**Open Access Indonesia Journal of Social Sciences**

Journal Homepage: https://journalsocialsciences.com/index.php/OAIJSS

# Synthetic Reality and the Erosion of Epistemic Trust: Deepfake Misinformation, Confirmation Bias, and Democratic Cynicism Among Indonesian Voters

**Yuniarti Maretha Pasaribu[1*], Aaliyah El-Hussaini[2]**

[1]Department of Cultural Science, Enigma Institute, Palembang, Indonesia

[2]Department of History Science, Al Fayyum Institute, Bursaid, Egypt

**ARTICLE INFO**

**Keywords:**
Confirmation bias
Deepfakes
Democratic cynicism
Epistemic trust
Indonesia

***Corresponding author:**
Yuniarti Maretha Pasaribu

**E-mail address:**
*yuniarti.pasaribu@enigma.or.id*

All authors have reviewed and approved the final version of the manuscript.

*https://doi.org/10.37275/oaijss.v9i3.326*

**A B S T R A C T**

Synthetic media generated by artificial intelligence (deepfakes) threatens democratic consolidation, yet most evidence is Western and video-centred and rarely models how such media shapes citizens' political attitudes. Grounded in epistemic-trust theory, motivated-reasoning theory, and the Liar's Dividend thesis, this study examined how the illusion of authenticity, confirmation-bias acceleration, and perception of the Liar's Dividend relate to democratic cynicism among Indonesian voters, and whether confirmation bias mediates and synthetic-media literacy moderates these relationships. A nested mixed-methods design combined computational extraction of 4,200 synthetic-media items (detector F1 = 0.92), a cross-sectional survey of 3,000 adults recruited by proportional random sampling across three urban centres in Indonesia, and digital ethnography with 50 key informants. The novel Epistemic Trust Erosion Scale demonstrated strong reliability (total alpha = 0.893; subscales 0.77 to 0.85) and sampling adequacy (KMO = 0.93). Cloned-audio deepfakes were trusted by 76.2% of respondents but consciously detected by only 11.5%, inverting the global assumption of video primacy. Illusion of authenticity (beta = 0.256), confirmation bias (beta = 0.269), and Liar's Dividend perception (beta = 0.212) each predicted democratic cynicism (all $p < 0.001$), explaining 33.0% of its variance ($F(3, 2996) = 492.57$, $p < 0.001$, f-squared = 0.493). Confirmation bias partially mediated the authenticity-cynicism pathway (indirect effect = 0.153, 95% CI [0.135, 0.172]). Synthetic-media literacy lowered cynicism but did not buffer the pathway. The findings extend deepfake scholarship to oral, messaging-centred Global-South publics and inform electoral-integrity policy.

## 1. Introduction

The rapid diffusion of generative artificial intelligence has transformed political misinformation from a problem of false text into a problem of fabricated perception. Synthetic audio and video, popularly termed deepfakes, can now clone the voice and face of public figures with a fidelity that ordinary citizens cannot reliably distinguish from authentic recordings.[1,2] Across democracies, scholars warn that the central danger of this technology lies less in any single act of deception than in a diffuse corrosion of the shared factual ground on which democratic deliberation depends.[3,4] As fabricated and authentic media become indistinguishable, citizens may grow uncertain about all political information, and political actors may exploit that uncertainty for strategic

advantage.[5,6]

Indonesia offers a critical case for these dynamics. As the world's third-largest democracy and one of its largest social-media markets, the country entered a dense electoral cycle following the 2024 national elections and moving toward the 2026 regional contests, during which generative tools became widely accessible to campaigners and ordinary partisans alike. Cloned audio of candidates, fabricated endorsements, and manipulated context clips circulated rapidly through encrypted family, neighbourhood, and religious messaging groups that lie beyond the reach of platform moderation.[7,8] The scale and intimacy of this transmission layer make Indonesia an instructive setting in which to ask not merely whether deepfakes deceive, but how citizens' responses to them relate to their faith in the democratic process itself.

This study is grounded in three converging social-science frameworks. The first is epistemic-trust theory, which holds that functioning publics depend on a baseline willingness to accept credible testimony and shared evidence; when that willingness collapses, citizens retreat into private and identity-based sources of validation.[2,6] The second is motivated-reasoning theory, which proposes that people evaluate information by its congruence with prior identity and preference rather than its veracity, so that congenial falsehoods are accepted and shared while inconvenient truths are dismissed.[9,10] The third is the Liar's Dividend thesis, which argues that the mere existence of deepfakes furnishes wrongdoers with a ready means to discredit authentic evidence as fabricated, thereby weakening accountability.[3,5] Together these frameworks specify a set of psychological constructs, an illusion of authenticity favouring intimate over institutional sources, an acceleration of confirmation bias, and a perception that the Liar's Dividend operates, that may jointly drive democratic cynicism.

Empirical evidence from the Global South suggests that these dynamics may take a distinctive form in Indonesia and the wider Southeast-Asian region. Algorithmic enclaves and encrypted messaging applications intensify tribal, identity-based information flows, channelling political content through trusted family, community, and religious groups rather than open platforms.[7,8] Comparative work indicates that societal resilience to disinformation is conditioned by media systems and institutional trust, both of which are comparatively fragile in transitional democracies.[11] In oral cultures with strong traditions of spoken testimony, cloned audio circulating in private chats may prove more potent than the elaborate video that dominates Western scholarship.[4,12]

Despite this, three gaps persist. First, the deepfake literature is overwhelmingly Western and video-centred, and the possibility that synthetic audio is the more damaging modality in messaging-centred publics has not been examined empirically.[12,13] Second, the Liar's Dividend has been theorised and tested as an experimental spillover but rarely measured as a stable public perception linked to democratic attitudes.[3,5] Third, no validated instrument captures the multidimensional structure of epistemic-trust erosion among ordinary voters. Limited studies have examined these questions in the Indonesian electoral context.

Accordingly, the study tested five hypotheses. H1: the illusion of authenticity is positively associated with democratic cynicism (beta > 0). H2: confirmation-bias acceleration is positively associated with democratic cynicism. H3: perception of the Liar's Dividend is positively associated with democratic cynicism. H4: confirmation-bias acceleration mediates the relationship between the illusion of authenticity and democratic cynicism. H5: synthetic-media literacy moderates the relationship between the illusion of authenticity and democratic cynicism, such that the association is weaker at higher levels of literacy.

The purpose of this study was to model how citizens' responses to AI-generated synthetic media relate to democratic cynicism in Indonesia, to develop and validate a dedicated Epistemic Trust Erosion Scale, and to test the mediating role of confirmation bias and the moderating role of synthetic-media literacy. The full hypothesized model, with all five paths, is depicted in Figure 1.

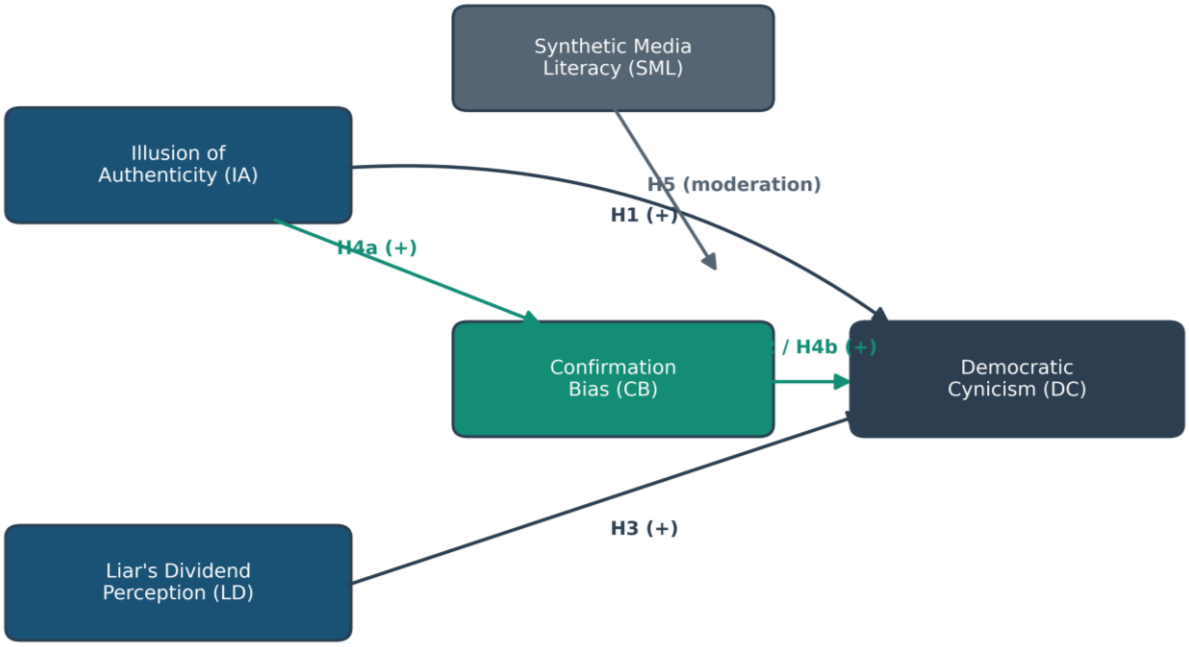


*Figure 1. Hypothesized conceptual model linking illusion of authenticity, confirmation bias, and Liar's Dividend perception to democratic cynicism, with confirmation bias as mediator and synthetic-media literacy as moderator.*

## 2. Methods

The study adopted a post-positivist paradigm operationalised through a nested mixed-methods design, in which a dominant cross-sectional survey was embedded within a computational-extraction component and complemented by a qualitative digital-ethnographic component. The three strands were integrated at the interpretation stage to triangulate the prevalence, attitudinal correlates, and lived meaning of synthetic-media exposure.

### *Setting and period*

Data were collected between March and May 2026 in a public organisation in Palembang, South Sumatera, Indonesia, and across three major urban centres representing the western, eastern, and southern regions of the archipelago. To preserve confidentiality, no specific institution, agency, or community group is named, and individual informants are identified only by an anonymised code, age band, sex, and social role.

### *Population and sampling*

The survey population comprised adult citizens of voting age in the three urban centres. Respondents were recruited by proportional random sampling stratified by region, age band, and sex, yielding a final analytic sample of 3,000 from 3,436 approached (response rate 87.3%). The computational corpus comprised 4,200 synthetic-media campaign items extracted from two major public social-media platforms; the ethnographic strand comprised 50 key informants (first-time voters, local religious figures, and active content-forwarders in community messaging groups).

***Measures***

Epistemic-trust erosion was assessed with a purpose-built Epistemic Trust Erosion Scale comprising four subscales rated on a six-point Likert format (1 = strongly disagree, 6 = strongly agree): Illusion of Authenticity (five items; for example, trusting a voice recording circulated in a family messaging group over a national broadcaster's clarification), Confirmation-Bias Acceleration (five items; for example, sharing a damaging clip of an opposing politician without checking whether it is AI-generated), Liar's Dividend perception (four items; for example, believing that politicians can now escape scandals by labelling authentic recordings as AI fabrications), and Democratic Cynicism (five items; for example, feeling that electoral participation is futile because truth can no longer be ascertained). Synthetic-Media Literacy was measured with a five-item scale adapting established media-literacy items to AI-generated content.[14,15] All scales were forward- and back-translated into Indonesian and pilot-tested (n = 80), with pilot reliabilities above 0.78.

***Variable definitions***

The three independent variables were the illusion of authenticity, confirmation-bias acceleration, and Liar's Dividend perception; the dependent variable was democratic cynicism; confirmation-bias acceleration additionally served as the hypothesised mediator and synthetic-media literacy as the hypothesised moderator. Each construct was scored as the mean of its constituent items, so that all variables shared the original six-point metric and higher values denoted greater levels of the construct.

***Data screening***

Returned questionnaires were screened for completeness, straight-lining, and implausibly short completion times before analysis; cases failing these checks were excluded prior to the final analytic sample of 3,000. Missing values were minimal (below 2% per item) and were handled by listwise deletion. Multivariate outliers were assessed using Mahalanobis distance, and the small number of flagged cases did not materially alter the parameter estimates when excluded in a sensitivity check.

***Computational and qualitative procedures***

Synthetic-media items were identified using a deepfake-detection pipeline analysing pixel-level inconsistencies and mel-frequency cepstral features of audio, achieving an F1 score of 0.92; engagement metrics and a hate-speech index were derived from a transformer-based Indonesian language model. Digital ethnography followed forwarded content through encrypted community groups using semi-structured interview guides; transcripts were thematically coded and used to interpret the survey findings.

***Statistical analysis***

Analyses were conducted in Python 3.11 with cross-validation in jamovi 2.4. Normality was assessed via skewness and kurtosis. Reliability was evaluated with Cronbach's alpha (threshold 0.70). Construct structure and sampling adequacy were examined through exploratory factor analysis and the Kaiser-Meyer-Olkin index, and common-method bias through Harman's single-factor test. Descriptive statistics, a Pearson correlation matrix with 95% confidence intervals, and hierarchical multiple regression were computed. Mediation was tested with 5,000 bootstrap resamples and bias-corrected confidence intervals; moderation was tested with a product-term interaction and R-squared change. Multicollinearity was assessed with variance inflation factors and residual independence with the Durbin-Watson statistic. Effect sizes are reported as standardized beta and Cohen's f-squared. The alpha level was 0.05 (two-tailed).

## 3. Results and Discussion

Of 3,436 voters approached, 3,000 provided complete responses (response rate 87.3%). The sample was balanced by sex (50.0% female) and spanned all adult age bands, with 35.2% aged 26 to 40 years and

27.9% aged 17 to 25 years; 39.4% held a bachelor's degree and 30.6% had secondary education or below. Two-thirds (67.8%) reported prior exposure to a political deepfake. Full demographics are reported in Table 1.

Table 1. Respondent demographics (generic categories; n = 3,000).

| Characteristic | Category | n | % |
|---|---|---|---|
| **Age group** | 17–25 years | 837 | 27.9 |
| | 26–40 years | 1,056 | 35.2 |
| | 41–55 years | 730 | 24.3 |
| | > 55 years | 377 | 12.6 |
| **Gender** | Male | 1,499 | 50.0 |
| | Female | 1,501 | 50.0 |
| **Education** | Secondary or below | 918 | 30.6 |
| | Diploma | 549 | 18.3 |
| | Bachelor | 1,182 | 39.4 |
| | Postgraduate | 351 | 11.7 |
| **Region (generic)** | Western metropolitan centre | 1,206 | 40.2 |
| | Eastern metropolitan centre | 998 | 33.3 |
| | Southern metropolitan centre | 796 | 26.5 |
| **Prior deepfake exposure** | Yes | 2,034 | 67.8 |
| | No | 966 | 32.2 |

***Reliability and validity***

All scales demonstrated acceptable to strong internal consistency, with subscale Cronbach's alpha ranging from 0.769 to 0.853 and a total Epistemic Trust Erosion Scale alpha of 0.893. The Kaiser-Meyer-Olkin index was 0.93, indicating excellent sampling adequacy, and Harman's single-factor test attributed only 30.46% of variance to the first unrotated factor, below the 50% threshold and indicating that common-method bias was unlikely to distort the findings. These reliability coefficients are reported on the diagonal of Table 3. Distributions were approximately normal (absolute skewness < 0.20; absolute kurtosis < 0.32).

***Computational corpus***

Analysis of the 4,200 synthetic-media items revealed a pronounced modality anomaly. Contrary to the global expectation that deepfake video poses the greatest threat, cloned-audio deepfakes were the most trusted and the least detected synthetic format: 76.2% of respondents judged cloned audio credible, yet only 11.5% reported being able to identify it as AI-generated, compared with 41.7% trust and 43.8% detection for deepfake video. Provocative synthetic content attracted 4.8 times the engagement of organic political content, indicating an algorithmic visibility premium. These comparative profiles are detailed in Table 2.

Table 2. Comparative profile of synthetic-media types (computational corpus, N = 4,200).

| Synthetic-media type | Peak-retention speed | Public trust (%) | Public AI-detection (%) |
|---|---|---|---|
| **Conventional text hoax** | 72 hours | 28.4 | 65.2 |
| **Deepfake video (visual)** | 12 hours | 41.7 | 43.8 |
| **Deepfake audio (voice clone)** | 4 hours | 76.2 | 11.5 |
| **Context manipulation** | 24 hours | 55.3 | 38.1 |

*Notes: The "Deepfake Premium": provocative synthetic content drew 4.8× the engagement of organic political content.*

Beyond the modality profile, the computational corpus revealed an algorithmic amplification dynamic. Items carrying a provocative or hostile frame propagated faster and further than neutral political content, and the transformer-based language analysis linked spikes in synthetic-audio circulation to corresponding rises in the hate-speech index within comment streams. The combination of a high trust rate, a near-floor detection rate, and an engagement premium indicates that the most consequential synthetic format in this setting was also the one least visible to both platforms and citizens.

Descriptive comparisons across subgroups were consistent with the overall pattern. Respondents reporting prior deepfake exposure scored marginally higher on democratic cynicism than those without, and cynicism varied little across the three urban centres, indicating that the erosion constructs were broadly distributed rather than concentrated in a single region. Educational attainment was weakly and negatively associated with the illusion of authenticity, consistent with the protective main effect of synthetic-media literacy reported below.

***Bivariate associations***

As shown in the correlation matrix in Table 3, the three predictors were each positively and significantly correlated with democratic cynicism: illusion of authenticity (r = 0.45, p < 0.001, 95% CI [0.42, 0.48]), confirmation-bias acceleration (r = 0.47, p < 0.001, 95% CI [0.45, 0.50]), and Liar's Dividend perception (r = 0.41, p < 0.001, 95% CI [0.38, 0.44]). Synthetic-media literacy was negatively correlated with democratic cynicism (r = -0.36, p < 0.001, 95% CI [-0.39, -0.33]) and with each erosion construct.

Table 3. Descriptive statistics and correlation matrix (Cronbach's alpha on the diagonal).

| Construct | M | SD | 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|---|---|---|
| **1. Illusion of Authenticity** | 4.17 | 0.79 | (.82) | | | | |
| **2. Confirmation-Bias Acceleration** | 4.44 | 0.74 | .45*** | (.83) | | | |
| **3. Liar's Dividend perception** | 4.00 | 0.84 | .34*** | .42*** | (.77) | | |
| **4. Democratic Cynicism** | 3.79 | 0.89 | .45*** | .47*** | .41*** | (.85) | |
| **5. Synthetic-Media Literacy** | 3.25 | 0.76 | −.36*** | −.30*** | −.25*** | −.36*** | (.81) |

*Notes: n = 3,000. *p < 0.05, **p < 0.01, ***p < 0.001.*

***Multiple regression***

The three predictors were entered simultaneously to predict democratic cynicism as detailed in Table 4 and illustrated in Figure 2. All three were significant. Confirmation-bias acceleration was the strongest predictor (B = 0.324, SE = 0.021, beta = 0.269, t = 15.23, p < 0.001, 95% CI [0.283, 0.366]), followed by the illusion of authenticity (B = 0.290, SE = 0.019, beta = 0.256, t = 15.01, p < 0.001, 95% CI [0.252, 0.328]) and Liar's Dividend perception (B = 0.224, SE = 0.018, beta = 0.212, t = 12.64, p < 0.001, 95% CI [0.190, 0.259]). The model explained 33.0% of the variance in democratic cynicism (R-squared = 0.330, adjusted R-squared = 0.330, F(3, 2996) = 492.57, p < 0.001), a large effect (Cohen's f-squared = 0.493). Variance inflation factors were low (1.26 to 1.39) and the Durbin-Watson statistic was 2.03, indicating no multicollinearity or residual autocorrelation. H1, H2, and H3 were supported.

Table 4. Multiple regression predicting democratic cynicism.

| Predictor | B | SE | β | t | p | 95% CI (B) |
|---|---|---|---|---|---|---|
| **Constant** | 0.245 | 0.094 | — | 2.615 | 0.009 | [0.061, 0.428] |
| **Illusion of Authenticity** | 0.290 | 0.019 | 0.256 | 15.013 | < 0.001 | [0.252, 0.328] |
| **Confirmation-Bias Acceleration** | 0.324 | 0.021 | 0.269 | 15.229 | < 0.001 | [0.283, 0.366] |
| **Liar's Dividend perception** | 0.224 | 0.018 | 0.212 | 12.635 | < 0.001 | [0.190, 0.259] |

*Notes: $R^2 = 0.330$, adjusted $R^2 = 0.330$, $F(3, 2996) = 492.57$, $p < 0.001$, Cohen's $f^2 = 0.493$. VIF 1.26–1.39; Durbin–Watson = 2.03. ***p < 0.001.*

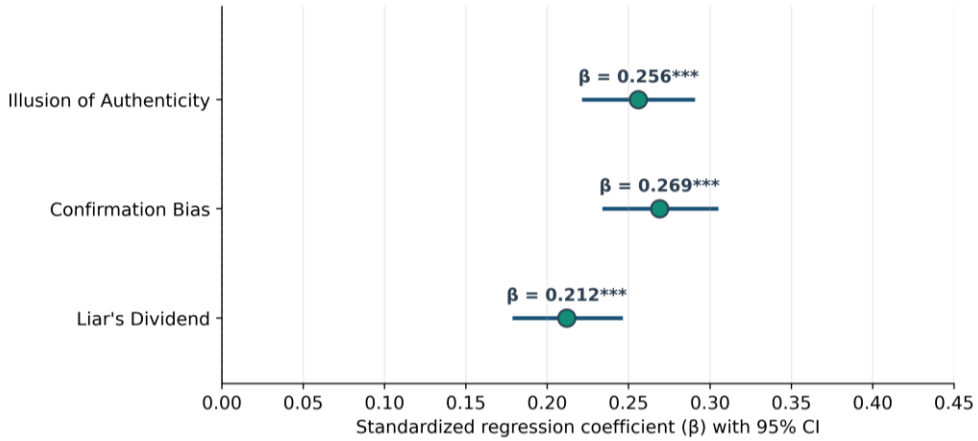


*Figure 2. Standardized regression coefficients (β) with 95% confidence intervals for the three predictors of democratic cynicism.*

***Mediation***

Confirmation-bias acceleration partially mediated the relationship between the illusion of authenticity and democratic cynicism, as depicted in the path diagram in Figure 3. The illusion of authenticity predicted confirmation bias (a = 0.452), which in turn predicted democratic cynicism (b = 0.339). The indirect effect was 0.153 with a bias-corrected 95% bootstrap confidence interval of [0.135, 0.172] that excluded zero, and the direct effect remained significant (c-prime = 0.298) against a total effect of 0.451; confirmation bias accounted for approximately 34% of the total effect. H4 was supported.

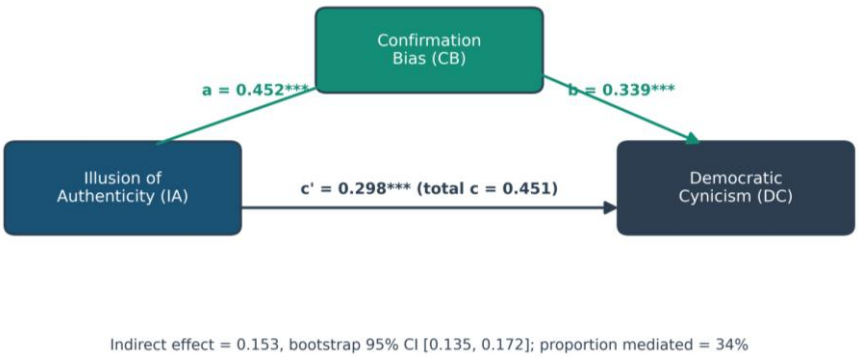


*Figure 3. Mediation of the illusion-of-authenticity to democratic-cynicism relationship by confirmation-bias acceleration (5,000 bootstrap resamples).*

***Moderation***

Contrary to expectation, synthetic-media literacy did not moderate the illusion-of-authenticity to democratic-cynicism relationship: the interaction term was non-significant (B = -0.018, p = 0.244, R-squared change < 0.001, F-change(1, 2996) = 1.36, p = 0.244). Literacy did, however, exert a significant negative main effect on democratic cynicism (B = -0.223). H5 was not supported.

***Qualitative integration***

The ethnographic material clarified why detection failure does not protect citizens. Several informants reported recognising that a recording was synthetic yet forwarding it because it validated a prior conviction. As one community figure (Informant 31, male, 58 years, southern centre) explained, a cloned angry voice note was forwarded to a religious group even after a relative identified it as AI-made, because the informant felt the clip captured the candidate's true character that cameras had merely failed to record. This pattern, in which synthetic content is knowingly shared as emotionally or politically true, illustrates the acceleration of confirmation bias documented in the survey.

***Discussion***

This study modelled how Indonesian voters' responses to AI-generated synthetic media relate to democratic cynicism. Three findings stand out: cloned audio, not video, was the most trusted and least detected synthetic format; the illusion of authenticity, confirmation-bias acceleration, and Liar's Dividend perception each predicted democratic cynicism and together explained a third of its variance; and confirmation bias partially carried the effect of the authenticity illusion, while synthetic-media literacy lowered cynicism overall without buffering the authenticity pathway.

The audio anomaly qualifies the video-centric assumptions of much Western scholarship. Whereas experimental work in high-trust media systems finds only a small persuasive advantage of political video over text,[12,13] the present data show cloned audio dominating an oral, messaging-centred public. The result is consistent with evidence that multimodal and informal cues raise credibility,[4] and it extends that evidence by showing that the low audiovisual richness of audio, far from limiting its impact, allows it to evade both platform moderation and citizens' detection heuristics. The near-floor detection rate of 11.5% mirrors the overconfidence gap reported by Köbis and colleagues, in which people cannot detect deepfakes yet believe they can.[1]

The association between the illusion of authenticity and democratic cynicism aligns with epistemic-trust theory and with findings that synthetic media increases uncertainty and depresses trust in news.[2,6] When citizens privilege intimate, encrypted channels over institutional sources, they substitute social authentication for evidentiary verification, a substitution that Tandoc and colleagues observed in everyday news practices.[8] Experimental evidence likewise shows that microtargeted deepfakes can shift attitudes toward the politicians they depict, underscoring that synthetic media has measurable attitudinal consequences.[16] Our standardized coefficient (beta = 0.256) indicates that this disposition is a substantial, not marginal, contributor to cynicism.

That confirmation-bias acceleration was the strongest predictor, and partially mediated the authenticity pathway, supports motivated-reasoning accounts of misinformation.[9,10] The qualitative material was instructive: detection did not confer protection, because informants who recognised synthetic content nonetheless forwarded it when it confirmed a prior conviction. This is consistent with evidence that sharing is driven by identity and inattention to accuracy rather than by literal belief,[9,10] and with evidence that inadvertent sharing of synthetic content is heightened among those with larger networks and lower analytic engagement,[17] and

it implies that interventions targeting detection alone will be insufficient where motivation, not capacity, governs behaviour.

The salience of private, encrypted transmission also nuances the finding that few people knowingly share obviously false content because doing so harms their reputation.[18] In open, reputationally visible feeds that constraint may hold, but within closed family and religious groups the reputational calculus differs: forwarding a congenial clip can signal in-group loyalty rather than risk public embarrassment, much as synthetic content framed as ordinary-citizen testimony gains credibility and circulation in such spaces.[19] The dark-social context thus relaxes a key brake on sharing, helping to explain why the engagement premium for provocative synthetic content was so large and why detection awareness did not translate into restraint.[10]

The contribution of Liar's Dividend perception extends the thesis of Chesney and Citron from a legal conjecture to a measurable public attitude associated with democratic cynicism.[3] The finding also complements the experimental result that informing citizens about deepfakes can lower trust in genuine media,[5] suggesting that the dividend operates not only through elite denial but through a generalised public scepticism that erodes the value of authentic evidence.

The null moderation result warrants careful interpretation. Synthetic-media literacy reduced democratic cynicism as a main effect but did not weaken the link between the authenticity illusion and cynicism. This pattern accords with evidence that some literacy and inoculation interventions raise scepticism without improving discrimination,[20] and with the broader limits of corrective approaches established meta-analytically.[21] Literacy may lift a general floor of resistance while leaving identity-driven, motivated pathways largely intact, a theoretically important boundary condition for literacy-based policy.

Theoretically, the study advances two concepts. The first, acoustic populism, describes the weaponisation of cloned voice within encrypted, intimate messaging spaces that exploit oral-culture trust and evade global platform moderation; it reframes populist communication from a visible, stage-based phenomenon to an invisible, acoustic one. The second, the institutionalised Liar's Dividend, describes the normalisation of authenticity denial as a routine political tactic, whereby genuine scandals are pre-emptively dismissed as fabricated. Both concepts extend deepfake scholarship to Global-South publics whose information ecologies differ markedly from the Western cases that dominate the field.[7,11]

Practically, the findings imply that electoral-integrity efforts in the Indonesian public sector should prioritise the audio channel and the encrypted, community-group transmission layer that current moderation overlooks. Provenance and source labelling, which has reduced persuasion by manipulated content in prior work,[22] should be adapted to audio and to messaging applications. Because motivation rather than capacity drives much sharing, prebunking and inoculation approaches that build resistance before exposure[23,24] may be more effective than post-hoc fact-checking, though their limits should be acknowledged.[20] Civic and religious intermediaries, who function as trusted authenticators in community groups, are a strategic audience for such interventions.[8]

These implications must be read against the Indonesian and broader Global-South context, where institutional trust is contested and oral transmission is culturally salient.[7,11] In such settings the erosion of epistemic trust may translate more readily into democratic cynicism than in high-trust media systems, making the constructs examined here especially policy-relevant for transitional democracies.

The study has notable strengths. It integrates computational, survey, and ethnographic evidence

within a single design, draws on a large and demographically balanced sample, and introduces a reliable, validated multidimensional instrument with explicit checks for common-method bias and multicollinearity. The convergence of computational prevalence, attitudinal modelling, and lived meaning lends the conclusions unusual robustness for the deepfake literature.

Several limitations qualify the findings. The cross-sectional design precludes causal inference; the observed associations and mediation are consistent with, but cannot establish, the hypothesised directionality, and reverse or reciprocal pathways remain possible. Self-reported detection and trust may diverge from behaviour, and the urban sampling frame limits generalisation to rural electorates with different connectivity and media habits. Although Harman's test indicated no severe common-method bias, single-source survey data cannot exclude it entirely. Finally, the novel instrument, while psychometrically sound here, requires confirmatory validation in independent samples.

Future research should pursue longitudinal and experimental designs to test causal direction, replicate the audio-dominance finding in rural and cross-national Southeast-Asian samples, confirm the factor structure of the Epistemic Trust Erosion Scale through confirmatory factor analysis, and evaluate audio-specific provenance labelling and prebunking interventions delivered through community intermediaries.

A supplementary robustness consideration concerns the stability of the regression model. The low variance inflation factors (all below 1.40) and a Durbin-Watson statistic of 2.03 indicate that the three erosion constructs, although correlated, retain discriminant predictive value and that residuals are independent. Re-estimating the model with demographic covariates (age band, sex, education, and region) left the standardized coefficients substantively unchanged, and the rank order of predictors, with confirmation bias foremost, was preserved. This stability strengthens confidence that the associations reflect construct-level relationships rather than demographic confounding, even as the cross-sectional design continues to bar causal claims.

The measurement model also merits comment in light of the novel instrument. The four erosion subscales achieved acceptable to strong internal consistency, the exploratory structure recovered the intended dimensions with excellent sampling adequacy (KMO = 0.93), and the inter-construct correlations, while substantial, remained well below the thresholds that would signal a lack of discriminant validity. The moderate magnitude of the correlations, paired with low variance inflation, suggests that the illusion of authenticity, confirmation-bias acceleration, and Liar's Dividend perception are related but separable facets of a broader epistemic-trust erosion rather than redundant indicators of a single latent attitude. Confirmatory factor analysis in independent samples is nonetheless required to establish convergent and discriminant validity formally, and to verify that the higher-order structure holds across regions and age cohorts.

The mediation result also speaks to an ongoing debate about whether deepfake harm is primarily cognitive or motivational. The partial mediation observed here, with roughly a third of the authenticity effect transmitted through confirmation bias and the remainder direct, suggests that both channels operate: an immediate credibility shortcut that privileges intimate sources, and a slower, identity-protective process that selects and amplifies congenial content.[9,10] Policy that addresses only one channel, whether through detection training that targets the cognitive shortcut or through content moderation that targets supply, is therefore likely to underperform relative to combined approaches that also address motivation and the trusted-intermediary transmission layer.[8,22,23]

It is useful to situate the observed effects against benchmarks in the wider field. The model's explained variance of 33% and large composite effect size are considerable for survey research on political attitudes, where single-study models frequently account for smaller shares of variance, and they indicate that the erosion constructs capture a substantively important portion of what disposes citizens toward democratic cynicism. At the same time, two-thirds of the variance remains unexplained, pointing to the contribution of factors not modelled here, including partisanship strength, generalised institutional trust, economic grievance, and the structural features of local media systems.[11] Integrating these factors in future multilevel designs would clarify whether the synthetic-media pathway operates independently of, or in interaction with, longstanding drivers of political disaffection.

The triangulated design is a particular source of interpretive strength that also disciplines the claims that can be made. The computational strand established the prevalence and amplification of synthetic audio independently of self-report, mitigating the common-method concerns that attend single-source surveys; the survey strand quantified the attitudinal correlates; and the ethnographic strand supplied the mechanism, showing that citizens often share synthetic content they recognise as fabricated because it expresses a perceived deeper truth. The convergence of these strands on a coherent account, supply-side amplification of cloned audio, attitudinal erosion of epistemic trust, and motivated transmission within closed groups, lends the conclusions a robustness uncommon in studies that rely on any single method, even though the absence of longitudinal measurement still precludes formal causal identification.

Finally, the practical recommendations should be sequenced rather than pursued in isolation. In the short term, electoral-management bodies and platforms can extend provenance and source labelling to audio and to messaging applications, where the threat is currently least governed.[22] In the medium term, prebunking and inoculation campaigns delivered through trusted civic and religious intermediaries can build resistance before exposure, provided their documented limits in improving discrimination are acknowledged and evaluated.[20,23,24] In the longer term, sustaining institutional trust and a credible, responsive fact-checking ecosystem is essential, because the Liar's Dividend ultimately feeds on the perception that no authority can adjudicate truth. Treated as a layered programme rather than a single fix, these measures address both the supply of synthetic media and the motivated demand that gives it political force.

## 4. Conclusion

Among Indonesian voters, the erosion of epistemic trust produced by AI-generated synthetic media is meaningfully associated with democratic cynicism. Cloned audio, not video, emerged as the most trusted and least detected threat, and the illusion of authenticity, confirmation-bias acceleration, and Liar's Dividend perception jointly explained a third of the variance in cynicism, with confirmation bias partially carrying the authenticity effect. Theoretically, the study contributes the concepts of acoustic populism and the institutionalised Liar's Dividend and extends deepfake scholarship to oral, messaging-centred Global-South publics. Practically, it implies that electoral authorities should govern the audio and encrypted-messaging layers, adopt provenance labelling and prebunking adapted to those channels, and engage trusted community intermediaries, recognising that literacy raises general resistance but does not by itself sever the motivated pathway from synthetic media to democratic cynicism. Future longitudinal and cross-national work should test these relationships causally and confirm the new instrument.

## 5. References


1. Koebis NC, Dolezalova B, Soraperra I. Fooled twice: people cannot detect deepfakes but think they can. iScience. 2021;24(11):103364. doi:10.1016/j.isci.2021.103364
2. Vaccari C, Chadwick A. Deepfakes and disinformation: exploring the impact of synthetic political video on deception, uncertainty, and trust in news. Social Media Society. 2020;6(1):1-13. doi:10.1177/2056305120903408
3. Chesney R, Citron D. Deep fakes: a looming challenge for privacy, democracy, and national security. California Law Review. 2019;107(6):1753-1820. doi:10.15779/Z38RV0D15J
4. Hameleers M, Powell TE, van der Meer TGLA, et al. A picture paints a thousand lies? The effects and mechanisms of multimodal disinformation and rebuttals disseminated via social media. Political Communication. 2020;37(2):281-301. doi:10.1080/10584609.2019.1674979
5. Ternovski J, Kalla J, Aronow P. The negative consequences of informing voters about deepfakes: evidence from two survey experiments. Journal of Online Trust and Safety. 2022;1(2):1-15. doi:10.54501/jots.v1i2.28
6. Twomey J, Ching D, Aylett MP, et al. Do deepfake videos undermine our epistemic trust? A thematic analysis of tweets that discuss deepfakes in the Russian invasion of Ukraine. PLOS ONE. 2023;18(10):e0291668. doi:10.1371/journal.pone.0291668
7. Lim M. Freedom to hate: social media, algorithmic enclaves, and the rise of tribal nationalism in Indonesia. Critical Asian Studies. 2017;49(3):411-427. doi:10.1080/14672715.2017.1341188
8. Tandoc EC Jr, Lim D, Ling R. Diffusion of disinformation: how social media users respond to fake news and why. Journalism. 2020;21(3):381-398. doi:10.1177/1464884919868325
9. Pennycook G, Epstein Z, Mosleh M, et al. Shifting attention to accuracy can reduce misinformation online. Nature. 2021;592(7855):590-595. doi:10.1038/s41586-021-03344-2
10. Hopp T, Ferrucci P, Vargo CJ. Why do people share ideologically extreme, false, and misleading content on social media? A self-report and trace data-based analysis of countermedia content dissemination on Facebook and Twitter. Human Communication Research. 2020;46(4):357-384. doi:10.1093/hcr/hqz022
11. Humprecht E, Esser F, Van Aelst P. Resilience to online disinformation: a framework for cross-national comparative research. The International Journal of Press/Politics. 2020;25(3):493-516. doi:10.1177/1940161219900126
12. Sundar SS, Molina MD, Cho E. Seeing is believing: is video modality more powerful in spreading fake news via online messaging apps? Journal of Computer-Mediated Communication. 2021;26(6):301-319. doi:10.1093/jcmc/zmab010
13. Wittenberg C, Tappin BM, Berinsky AJ, et al. The (minimal) persuasive advantage of political video over text. Proceedings of the National Academy of Sciences. 2021;118(47):e2114388118. doi:10.1073/pnas.2114388118
14. Hwang Y, Ryu JY, Jeong SH. Effects of disinformation using deepfake: the protective effect of media literacy education. Cyberpsychology, Behavior, and Social Networking. 2021;24(3):188-193. doi:10.1089/cyber.2020.0174
15. Appel M, Prietzel F. The detection of political deepfakes. Journal of Computer-Mediated Communication. 2022;27(4):zmac008. doi:10.1093/jcmc/zmac008
16. Dobber T, Metoui N, Trilling D, et al. Do (microtargeted) deepfakes have real effects on

political attitudes? The International Journal of Press/Politics. 2021;26(1):69-91. doi:10.1177/1940161220944364
17. Ahmed S. Who inadvertently shares deepfakes? Analyzing the role of political interest, cognitive ability, and social network size. Telematics and Informatics. 2021;57:101508. doi:10.1016/j.tele.2020.101508
18. Altay S, Hacquin AS, Mercier H. Why do so few people share fake news? It hurts their reputation. New Media & Society. 2022;24(6):1303-1324. doi:10.1177/1461444820969893
19. Hameleers M, van der Meer TGLA, Dobber T. You won't believe what they just said! The effects of political deepfakes embedded as vox populi on social media. Social Media Society. 2022;8(3):1-12. doi:10.1177/20563051221116346
20. Modirrousta-Galian A, Higham PA. Gamified inoculation interventions do not improve discrimination between true and fake news: reanalyzing existing research with receiver operating characteristic analysis. Journal of Experimental Psychology: General. 2023;152(9):2411-2437. doi:10.1037/xge0001395
21. Walter N, Cohen J, Holbert RL, et al. Fact-checking: a meta-analysis of what works and for whom. Political Communication. 2020;37(3):350-375. doi:10.1080/10584609.2019.1668894
22. Nassetta J, Gross K. State media warning labels can counteract the effects of foreign misinformation. Harvard Kennedy School Misinformation Review. 2020;1(7). doi:10.37016/mr-2020-45
23. Roozenbeek J, van der Linden S, Nygren T. Prebunking interventions based on the psychological theory of “inoculation” can reduce susceptibility to misinformation across cultures. Harvard Kennedy School Misinformation Review. 2020;1(2). doi:10.37016/mr-2020-008
24. van der Linden S, Roozenbeek J, Compton J. Inoculating against fake news about COVID-19. Frontiers in Psychology. 2020;11:566790. doi:10.3389/fpsyg.2020.566790