IIAR JOURNALS

*Innovating Knowledge. Advancing Research.*

IIAR Journal of Communication, Media, and Digital Society

Vol 1, No 1, 2026

# Synthetic Media and the Epistemics of Public Communication: A Theoretical Account of Deepfakes, Authenticity Infrastructure and Collective Belief

Atolagbe Umar Kawu[1], Nwogboji Emmanuel Usulor[2]

[1,2] Veristzon Global Research Lab

## ABSTRACT

*For most of the twentieth century, audio and video recordings were among the few kinds of media that publics accepted as evidence without asking who stood behind them. Generative models have ended that arrangement. Convincing synthetic voices, faces and scenes can now be produced cheaply and at scale, and the public conversation about this change has concentrated on one danger: that people will be fooled. This conceptual, non-empirical paper argues that the more serious danger lies elsewhere. Drawing on peer-reviewed scholarship published between 2023 and 2026 in social epistemology, disinformation and media-effects research, deepfake detection and political deepfake studies, including work on African elections, and research on provenance and content authentication, the paper distinguishes two harms. The deception harm is the familiar one, in which people come to believe something false because a fabricated recording looks real. The scepticism harm is the loss of belief in things that are true, because the mere possibility of fabrication gives every viewer a reason to doubt and every exposed actor an excuse. The synthesis suggests that the scepticism harm is larger and harder to remedy. It is produced by general awareness that fakes exist rather than by any particular fake, it survives correction, and it is actively exploited by those who dismiss authentic evidence against them, a pattern known as the liar's dividend. From this distinction the paper develops an authenticity-infrastructure framework, which holds that a public's epistemic health depends less on the ability to detect fakes than on the availability of provenance, attestation by accountable parties, and trusted intermediaries that can establish what is real. Six propositions specify the conditions under which publics keep the capacity to agree on facts. The paper then examines conditions in Africa that heighten exposure to the scepticism harm, namely low institutional trust, the circulation of political content in closed messaging groups, and limited platform investment in content integrity for African languages and markets. It concludes with recommendations for newsrooms, electoral bodies, regulators and platforms.*

***Keywords:*** deepfakes; synthetic media; disinformation; authenticity; provenance

# Synthetic Media and the Epistemics of Public Communication: A Theoretical Account of Deepfakes, Authenticity Infrastructure and Collective Belief

## 1. INTRODUCTION

A recording of a politician's voice reaches a WhatsApp group a week before an election. In it, the candidate appears to admit to rigging the count. Within hours, one of two things can go wrong. Members of the group may believe a fabrication. Or, if the recording is genuine, the candidate may declare it a deepfake, and many members may accept that explanation because they know such fakes are now easy to make. Public debate about synthetic media has dwelt almost entirely on the first of these failures. This paper is about the second.

Recordings once enjoyed a special position in public communication. A written report depends on the honesty of its author, but audio and video seemed to carry their own warrant: they showed what happened, or at least they were hard to fake well. That position was always partly a social achievement rather than a property of the technology (Habgood-Coote, 2023), yet it was real enough to shape courts, newsrooms and ordinary arguments about what had happened. Generative models now produce convincing voices, faces and scenes at negligible cost, and the presumption that a recording is evidence no longer holds by default (Ferrara, 2026).

Research on this change has grown quickly but along separate tracks. Experiments ask whether people believe deepfakes and whether fact-checks reverse the damage (Barari et al., 2025; Dan, 2026), computer scientists document how poorly detectors generalise (Kaur et al., 2024), philosophers ask whether deepfakes threaten knowledge itself (Habgood-Coote, 2023; Harris, 2024), and studies of African elections examine disinformation on messaging apps (Asiedu, 2026; Gondwe, 2025). These literatures rarely meet, and most framing still treats being fooled as the central harm.

The paper asks three questions. What kinds of epistemic harm does synthetic media produce, and how do they differ? Why is detection an insufficient response? And what conditions allow a public, particularly in African settings, to keep its capacity to agree on facts? It answers them by developing a distinction between a deception harm and a scepticism harm, an authenticity-infrastructure framework, and six propositions. The argument proceeds through a literature review, a statement of method, the framework and its propositions, a discussion of African conditions, and recommendations.

## 2. LITERATURE REVIEW

### 2.1 CONCEPTUAL REVIEW

*Synthetic media* refers to audio, images, video and text produced or substantially altered by generative models so that they appear to record real people or events. *Deepfakes* are the subset that place a real person's face or voice in words or actions that never occurred. *Cheapfakes* are cruder manipulations, such as misleading edits, that need no machine learning; experimental evidence suggests they can damage a politician's reputation as much as deepfakes (Dan, 2026).

*Epistemic trust* is the reliance one person places on another's communication as a source of belief. Kotsonis (2024) treats it as a virtue: the good truster depends on others while remaining skilled at judging when such dependence is warranted. On this view, the opposite of credulity is not blanket doubt: a public that trusts nothing is as badly placed as one that trusts everything.

*Authenticity* in this paper means that a piece of media is what it claims to be: that it was captured where, when and by whom it says, and has not been materially altered. *Provenance* is the documented history of a piece of media, from capture through editing to publication. Content-credential systems, such as those developed by the Coalition for Content Provenance and Authenticity, attach signed metadata to media so that this history can be checked (Wersényi & Koech, 2026). *Attestation* is a statement by an identifiable and accountable party that a piece of media is authentic. *Authenticity infrastructure* is the combination of provenance, attestation and

© IIARJournal. All rights reserved.

trusted intermediaries through which a public establishes what is real.

The *liar's dividend* names the advantage that people gain when they can dismiss authentic evidence as fabricated, because the public knows fabrication is possible (Grohmann et al., 2026; Schiff et al., 2025).

## 2.2 THEORETICAL REVIEW

Three bodies of theory inform the framework. The first is social epistemology, which holds that most of what people know depends on the testimony of others and on institutions that sort reliable testimony from unreliable. Habgood-Coote (2023) uses this tradition to criticise the claim that deepfakes will bring about an epistemic apocalypse. On his account, recordings never spoke for themselves; their authority rested on social practices of capture, custody and vouching. The problem deepfakes create is therefore a problem for those practices, and the solution lies in repairing them rather than in technology alone. Harris (2024) reaches a related point from the side of technology. He argues that reactive approaches, which use AI to identify synthetic media, face a version of the ancient problem of the criterion, with sceptical implications, and that proactive approaches, which create verified records of authentic content, resist this objection better but place a heavy burden on users to verify their own content, a burden likely to deepen existing inequalities.

The second body of theory comes from research on misinformation and media effects. Here a recurring warning is that the fight against misinformation can itself produce harm. Altay et al. (2023) argue that alarmist accounts of misinformation overstate its reach and effects, and that such accounts may feed general distrust. Simon et al. (2023) apply this reasoning to generative AI, arguing that fears about its effect on misinformation are overblown because the supply of false content was never the main constraint; demand and attention were. Van der Meer et al. (2023) show experimentally that warnings about the threat of misinformation can lower the credibility of accurate news. The most important effects of synthetic media may therefore operate through beliefs about the information environment rather than through exposure to particular fakes.

The third body of theory concerns visual persuasion. Images and recordings are persuasive partly because they appear to be caused by what they depict. Weikmann and Lecheler (2023) synthesise research showing that visual disinformation draws on this perceived indexicality, and that its effects therefore differ from those of text. Once indexicality can no longer be assumed, the effect extends to every recording. Ferrara (2026) frames this as a shift from isolated synthetic artefacts to synthetic realities, in which content, identity, interaction and institutions can all be manufactured, so that the shared ground for verification erodes.

## 2.3 EMPIRICAL REVIEW

Evidence on the deception harm is more reassuring than public debate suggests. Barari et al. (2025) created deepfakes with industry partners and a professional actor and found that they were about as credible as the same misinformation presented in text or audio, not uniquely deceptive. Hameleers et al. (2024a) found that deepfakes with plausible content were judged far more credible than implausible ones, yet the implausible versions did more to delegitimise the politician they targeted, so a deepfake need not be believed to cause damage. When audiences do doubt a deepfake, the reason is more often that its content departs from what they expect of the speaker than that they notice technical glitches, and many still struggle to tell it apart from a related authentic video (Hameleers et al., 2024b). Dan (2026) found that fake scandal videos damaged an innocent politician's reputation, but that a journalistic fact-check substantially reduced and in some cases eliminated the damage. Łabuz and Nehring (2024) document deepfakes across election campaigns in 2023.

Evidence on human detection is less encouraging. A meta-analysis of 56 papers found that people's ability to detect deepfakes is, on average, little better than chance (Diel et al., 2024). People rely more on how something is said than on what is said (Groh et al., 2024), and they judge deepfakes through heuristic cues and context rather than systematic inspection (Shin, 2026), so detection will become harder as visual and vocal quality improves. Machine detectors trained on one generation method often fail on another (Kaur et al., 2024).

Evidence on the scepticism harm is growing. Barari et al. (2025) found that participants often mistook authentic videos for deepfakes when the video

© IIARJournal. All rights reserved.

showed a politician from their own party in a scandal. Weikmann et al. (2025) found that revealing a deepfake lowered the credibility participants assigned to all media formats and reduced their confidence in their own ability to tell real from fake, including among those who had seen only authentic material. In a United States survey, Ahmed (2023) found that exposure to deepfakes and concern about them were both positively related to scepticism towards news on social media. Analysing tweets about deepfakes during the Russian invasion of Ukraine, Twomey et al. (2023) found that much of the misinformation in their data came from people labelling real media as deepfakes, that deepfake scepticism was linked to conspiratorial beliefs, and that efforts to raise awareness of deepfakes may undermine trust in legitimate videos. Altay and Gilardi (2024) found that labelling headlines as AI-generated reduced belief in them even when they were true or written by humans, because people assumed full automation. Journalism research reports that audiences trust labelled AI content less (Li & Lu, 2026; Toff & Simon, 2025).

The liar's dividend has now been tested directly. In five survey experiments in the United States, Schiff et al. (2025) found that politicians who claimed that a scandal story was misinformation gained support, and gained more than by staying silent or apologising. The strategy worked against text-based reports but was largely ineffective against video evidence. A later experiment reached a different result for video. Grohmann et al. (2026) found that a politician's false claim that genuine video evidence was a deepfake raised perceptions of his leadership ability, and that this effect worked through viewers misidentifying the authentic video as fake. Deepfake perception was also negatively associated with general trust in media.

Evidence on provenance is mainly technical but has begun to include users. In an experiment with 595 participants, Feng et al. (2023) found that provenance information often lowered trust in deceptive media, particularly when it revealed that an image was composited. When provenance information was shown to be incomplete or invalid, however, some participants came to disbelieve honest media, and many did not fully understand what provenance meant. Wersényi and Koech (2026) review content credentials and note that only a few devices and applications yet implement them. Photojournalists see generative AI as continuous with older threats of image manipulation (Matich et al., 2025).

African evidence centres on elections and messaging apps. In South Africa's 2024 elections, an analysis of more than 22,000 messages from 47 politically active WhatsApp groups found AI-generated deepfakes alongside fear appeals and content that mimicked credible journalism; disinformation spread through emotional resonance and relational trust, and corrective content was often dismissed (Gondwe, 2025). In Nigeria's 2023 presidential campaign, campaign media units both produced disinformation and responded to it, including by tagging unwelcome information as fake news (Ibrahim & Ekdale, 2026). In Kenya's 2022 election, fact-checking struggled because false information moved faster than verification could, a dynamic described as the timing trap (Asiedu, 2026). A study of the UNDP-supported iVerify initiative across five sub-Saharan countries found that fact-checkers acted as gatekeepers selecting which claims to verify (Tsetsura et al., 2026). West African parties organise extensively through WhatsApp (Fisher et al., 2024), Nigerian protest movements have coordinated through mobile social networking apps (Uwalaka, 2023), and mobile connectivity can lower rural trust in government by strengthening ties to relatives who report on urban hardship (Yeandle, 2026).

### 2.4 GAPS IN THE LITERATURE

Three gaps follow. First, deception and scepticism are studied in separate experiments, and no account explains how they relate or which matters more. Second, detection receives far more attention than the social arrangements that establish authenticity, even though the evidence shows both human and machine detection to be unreliable. Third, African research documents disinformation on messaging apps and the limits of fact-checking, but it has not yet been connected to the epistemic theory of synthetic media or to provenance. This paper addresses all three.

## 3. MATERIALS AND METHODS

**Research design.** This is a theoretical, non-empirical article. It uses integrative conceptual synthesis to build a framework and to derive propositions for later empirical testing. No human participants were

© IIARJournal. All rights reserved.

involved and no primary data were collected, so population, sampling and instrument validation in the empirical sense do not apply.

**Source identification.** Literature was identified through structured searches of EBSCOhost databases accessed through the University of South Africa (Unisa) Library (Academic Search Ultimate, Business Source Ultimate and APA PsycArticles), the Crossref metadata service and the OpenAlex scholarly index. Search strings combined "deepfake", "synthetic media" or "AI-generated" with terms such as "epistemic", "trust", "liar's dividend", "provenance", "content credentials", "election", "WhatsApp", "fact-checking" and "Africa". Searches were run on 20 and 21 October 2026.

**Inclusion and exclusion criteria.** Peer-reviewed journal articles published between 2023 and 2026 were included if they addressed the effects of synthetic media on belief and trust, detection, provenance, the epistemology of recordings, or disinformation and verification in Africa. Preprints, conference papers, magazine articles, editorials, book reviews, grey literature, items without a resolvable digital object identifier (DOI), and outlets whose peer-review practices could not be confirmed were excluded.

**Verification.** Each candidate was checked against its Crossref record for DOI, authors, title, journal, volume, issue and pagination, and was screened for retractions, corrections and expressions of concern through Crossref update metadata. One otherwise relevant article was excluded because a correction notice was attached to it. Where an article is indexed in the EBSCOhost databases, its peer-reviewed status and abstract were also confirmed there. Publication year follows the issue year where one exists. Thirty-seven articles met all criteria and are cited.

**Analytic procedure.** Sources were coded for the type of harm studied (belief in false content or disbelief in true content), the mechanism proposed, the remedy tested, and the setting. Mechanisms that recurred across settings were stated as propositions, and African studies were used to judge whether each was likely to be stronger, weaker or different in form in African information environments.

**Bases of judgement and trustworthiness.** Conceptual work is judged by clarity, parsimony, explanatory reach and falsifiability. Each proposition is stated so that it could prove wrong, and the discussion identifies evidence that would weaken it.

**AI usage disclosure.** Generative AI tools were used to assist with literature retrieval, metadata checking and language editing. The arguments, interpretations and conclusions were developed and reviewed by the author(s), who take full responsibility for the content. No AI tool is listed as an author, and no source was cited without verification against its publisher record.

**Limitations.** The synthesis is restricted to English-language journal articles from a four-year window. Experimental evidence on the scepticism harm comes mostly from Europe and North America, and peer-reviewed research on deepfakes in African elections is still scarce, so some African claims rest on studies of disinformation more broadly. The framework identifies mechanisms and does not estimate their size.

## 4. RESULTS AND DISCUSSION

### 4.1 RESULTS: TWO HARMS AND AN AUTHENTICITY-INFRASTRUCTURE FRAMEWORK

The synthesis yields two linked results. The first is a distinction between the deception harm and the scepticism harm, summarised in Table 1. The second is a framework that explains what protects a public against both, shown in Figure 1.

Table. 1. The deception harm and the scepticism harm compared

| Feature | Deception harm | Scepticism harm |
|---|---|---|
| **Definition** | Believing a fabricated recording | Disbelieving an authentic recording |
| **Cause** | Exposure to a particular fake | Awareness that fakes are possible |
| **Scope** | Local to the item and its audience | General: spreads to all recordings and to self-confidence |
| **Response to correction** | Largely reversible by fact-checks | Resistant; there is no single falsehood to correct |
| **Strategic use** | Fabricators create fakes | Exposed actors deny real evidence (liar's dividend) |
| **Effect of warnings and labels** | Reduced | Increased |

*Note.* Compiled by the author(s) from the synthesis in Sections 2.3 and 4.1.

© IIARJournal. All rights reserved.

**The deception harm.** The deception harm occurs when a fabricated recording leads people to believe something false. It is the harm that detection tools, fact-checks and media-literacy campaigns are designed to prevent. The evidence reviewed above suggests that it is real but bounded. Deepfakes are roughly as credible as other forms of misinformation (Barari et al., 2025), and people bring prior knowledge to bear when a deepfake contradicts what they know of the speaker (Hameleers et al., 2024b). Its effects can also be reversed. A fact-check substantially reduced the reputational damage of fake scandal videos in Dan's (2026) experiment. Because it is attached to particular items, much of it can be undone once a fake is identified and corrected, at least for those who see the correction.

**The scepticism harm.** The scepticism harm occurs when people stop believing true recordings because they know fakes are possible. It differs from the deception harm in four ways. It is produced by awareness of the technology rather than by any particular fake, which is why Weikmann et al. (2025) observed reduced trust among participants who had seen only authentic media. It is general: doubt spreads from one recording to all recordings, and to viewers' confidence in their own judgement. It survives correction, because there is no single falsehood to correct; one can debunk a fake but cannot easily "rebunk" a truth that someone has chosen to doubt. And it can be exploited strategically. Politicians and campaigns can invoke the possibility of fabrication to escape accountability (Grohmann et al., 2026; Schiff et al., 2025), and campaign units in Nigeria have used fake-news tagging as a deliberate response to unwelcome information (Ibrahim & Ekdale, 2026).

The scepticism harm is also partly self-inflicted by the public conversation about deepfakes. Warnings about misinformation can lower the credibility of accurate news (van der Meer et al., 2023), labels intended to inform can reduce belief in true content (Altay & Gilardi, 2024), alarmist framing can increase distrust out of proportion to actual exposure (Altay et al., 2023), and public efforts to raise awareness of deepfakes may themselves undermine trust in legitimate video (Twomey et al., 2023). Each warning that "anything could be fake" adds to the stock of doubt on which the liar's dividend draws.

**From detection to infrastructure.** If the scepticism harm is the larger problem, detection cannot be the main remedy. Detection asks "is this fake?"; it does nothing for an authentic recording that is disbelieved, and its unreliability makes "no manipulation detected" weak reassurance (Diel et al., 2024; Kaur et al., 2024). A public needs ways to establish positively that something is real, and the framework identifies three.

*Provenance* establishes where a piece of media came from and what has been done to it. Content credentials and similar systems sign media at capture and record subsequent edits (Wersényi & Koech, 2026). Provenance shifts the question from "can I spot a flaw?" to "can I see a chain of custody?". Harris (2024) argues that such proactive approaches resist the sceptical regress that undermines detection, and provenance information can help users discount deceptive media (Feng et al., 2023). Yet provenance has limits. It can prove that signed media is authentic, but it cannot prove that unsigned media is fake, and while adoption remains thin, most authentic media will carry no credentials at all.

*Attestation* attaches an accountable party to a claim of authenticity. A newsroom that states that its reporter recorded a video at a given place and time, and stakes its reputation on the claim, provides what metadata cannot: a party who can be held responsible if the claim is false. Habgood-Coote's (2023) argument that the authority of recordings always rested on social practices of vouching implies that attestation is not a stopgap but the core of authenticity. Aharoni (2026) similarly argues that news trust should be understood as embedded in wider structures rather than as an audience evaluation alone.

*Trusted intermediaries* carry provenance and attestation to the people who need them. They include newsrooms, fact-checkers, electoral bodies, civil-society monitors and, in closed messaging environments, the individuals whose word carries weight within a group. In South African WhatsApp groups, corrections were often dismissed when they came from outside the group (Gondwe, 2025), and Yeandle (2026) shows how personal networks shape political trust. An authenticity signal that does not pass through a trusted node may never be heard.

© IIARJournal. All rights reserved.

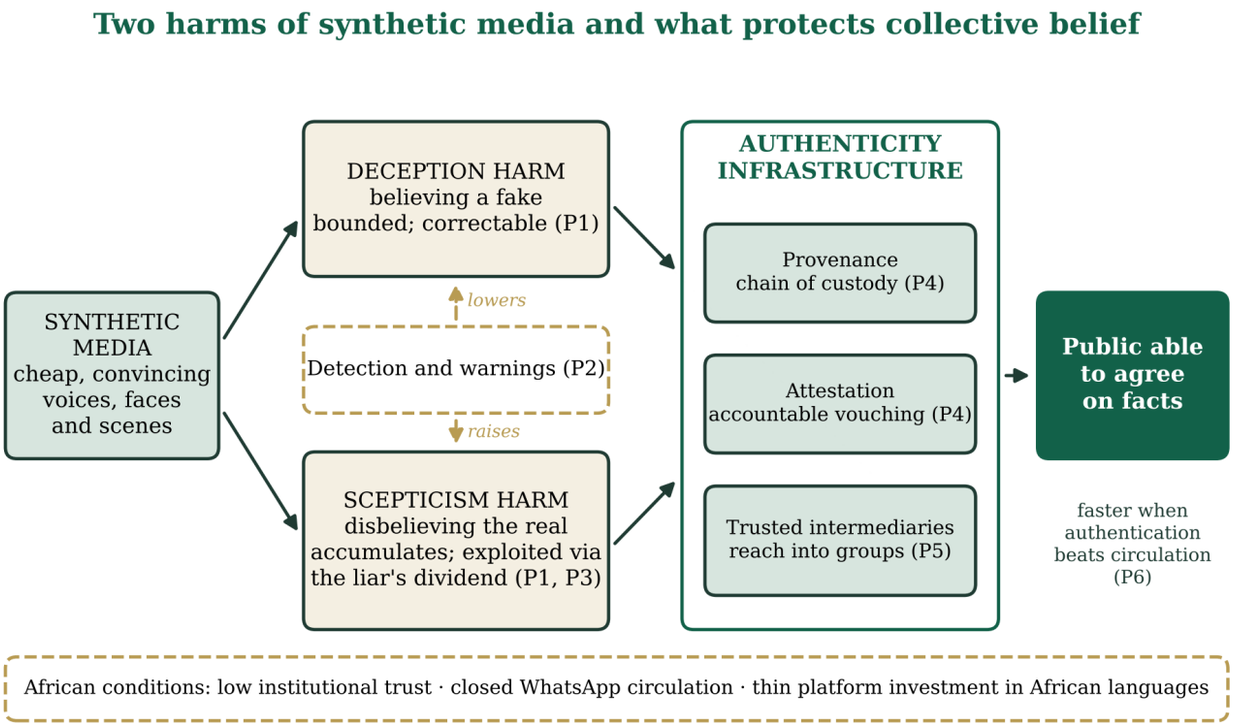



**Figure. 1.** The authenticity-infrastructure framework: synthetic media produces a deception harm and a larger scepticism harm; detection addresses only the first, whereas provenance, attestation and trusted intermediaries sustain the public's capacity to agree on facts. Source: developed by the author(s).

The framework's central claim is that a public's epistemic health depends on the availability and reach of this infrastructure more than on its ability to detect fakes. From it, six propositions follow.

**Proposition 1: The deception harm of synthetic media is bounded and largely correctable, whereas the scepticism harm accumulates and resists correction.** The first part follows from evidence that deepfakes are not uniquely credible and that fact-checks can reverse their effects (Barari et al., 2025; Dan, 2026). The second follows from evidence that awareness of fakes lowers trust in authentic media and self-confidence in judgement (Ahmed, 2023; Weikmann et al., 2025). The proposition would be weakened by longitudinal evidence that general media trust recovers quickly after deepfake incidents, or that deepfakes produce lasting false beliefs that corrections cannot reach.

**Proposition 2: Strategies built on detection and warning reduce the deception harm at the cost of increasing the scepticism harm.** Detection is unreliable for humans and machines (Diel et al., 2024; Kaur et al., 2024), and warnings and labels spread doubt to true content (Altay & Gilardi, 2024; van der Meer et al., 2023). A strategy that teaches the public to suspect everything will lower belief in fakes and in truths alike. The proposition would be weakened if interventions were found that raise discernment, meaning correct acceptance as well as correct rejection, without lowering trust in authentic content.

**Proposition 3: The liar's dividend grows with public awareness of synthetic media, partisan motivation and low institutional trust.** Claims of misinformation protect politicians against text-based reports (Schiff et al., 2025), and more recent evidence shows that false deepfake claims also pay a dividend against video by leading viewers to misidentify real footage as fake (Grohmann et al., 2026). The contrast between these studies is consistent with a dividend that extends to video as awareness of synthetic media spreads. Partisans are also especially likely to doubt authentic videos of their own side's scandals (Barari et al., 2025). Where institutions that could contradict a politician's denial are themselves distrusted, the denial meets less resistance. The proposition would be weakened by evidence that the dividend is small or short-lived in low-trust settings.

© IIARJournal. All rights reserved.

**Proposition 4: Positive authentication through provenance and attestation protects collective belief more effectively than negative detection, but only when it is widely adopted and anchored in accountable institutions.** Provenance helps users discount deceptive media (Feng et al., 2023), and attestation supplies accountability that technical signals lack (Habgood-Coote, 2023). Thin adoption, however, creates a new risk. Incomplete or invalid provenance can lead people to disbelieve honest media (Feng et al., 2023), and proactive verification places burdens on users that track existing inequalities (Harris, 2024). If authentic media without credentials is treated as suspect, provenance could deepen the scepticism harm for those who cannot afford compatible devices or whose media passes through platforms that strip metadata (Wersényi & Koech, 2026). The proposition would be weakened by evidence that provenance labels are ignored or that their absence produces more unwarranted doubt than their presence removes.

**Proposition 5: In closed messaging environments, relational trust takes the place of institutional trust, so authenticity infrastructure works only when it reaches the trusted nodes within groups.** Disinformation in South African WhatsApp groups gained traction through relational trust, and external corrections were often rejected (Gondwe, 2025). Parties in West Africa organise through WhatsApp (Fisher et al., 2024), and personal networks shape political trust (Yeandle, 2026). In such settings, a newsroom's attestation matters only if a group member who is trusted passes it on. The proposition would be weakened by evidence that platform-level labels or forwarded fact-checks change belief within closed groups without such brokers.

**Proposition 6: The capacity of a public to agree on facts depends on the speed of authentication relative to the speed of circulation.** The timing trap in Kenya's 2022 election shows that verification loses when false content outpaces it (Asiedu, 2026). The same logic applies to authentic content that is being denied: the longer a recording remains unattested, the longer the liar's dividend can be collected, so authentication at capture changes the balance. The proposition would be weakened by evidence that late authentication restores belief as fully as early authentication.

Table 2 relates the framework to African conditions.

Table. 2. Authenticity infrastructure under African conditions

| Element | African constraint | Risk | Response |
|---|---|---|---|
| **Provenance** | Few devices with content credentials; media forwarded through apps that strip metadata | Authentic media without credentials treated as suspect | Open, low-cost standards; platforms required to preserve metadata |
| **Attestation** | Low trust in electoral bodies, governments and parts of the press | Denials of real evidence meet little resistance | Rapid joint attestation by newsrooms, observers and electoral bodies |
| **Trusted intermediaries** | Belief formed in closed WhatsApp groups through relational trust | Outside corrections dismissed | Route authentication through trusted group members and local fact-checkers |
| **Speed** | Verification slower than circulation during elections | Liar's dividend collected before authentication | Authenticate at capture; prepare election-period authentication desks |

*Note.* Developed by the author(s) from the propositions in Section 4.1 and the discussion in Section 4.2.

## 4.2 DISCUSSION

The deception–scepticism distinction reframes the problem. The dominant framing asks how to stop people from being fooled. The evidence suggests that people are fooled less often, and less durably, than feared, and that a larger and more persistent cost arises when people stop believing what is true. Deception is not harmless; fake scandal videos do damage (Dan, 2026), and deepfakes are being used in election campaigns (Gondwe, 2025; Łabuz & Nehring, 2024). The claim is about priorities. A response designed only around deception will tend to make the scepticism harm worse, because its main tools, warnings, labels and detection, all work by raising doubt.

This conclusion fits Habgood-Coote's (2023) argument against the apocalyptic reading of deepfakes but gives it a practical direction: if the authority of recordings always rested on social practices, the task is to rebuild those practices for a world in which capture is no longer proof. The framework argues that provenance, attestation and

© IIARJournal. All rights reserved.

intermediaries each need the others. Provenance without attestation is a technical signal that few people can read. Attestation without provenance asks the public to trust institutions on their word alone. Both are ineffective if they do not reach the groups where belief is formed.

The framework also clarifies the place of journalism. Newsrooms have long performed attestation, and the custody and verification practices of photojournalists (Matich et al., 2025) now look like a public good rather than a professional habit. The disclosure literature adds a caution: labelling AI involvement lowers trust (Li & Lu, 2026; Toff & Simon, 2025). Newsrooms may serve the public better by attesting positively to what is authentic than by flagging what is synthetic.

**African conditions.** Three features of many African information environments heighten exposure to the scepticism harm. The first is low institutional trust. Where electoral bodies, governments and sometimes the press are distrusted, the institutions that would normally attest to authenticity lack the standing to do so, and a politician's claim that damaging evidence is fake meets less resistance. Campaigns in Nigeria have already used fake-news tagging as a tactic (Ibrahim & Ekdale, 2026). The second is the circulation of political content in closed messaging groups. WhatsApp is central to political organisation in West Africa (Fisher et al., 2024) and to protest coordination in Nigeria (Uwalaka, 2023), and in South Africa it carried AI-generated deepfakes during the 2024 elections (Gondwe, 2025). Content there travels without metadata, beyond platform labels, and through relationships that outweigh institutional signals. The third is limited platform investment in content integrity for African languages and markets. Udupa et al. (2023) document the severe limits of machine-learning moderation in keeping pace with extreme speech and propose "ethical scaling", a moderation process built on community participation, as a response to the colonial structures that shape both speech and its governance. Fact-checking initiatives such as iVerify depend on external support and must choose which claims to check (Tsetsura et al., 2026), while the timing trap limits what they can achieve during elections (Asiedu, 2026).

These conditions do not make African publics more gullible. Low trust and active campaigning against "fake news" are more likely to raise scepticism than credulity, which is why the scepticism harm deserves priority in African settings. A public that already doubts institutions, and that learns that recordings can be faked, may find it harder than most to agree on what happened in a disputed election.

**Implications for theory.** The value of a recording as evidence depends on infrastructure that makes its authenticity checkable and accountable. This shifts attention from the properties of media to the systems through which media is vouched for, and it implies that measures of a public's epistemic health should capture its ability to accept truths as well as to reject falsehoods.

**Counterarguments.** One objection holds that publics should doubt recordings more than they once did. The framework agrees that calibrated doubt is a virtue (Kotsonis, 2024); its concern is with uncalibrated doubt that treats all recordings alike. A second objection holds that provenance will create a two-tier system in which only well-resourced actors can prove authenticity. That risk is built into Proposition 4 and argues for open standards and institutional attestation that does not depend on owning particular devices.

## 5. CONCLUSION AND RECOMMENDATIONS

Synthetic media has ended the period in which recordings could serve as evidence without further warrant. The most serious consequence is not that people will believe fakes but that they will stop believing what is real, while those exposed by genuine evidence gain a ready excuse. The scepticism harm is larger and harder to remedy than the deception harm, and detection-centred responses can make it worse. Publics keep their capacity to agree on facts when authenticity infrastructure, made up of provenance, attestation and trusted intermediaries, can establish positively and quickly what is real. In African settings, low institutional trust, closed messaging circulation and thin platform investment make the scepticism harm especially pressing.

Several recommendations follow. Newsrooms should treat attestation as a core public function: they

© IIARJournal. All rights reserved.

should document the custody of original recordings, adopt content credentials where feasible, and publish clear statements of authenticity for important material rather than relying mainly on labels for synthetic content. Electoral bodies and civil-society monitors should establish rapid authentication arrangements for election periods, including channels through which parties, journalists and observers can submit disputed recordings and receive an attested finding quickly. Regulators should support open provenance standards and require platforms to preserve provenance metadata, while ensuring that missing credentials are never treated as proof of fabrication. Platforms should invest in content integrity for African languages and markets and help verified authenticity information travel into messaging groups. Public communicators, researchers included, should stop implying that anything could be fake and explain instead how authentic material can be recognised. Future research should measure the scepticism harm directly and test provenance and attestation in African messaging environments.

## 6. CONTRIBUTION TO EXISTING KNOWLEDGE

The paper contributes in four ways. Conceptually, it separates the deception harm from the scepticism harm and argues that the latter is larger, more persistent and more open to strategic exploitation. Theoretically, it develops an authenticity-infrastructure framework that integrates provenance, attestation and trusted intermediaries and explains why detection alone cannot sustain collective belief. Analytically, it states six falsifiable propositions on the conditions under which publics retain the capacity to agree on facts, including the role of relational trust in closed messaging groups and the race between authentication and circulation. Contextually, it brings African evidence on elections, messaging apps and fact-checking into the epistemic theory of synthetic media.

## 7. REFERENCES

Aharoni, T. (2026). Beyond audience evaluations: Re-thinking news trust in digital environments. *Journalism Studies*, *27*(3), 340–359. https://doi.org/10.1080/1461670X.2025.2588239

Ahmed, S. (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

Altay, S., Berriche, M., & Acerbi, A. (2023). Misinformation on misinformation: Conceptual and methodological challenges. *Social Media + Society*, *9*(1). https://doi.org/10.1177/20563051221150412

Altay, S., & Gilardi, F. (2024). People are skeptical of headlines labeled as AI-generated, even if true or human-made, because they assume full AI automation. *PNAS Nexus*, *3*(10), pgae403. https://doi.org/10.1093/pnasnexus/pgae403

Asiedu, M. (2026). The timing trap: Misinformation, disinformation, and the limits of fact-checking in Kenya's 2022 elections. *South African Journal of International Affairs*, *33*(2), 271–298. https://doi.org/10.1080/10220461.2026.2683199

Barari, S., Lucas, C., & Munger, K. (2025). Political deepfakes are as credible as other fake media and (sometimes) real media. *The Journal of Politics*, *87*(2), 510–526. https://doi.org/10.1086/732990

Dan, V. (2026). Deepfakes as a democratic threat: Experimental evidence shows noxious effects that are reducible through journalistic fact checks. *The International Journal of Press/Politics*, *31*(3), 525–550. https://doi.org/10.1177/19401612251317766

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

Feng, K. J. K., Ritchie, N., Blumenthal, P., Parsons, A., & Zhang, A. X. (2023). Examining the impact of provenance-enabled media on trust and accuracy perceptions. *Proceedings of the ACM on Human-Computer Interaction*, *7*(CSCW2), 1–42. https://doi.org/10.1145/3610061

Ferrara, E. (2026). The generative AI paradox: GenAI and the erosion of trust, the corrosion of information verification, and the demise of truth. *Future Internet*, *18*(2), 73. https://doi.org/10.3390/fi18020073

Fisher, J., Gadjanova, E., & Hitchen, J. (2024). WhatsApp and political communication in West Africa: Accounting for differences in parties' organization and message discipline online. *Party*

© IIARJournal. All rights reserved.

*Politics*, *30*(5), 934–948. https://doi.org/10.1177/13540688231188690

Gondwe, G. (2025). AI-driven disinformation and political influence on WhatsApp in South Africa's 2024 elections. *The International Journal of Press/Politics*. Advance online publication. https://doi.org/10.1177/19401612251395434

Grohmann, L., Halle, F. A., & Appel, M. (2026). Deepfake! A liar's dividend for audiovisual material. *Psychology of Popular Media*. Advance online publication. https://doi.org/10.1037/ppm0000665

Groh, M., Sankaranarayanan, A., Singh, N., Kim, D. Y., Lippman, A., & Picard, R. (2024). Human detection of political speech deepfakes across transcripts, audio, and video. *Nature Communications*, *15*, 7629. https://doi.org/10.1038/s41467-024-51998-z

Habgood-Coote, J. (2023). Deepfakes and the epistemic apocalypse. *Synthese*, *201*(3), 103. https://doi.org/10.1007/s11229-023-04097-3

Hameleers, M., van der Meer, T. G. L. A., & Dobber, T. (2024a). Distorting the truth versus blatant lies: The effects of different degrees of deception in domestic and foreign political deepfakes. *Computers in Human Behavior*, *152*, 108096. https://doi.org/10.1016/j.chb.2023.108096

Hameleers, M., van der Meer, T. G. L. A., & Dobber, T. (2024b). They would never say anything like this! Reasons to doubt political deepfakes. *European Journal of Communication*, *39*(1), 56–70. https://doi.org/10.1177/02673231231184703

Harris, K. R. (2024). Synthetic media detection, the wheel, and the burden of proof. *Philosophy & Technology*, *37*(4), 131. https://doi.org/10.1007/s13347-024-00821-0

Ibrahim, A. A., & Ekdale, B. (2026). Political campaign responses to information disorder: A case study of the 2023 Nigerian presidential elections. *The International Journal of Press/Politics*, *31*(3), 595–615. https://doi.org/10.1177/19401612251324985

Kaur, A., Noori Hoshyar, A., Saikrishna, V., Firmin, S., & Xia, F. (2024). Deepfake video detection: Challenges and opportunities. *Artificial Intelligence Review*, *57*(6), 159. https://doi.org/10.1007/s10462-024-10810-6

Kotsonis, A. (2024). Epistemic trust in the age of misinformation. *Journal of Philosophical Research*, *49*, 31–47. https://doi.org/10.5840/jpr202487218

Łabuz, M., & Nehring, C. (2024). On the way to deep fake democracy? Deep fakes in election campaigns in 2023. *European Political Science*, *23*(4), 454–473. https://doi.org/10.1057/s41304-024-00482-9

Li, Y., & Lu, H. (2026). Perceived legitimacy matters: Building public trust and acceptance of AI-generated news images through strategic AI disclosure. *Journalism & Mass Communication Quarterly*. Advance online publication. https://doi.org/10.1177/10776990261462536

Matich, P., Thomson, T. J., & Thomas, R. J. (2025). Old threats, new name? Generative AI and visual journalism. *Journalism Practice*, *19*(10), 2402–2421. https://doi.org/10.1080/17512786.2025.2451677

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/10.1017/S0003055423001454

Shin, D. (2026). Seeing through the fake: How users detect and interpret deepfakes. *Information, Communication & Society*, *29*(2), 473–491. https://doi.org/10.1080/1369118X.2025.2518260

Simon, F. M., Altay, S., & Mercier, H. (2023). Misinformation reloaded? Fears about the impact of generative AI on misinformation are overblown. *Harvard Kennedy School (HKS) Misinformation Review*, *4*(5). https://doi.org/10.37016/mr-2020-127

Toff, B., & Simon, F. M. (2025). "Or they could just not use it?": The dilemma of AI disclosure for audience trust in news. *The International Journal of Press/Politics*, *30*(4), 881–903. https://doi.org/10.1177/19401612241308697

Tsetsura, K., Adesina, K., & Aderemi, A. (2026). Comparing fact-checking efforts across five countries in sub-Saharan Africa: The case of UNDP's iVerify. *Communication & Society*, *39*(1), 501–521. https://doi.org/10.15581/003.39.1.033

Twomey, J., Ching, D., Aylett, M. P., Quayle, M., Linehan, C., & Murphy, G. (2023). Do deepfake videos undermine our epistemic trust? A thematic analysis of tweets that discuss deepfakes in the Russian invasion of Ukraine. *PLOS ONE*, *18*(10),

*© IIARJournal. All rights reserved.*

e0291668. https://doi.org/10.1371/journal.pone.0291668

Udupa, S., Maronikolakis, A., & Wisiorek, A. (2023). Ethical scaling for content moderation: Extreme speech and the (in)significance of artificial intelligence. *Big Data & Society*, *10*(1). https://doi.org/10.1177/20539517231172424

Uwalaka, T. (2023). Mobile internet and contentious politics in Nigeria: Using the organisational tools of mobile social networking applications to sustain protest movements. *Journalism and Media*, *4*(1), 396–412. https://doi.org/10.3390/journalmedia4010026

van der Meer, T. G. L. A., Hameleers, M., & Ohme, J. (2023). Can fighting misinformation have a negative spillover effect? How warnings for the threat of misinformation can decrease general news credibility. *Journalism Studies*, *24*(6), 803–823. https://doi.org/10.1080/1461670X.2023.2187652

Weikmann, T., Greber, H., & Nikolaou, A. (2025). After deception: How falling for a deepfake affects the way we see, hear, and experience media. *The International Journal of Press/Politics*, *30*(1), 187–210. https://doi.org/10.1177/19401612241233539

Weikmann, T., & Lecheler, S. (2023). Visual disinformation in a digital age: A literature synthesis and research agenda. *New Media & Society*, *25*(12), 3696–3713. https://doi.org/10.1177/14614448221141648

Wersényi, G., & Koech, V. (2026). Content credentials: Trust issues, technical solutions and future perspectives. *Infocommunications Journal*, *18*(1), 62–70. https://doi.org/10.36244/icj.2026.1.7

Yeandle, A. (2026). The political consequences of Africa's mobile revolution. *American Journal of Political Science*, *70*(3), 1042–1061. https://doi.org/10.1111/ajps.12991

*© IIARJournal. All rights reserved.*