

Characteristics and key findings of studies reviewed on deepfakes, synthetic media, and consensus, trust, or detection outcomes.

Author(s) and Year	Focus / Design	Key Finding Relevant to Consensus, Trust, or Detection
Goodfellow et al. (2014)	Technical/foundational (generative adversarial networks)	Introduced the generator–discriminator architecture that underpins present-day deepfake creation, paving the way for increasingly convincing synthetic imagery.
Ho et al. (2020)	Technical/foundational (diffusion models)	Extended synthesis quality beyond GANs, further narrowing the perceptual gap between authentic and synthetic visual content.
Köbis, Doležalová, and Soraperra (2021)	Pre-registered behavioural experiment (N = 210)	Showed that participants left to judge unaided could not reliably spot deepfakes; accuracy improved with neither awareness prompts nor financial incentives, and participants consistently overrated their own ability.
Groh, Epstein, Firestone, and Picard (2022)	Large-scale crowd-detection experiment	Showed that human deepfake-detection accuracy can improve with repeated feedback but remains inconsistent without structured training or machine assistance.
Nightingale and Farid (2022)	Controlled experiment (AI-synthesized faces)	Found that AI-generated faces were rated as indistinguishable from, and in some cases more trustworthy than, authentic faces among untrained viewers.
Diel et al. (2024)	Systematic review and meta-analysis (56 studies)	Pooled findings on human deepfake-detection performance across the field, concluding that accuracy fluctuates widely and often sits near chance level, with training-based interventions offering the most dependable gains.
Somoray, Miller, and Holmes (2025)	Systematic review	Concluded that interactive, feedback-based media-literacy training outperforms passive awareness campaigns or content-labelling alone in improving detection outcomes.

Schlippe, Wölfel, and Mabokela (2025)	Cross-cultural detection experiment (South Africa)	Found that contextual/cultural familiarity aided recognition of authentic content but did not reliably improve detection of AI-generated fabrications, and noted a scarcity of African-context research.
Chesney and Citron (2019)	Legal and policy analysis	Framed deepfakes as a structural threat to privacy, democratic discourse, and national security, foreseeing difficulties for institutions attempting to adapt.
UNESCO (2025)	Expert policy commentary	Argued that purely technical verification responses are insufficient on their own and that synthetic content is eroding institutional and epistemic trust globally.
Wonodi et al. (2022)	Mixed-methods study (22 focus groups, 24 key informant interviews; six Nigerian geopolitical zones)	Found that exposure to COVID-19 misinformation and conspiracy theories was nearly universal among Nigerian respondents, rooted more in distrust of government than in information deficits alone.
Erim et al. (2025)	Programme evaluation (Niger State, Nigeria)	Found that training community fellows drawn from media, faith-based, and traditional leadership structures improved participants' ability to identify and counter health misinformation at community level.
Dubawa (2026)	Fact-checking investigative report (Nigeria)	Documented multiple AI-generated deepfakes that deceived Nigerian audiences in 2025, noting that forensic verification tools — not unaided observation — were required to confirm manipulation.
Dubawa / PRNigeria (2025)	Fact-checking investigative report (Nigeria)	Documented a deepfake falsely depicting Nigerian soldiers in Benue State, illustrating how synthetic media can be weaponized to inflame ethnic and regional tension ahead of Nigeria's 2027 elections.
Olisa Agbakoba Legal (2025/2026)	Legal analysis (Nigeria)	Identified weak enforcement mechanisms, low public digital literacy, and a largely unregulated social-media environment as key gaps in Nigeria's institutional readiness against deepfakes.

DISCUSSION

The pattern of findings summarized in Table 1 supports H1 and H2: across methodologically diverse studies, deepfake exposure is consistently associated with reduced ability to distinguish authentic from fabricated content, and unaided detection accuracy repeatedly approximates chance level rather than reflecting reliable human vigilance (Köbis et al., 2021; Nightingale & Farid, 2022; Diel et al., 2024). This is consistent with Media Richness Theory's expectation that visually rich, high-fidelity synthetic content carries strong persuasive weight relative to text-based falsehoods (Daft & Lengel, 1986), and with Information Manipulation Theory's account of deception succeeding by exploiting the ordinarily reasonable assumption that video or image content is authentic (McCornack, 1992). The recurring overconfidence finding, in which detection confidence exceeds actual accuracy (Köbis et al., 2021), carries particular significance for institutions in contexts such as Taraba State that may rely on staff's self-assessed ability to recognize manipulated health, agricultural, or civic imagery without structured training.

The evidence also supports H3: interventions built around passive awareness or content labelling show limited effect, while feedback-driven, structured training consistently produced the strongest improvement across the reviews examined here (Groh et al., 2022; Somoray et al., 2025; Diel et al., 2024). This carries direct implications for Taraba State's institutions, where digital-literacy programming, if introduced, should be designed around interactive practice and feedback rather than one-off sensitization talks or poster campaigns, mirroring the demonstrated ineffectiveness of purely informational approaches documented in the wider literature.

Regarding H4, the Nigeria-specific evidence indicates that institutional readiness lags behind the sophistication of circulating content. Nigerian fact-checkers report needing specialized forensic tools, rather than unaided judgement, to confirm manipulation (Dubawa, 2026), and legal analysts point to limited enforcement capacity and an under-regulated information environment (Olisa Agbakoba Legal, 2025/2026). At the same time, the qualitative evidence from Niger State's community-fellow programme (Erim et al., 2025) and the pattern of government-distrust-driven misinformation belief documented by Wonodi et al. (2022) together suggest, in keeping with Berger and Luckmann's (1966) social-constructionist framework, that credibility within under-resourced Nigerian settings is built largely through trusted community messengers, religious leaders, traditional rulers, health workers, and radio presenters, rather than through technical verification alone. For Taraba State specifically, where public communication continues to depend heavily on radio and community gatherings alongside growing smartphone-based social media use, this suggests that community-anchored verification and messenger training may offer a more immediately feasible intervention than purely technological countermeasures, echoing but extending the community-fellow evidence from the neighbouring state.

A limitation running through the reviewed literature, and inherited by this review, is the concentration of controlled detection experiments in Western, English-speaking, and often university-recruited samples (Schlippe et al., 2025), which restricts confident generalization of exact detection-accuracy figures to Nigerian or Taraba-specific populations. The directional findings, that unaided detection is unreliable, that awareness alone is insufficient, and that trust is socially negotiated, are nonetheless corroborated by the Nigeria-specific investigative and programme-evaluation literature reviewed here, even though that literature does not report standardized detection-accuracy percentages comparable to the controlled experimental studies.

CONCLUSION

This review set out to synthesize what the existing literature indicates about how AI-generated deepfakes and synthetic media affect scientific and public consensus, human detection capacity, and institutional readiness, and to interpret that evidence for Taraba State, Nigeria. The specialized evidence indicates that deepfakes plausibly and measurably threaten scientific and public consensus by exploiting the perceptual richness of visual media and the ordinarily reasonable assumption that video and image content is authentic, that unaided human detection cannot be relied upon regardless of awareness or incentive, that structured feedback-based training is the most consistently effective known countermeasure, and that in the Nigerian context specifically, institutional verification capacity remains limited while community-anchored trust structures continue to play an outsized role in what is accepted as credible. On this basis, the threat documented in the literature is best addressed through coordinated institutional, educational, and community-level responses rather than technological intervention alone.

RECOMMENDATION

For Taraba State University and other state tertiary institutions, the evidence supports integrating interactive, feedback-based visual- and digital-literacy modules into research-methods and science-communication curricula, rather than one-off awareness sessions, and establishing a low-cost image/video verification consultation point for staff and postgraduate researchers using freely available forensic tools already referenced in Nigerian fact-checking practice (Dubawa, 2026).

For health authorities, including the Federal Medical Centre Jalingo and the Taraba State Ministry of Health, the Niger State community-fellow model (Erim et al., 2025) offers a directly adaptable framework for building rapid-response capacity around suspected manipulated health imagery or claims, given the demonstrated near-universal exposure to health misinformation documented among Nigerian publics (Wonodi et al., 2022).

For Taraba State's broadcast media, the evidence supports adopting basic verification checklists for user-submitted images and video prior to broadcast and participating in national fact-checking networks such as Dubawa to access verification tools beyond individual station capacity, addressing the gap between nominal awareness and applied verification skill documented in the wider literature.

For Taraba State policymakers and the House of Assembly, the evidence supports monitoring and contributing state-level input to Nigeria's evolving national online-harms framework and supporting funding for interactive, feedback-based digital- and scientific-literacy campaigns targeting rural Local Government Areas, given the documented insufficiency of passive awareness campaigns.

For community structures in Jalingo, Wukari, Bali, and other Local Government Areas, the evidence supports leveraging existing trusted community gatekeepers, religious leaders, traditional rulers, and radio presenters, as first-line counter-misinformation messengers, in keeping with both the social-constructionist framework and the proven success of the community-fellow model in a comparable Nigerian setting (Erim et al., 2025).

Future primary research grounded in Taraba State should build directly on this review by empirically testing unaided and trained detection accuracy among the state's researchers, health workers, journalists, policymakers, and general public, and by evaluating a Taraba-adapted community-fellow or verification-training programme, so that the propositions evaluated narratively here can be tested with locally collected data.

REFERENCES

1. Berger, P. L., & Luckmann, T. (1966). The social construction of reality: A treatise in the sociology of knowledge. Doubleday.
2. Chesney, R., & Citron, D. (2019). Deep fakes: A looming challenge for privacy, democracy, and national security. *California Law Review*, 107(6), 1753–1820.
3. Daft, R. L., & Lengel, R. H. (1986). Organizational information requirements, media richness and structural design. *Management Science*, 32(5), 554–571. https://doi.org/10.1287/mnsc.32.5.554
4. Diel, A., Lalgi, T., Schröter, I. C., MacDorman, K. F., Teufel, M., & Bäuerle, A. (2024). Human performance in detecting deepfakes: A systematic review and meta-analysis of 56 papers. *Computers in Human Behavior Reports*, 16, 100538. https://doi.org/10.1016/j.chbr.2024.100538
5. Dubawa. (2026, January 6). Six AI-generated deepfakes that deceived Nigerians in 2025. https://dubawa.org/six-ai-generated-deepfakes-that-deceived-nigerians-in-2025/
6. Dubawa. (2025, July 31). AI-generated disinformation looms ahead of Nigeria's 2027 elections. https://dubawa.org/ai-generated-disinformation-looms-ahead-of-nigerias-2027-elections/
7. Erim, A., Oko, S., Biose, S., Agbaoye, K., Nzedibe, O. E., Nwankwo, A., Gimba, P., & Ihekweazu, V. (2025). Tackling infectious disease outbreak and vaccination misinformation: A community-based strategy in Niger State, Nigeria. BMC Health Services Research, 25(1), 513. https://doi.org/10.1186/s12913-025-12683-z
8. Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., & Bengio, Y. (2014). Generative adversarial nets. *Advances in Neural Information Processing Systems*, 27, 2672–2680.
9. Groh, M., Epstein, Z., Firestone, C., & Picard, R. (2022). Deepfake detection by human crowds, machines, and machine-informed crowds. Proceedings of the National Academy of Sciences, 119(1), e2110013119. https://doi.org/10.1073/pnas.2110013119
10. Ho, J., Jain, A., & Abbeel, P. (2020). Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems*, 33, 6840–6851.
11. 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/10.1016/j.isci.2021.103364
12. McCornack, S. A. (1992). Information manipulation theory. *Communication Monographs*, 59(1), 1–16. https://doi.org/10.1080/03637759209376245
13. Nightingale, S. J., & Farid, H. (2022). AI-synthesized faces are indistinguishable from real faces and more trustworthy. Proceedings of the National Academy of Sciences, 119(8), e2120481119. https://doi.org/10.1073/pnas.2120481119

14. Olisa Agbakoba Legal. (2025/2026). Truth in the age of deepfakes: Legal frontiers against digital deception and image abuse. https://oal.law/truth-in-the-age-of-deepfakes-legal-frontiers-against-digital-deception-and-image-abuse/
15. Page, M. J., McKenzie, J. E., Bossuyt, P. M., Boutron, I., Hoffmann, T. C., Mulrow, C. D., Shamseer, L., Tetzlaff, J. M., Akl, E. A., Brennan, S. E., Chou, R., Glanville, J., Grimshaw, J. M., Hódhegyi, A., Lalu, M. M., Li, T., Loder, E. W., Mayo-Wilson, E., McDonald, S., … Moher, D. (2021). The PRISMA 2020 statement: An updated guideline for reporting systematic reviews. *BMJ*, 372, n71. https://doi.org/10.1136/bmj.n71
16. Schlippe, T., Wölfel, M., & Mabokela, K. R. (2025). A cross-cultural assessment of human ability to detect LLM-generated fake news about South Africa (arXiv:2511.17682). arXiv. https://arxiv.org/abs/2511.17682
17. Somoray, K., Miller, D. J., & Holmes, M. (2025). Human performance in deepfake detection: A systematic review. *Human Behavior and Emerging Technologies*, 2025(1), Article 1833228. https://doi.org/10.1155/hbe2/1833228
18. TheCable Fact Check. (2025, August 8). The rise of deepfakes and the fight for digital truth. https://factcheck.thecable.ng/the-rise-of-deepfakes-and-the-fight-for-digital-truth/
19. UNESCO. (2025, October 27). Deepfakes and the crisis of knowing. https://www.unesco.org/en/articles/deepfakes-and-crisis-knowing
20. Wonodi, C., Obi-Jeff, C., Adewumi, F., Keluo-Udeke, S. C., Gur-Arie, R., Krubiner, C., Jaffe, E. F., Bamiduro, T., Karron, R., & Faden, R. (2022). Conspiracy theories and misinformation about COVID-19 in Nigeria: Implications for vaccine demand generation communications. *Vaccine*, 40(13), 2114–2121. https://doi.org/10.1016/j.vaccine.2022.02.005