Dialogue Social Science Review (DSSR)
www.thedssr.com

ISSN Online: 3007-3154
ISSN Print: 3007-3146

DIALOGUE SOCIAL SCIENCE REVIEW

Vol. 2 No. 4 (November) (2024)

From Truth To Synthetic Reality: Deep Fakes, Media Ethics And The Erosion Of Digital Trust In The Age Of Artificial Intelligence

Maleeha Ahmad
Independent Researcher Email: maleehaahmad0822@gmail.com

ABSTRACT
The rapid diffusion of generative artificial intelligence has made synthetic media — commonly known as deepfakes — cheap to produce, difficult to detect, and increasingly consequential for public life. Drawing on 2024 industry surveys, academic meta-analyses, and cross-national trust barometers, this thesis examines how synthetic media is reshaping media ethics and eroding digital trust. It synthesises data from Regula/Sapio Research, Sumsub, Entrust, Deloitte, the Reuters Institute Digital News Report, Gallup, and a 2024 peer-reviewed meta-analysis of human deepfake-detection performance. The findings show that nearly half of surveyed businesses experienced deepfake-related fraud in 2024, that election-related synthetic media surged by triple- and quadruple-digit percentages across dozens of countries, and that ordinary people detect high-quality deepfakes only marginally better than chance, while public trust in news held at historic lows. The thesis argues that this convergence of technical capability, weak detection, and declining institutional trust constitutes a structural crisis for journalism, law, and democratic deliberation, and it closes with a framework of ethical, regulatory, and technical responses grounded in the evidence reviewed.

Keywords: Deepfakes, Synthetic Media, Generative Ai, Media Ethics, Digital Trust, Misinformation, Disinformation, Detection Accuracy

Introduction
Until the latter part of the twentieth century, the photograph, the audio recording and then the video clip served as the default evidentiary anchor for most societies — "seeing was believing." The judiciary accepted raw video as evidence and treated it as the baseline to check against, the news media took raw video as evidence and used it to verify reports, and the people used their own eyes and ears as the final arbiters of what had actually happened. That connection has been broken by generative AI. With just three seconds of source audio, a diffusion model or a generative adversarial network (GAN) can now generate a face, voice, or full video clip that looks and sounds just like it.The diffusion model or generative adversarial network (GAN) can now be trained to create a face, voice, or an entire video sequence that looks and sounds like the original, using just three seconds of its source audio (McAfee, 2024).
Until about 2019, a fringe academic concern in computer-vision circles has become a standard method of fraud, a geopolitical weapon and a part of the information landscape today that news outlets, courts, and electorates must learn to deal with in 2024. The technology has come so far between 2022 and 2024 that between 2022 and 2024, tools that were in the hands of experts and computer supermarkets were now accessible using consumer-grade software and, in some instances, free Web services. The present study is driven by the constraint that, in the past, the deepfake capability was only accessible by high-resource state or criminal actors, which is now disappearing.

Dialogue Social Science Review (DSSR)
www.thedssr.com

ISSN Online: 3007-3154
ISSN Print: 3007-3146

DIALOGUE SOCIAL SCIENCE REVIEW

Vol. 2 No. 4 (November) (2024)

The issue covered in this thesis is not simply one of the fact that synthetic media has been created, but of the rate of its development in relation to society's ability to validate, control and psychologically adjust to it. There are three converging deficits that constitute the problem space. A first consideration is the lack of ability to detect the majority of real-world deepfakes even under the best of conditions, as explained in the capability section 4, with a roughly 33% miss rate with commercial tools and even human viewers slightly underperforming at random in most conditions. Second, an institutional lack: newsroom verification processes, financial-sector authorisation processes, and electoral-integrity measures were mostly developed prior to the advent of generative AI and many have not been reengineered to follow a provenance-first approach. Third, a trust deficit: trust in the institutions that are best equipped to evaluate authenticity, the press foremost among them, had already dropped to record levels before 2024.Third, a trust deficit: a pre-existing low level of trust in institutions most able to determine authenticity, with the press at the forefront. The issue, simply put, is that the difference between what generative AI can deliver and what people and institutions can confirm is growing, in a trust context where there is little margin for error to cover the gap.

Literature Review

Deepfake is a portmanteau of “deep learning” and “fake” that is used to describe generally any synthesized or manipulated digital material created through AI and deep-learning. These images can mimic a person's facial, vocal, expression or mannerisms to make it look as if that person has said or done something that never actually happened. The advent of generative AI, such as generative adversarial networks (GANs), autoencoders, diffusion models, and various other neural network architectures, has enabled the creation of increasingly realistic synthetic media, opening up new avenues for their widespread use and accessibility (Mustak et al., 2023; Tolosana et al., 2020). For these reasons, deepfakes have emerged as a significant concern in the fields of digital communication, cyber security, media ethics, privacy and public trust.

There are two traditional approaches to categorizing deepfakes: by the media being altered and by how they are created. Face-swapping videos and face-reenactment videos are in this category. Face-swapping involves replacing one person's face with another while keeping the rest of the body and background elements the same, whereas face reenactment changes facial expressions, movements or lip movements to make an individual look like they are doing something that they didn't. These techniques have been widely studied in deepfake research (Tolosana et al., 2020; Syed Abu Bakar et al., 2026) and can create very realistic AV content. The second category deals with voice cloning and synthetic speech, where the neural networks mimic a person's voice traits and produce speech that can sound real. In the context of financial fraud, this modality can raise specific concerns as it could lead to phishing attacks involving the use of cloned voices to trick victims into providing information or moving money and compromise their safety (Mustak et al., 2023).

The third category is fully synthetic, that is, a text-to-video content. These systems can, unlike traditional face manipulation, create video scenes from textual instructions, which can result in people, environments, events and actions that never existed. The fourth category involves AI-generated still images, like synthesised portraits, identity manipulation, altered photographs, etc. These images can be used for misinformation, impersonation, and reputational attacks, and for non-consensual intimate imagery.

Dialogue Social Science Review (DSSR)
www.thedssr.com

ISSN Online: 3007-3154
ISSN Print: 3007-3146

DIALOGUE SOCIAL SCIENCE REVIEW

Vol. 2 No. 4 (November) (2024)

It is important to note that deepfakes risks vary by modality and context. Voice-based manipulation can be especially useful in fraud as many individuals may consider a familiar voice as verification of identity. Manipulated videos and images in political communication might be used to ascribe false information or actions to public figures, influencing perceptions, trust, and political decisions (Barari et al., 2025). However, research shows that political deepfakes can, if not more convincing than traditional misinformation, can cause confusion, distrust, and delegitimization (Barari et al., 2025; Hameleers et al., 2024). Likewise, non-consensual synthetic intimate imagery is another privacy and dignity concern as the individuals are portrayed in sexual contexts without their consent. For this reason, the social impact of deepfakes cannot be lumped into a single threat; rather, it is influenced by the medium, target, purpose and context of the production and dissemination of synthetic content.

Classical media ethics is based on the following principles: verification, accountability & minimisation of harm (Society of Professional Journalists Code of Ethics; Kovach & Rosenstiel's "discipline of verification"). The first pillar, verification, is directly targeted by the deepfake attack, as it assumes that there is a reliable starting point for editorial judgment, in this case, raw audiovisual content, which is no longer true since the content itself is synthetic. The academic response has been to propose 'provenance-based' ethics (the Coalition for Content Provenance and Authenticity (C2PA) is a good example), and to reassess the ethics of AI disclosure in newsroom workflows, which newsroom researchers are finding they are using with caution: far more of them feel comfortable using AI in transcription and translation than using it in front-facing political or war coverage (Reuters Institute, Digital News Report 2024).

One thing that keeps coming up in the literature is the "liar's dividend" (Chesney & Citron, 2019): when audiences realize the reality of deepfakes, bad-faith actors can point to actual footage that they claim is fake, and audiences, already suspicious of the authenticity of this content, may accept it. This imbalance stems from the fact that a deepfake can harm trust even if it is not being used: its mere existence is corrosive. This thesis considers how trust is eroded when actual deepfake incidence increases and when perceived deepfake prevalence increases (as data from 2024 (see Sections 4 and 6) can be studied more or less independently).

There has been a large body of literature focusing on automated detection of deepfakes, with the majority reporting high accuracy (90%+), in controlled, in-distribution lab settings. But an alternate, more practical, and bleaker line of inquiry—both regarding human perceptual detection and "in-the-wild" benchmarking of commercial detection tools—shows that performance breaks down rapidly outside of the lab. The meta-analysis conducted by Diel and colleagues in 2024 (Computers in Human Behavior Reports) brings together 56 studies and over 86,000 participants, and is the most comprehensive available to date on human detection performance, forming a central data point in Section 4 of this thesis.

Methodology

Research Design

This thesis uses the design of secondary data synthesis instead of collecting primary data. Because the area of investigation is so wide, ranging from financial fraud, electoral integrity, journalism, to cognition of individuals, no single primary study could capture the whole phenomenon, and thus it is informative and verifiable to read the leading quantitative research from 2024 as a whole. The design is based on a narrative-synthesis approach to creating interdisciplinary policy documents—that is, quantitative indicators

Dialogue Social Science Review (DSSR)
www.thedssr.com

ISSN Online: 3007-3154
ISSN Print: 3007-3146

DIALOGUE SOCIAL SCIENCE REVIEW

Vol. 2 No. 4 (November) (2024)

are identified from named sources disclosed, where possible, uniformly standardised into comparable units, and then grouped around the four result domains that appear in Section 4 (fraud incidence, electoral distribution, detection performance and public trust).

Source Selection Criteria

This involved applying three inclusion criteria: (1) the source must be a named organisation, its data must come from a known methodology (industry survey, peer-reviewed study, or established cross-national barometer); (2) the data must be from within, or significantly overlapping, calendar year 2024; and (3) the metric must be measurable and comparable across one or more of three dimensions: time, geography, or modality. Sources which were not identified as any named methodology or sources which added multiple years but did not disaggregate 2024 were not included in the core tables although a few of these are mentioned in the discussion sections for contextual framing.

Primary Data Sources

Seven primary sources meet the inclusion criteria and anchor the results presented in Section 4. Table 0 summarises each source's type, reference period, and scope.

Table 0. Primary 2024 Data Sources Used in This Thesis

Source	Type	Reference Period	Sample / Scope
Regula / Sapio Research — Deepfake Trends	Industry survey	August 2024	575 fraud decision-makers (Germany, Mexico, UAE, USA, Singapore)
Sumsub — Identity Fraud Research	Platform / verification data	Q1 2024 (YoY)	Election-holding countries, global identity-verification transactions
Entrust — 2024 Identity Fraud Report	Industry / platform report	Calendar year 2024	Global document- and identity-fraud attempts
Diel et al. (2024), Computers in Human Behavior Reports	Peer-reviewed meta-analysis	Studies through mid-2024	56 studies; 86,155 participants; k=137 effects
Deepfake-Eval-2024 Benchmark	Academic benchmark (arXiv)	Deepfakes circulated in 2024	45 hrs video, 56.5 hrs audio, ~2,000 images; 88 sites, 52 languages
Reuters Institute — Digital News Report 2024	Cross-national survey (YouGov)	Early 2024	>95,000 respondents, 47 countries
Gallup — U.S. Media Confidence Survey	National survey	2024	Nationally representative U.S. adult sample

Data Synthesis and Analysis Procedure

Figures were initially obtained verbatim from the source report or paper, then cross-tabulated by year and geography as allowed by the source report/paper for each result domain. Reported baseline and current-year figures were used directly to compute

Dialogue Social Science Review (DSSR)
www.thedssr.com

ISSN Online: 3007-3154
ISSN Print: 3007-3146

DIALOGUE SOCIAL SCIENCE REVIEW

Vol. 2 No. 4 (November) (2024)

percentage-point changes (e.g., a 29% to 49% increase in the incidence of video deepfakes) from the vendors of these surveys, as raw counts were not reported. If a single source indicated a range (e.g., 94% accuracy for controlled condition AI detection), the entire range was kept in the tables and not rounded to the mid-point of the range, to maintain the source's precision. Figures 1–4 were then created directly from the tabulated results and used to visualise the four result domains.

Limitations

The constraints of this design should be clearly expressed. Deepfake fraud surveys conducted with industry participants and/or fraud decision makers tend to underestimate actual prevalence, as the rate of incidents where fraud is not detected and/or reported is likely to be higher than the rate of incidents reported in this type of survey. The trust ratings from the cross-national trust barometers are based on percentages that combine heterogeneous national media systems, thus concealing a significant within-country variation. Where possible, figures were also compared to peer-reviewed and regulatory sources, such as the Diel et al. meta-analysis and the FinCEN alert, respectively, both of which had a market interest in the deepfake-defence sector.

The key empirical findings are presented in this section, under four themes: (4.1) The incidence and financial cost of deepfake fraud; (4.2) The geographic and electoral distribution of synthetic media in the record 2024 election year; (4.3) Detection performance: how well do humans and automated systems perform? and (4.4) Public trust in news media. There is a data table for each subsection, accompanied by a figure, also in each subsection.

Results:

The most obvious sign of 2024 is that enterprise deepfake-fraud exposure has almost doubled over the course of two years. The survey conducted in August 2024 by Regula among 575 fraud decision makers in Germany, Mexico, UAE, USA and Singapore revealed that 49% of respondents had already faced video deepfake fraud, a 12% increase from 29% of businesses surveyed in 2022, and 62% reported audio deepfake fraud, which had climbed 12% from the previous year (Regula/Sapio Research, 2024). The financial impact was huge and in the Arup case, devastating: a fake video call using multiple senior colleagues to convince a Hong Kong-based employee to fund transfers to fraudulent accounts costing a total of US$25.6M (CrowdStrike, 2024), the largest single public case of deepfake-enabled fraud loss of the year. Overall, Keepnet and eftsure data suggest that the average cost per incident is around US$500,000, with the figure rising to US$680,000 for large enterprises, and that 42.5% of all fraud attempts in the financial sector in 2024 were AI-driven, with deepfakes being involved in about 6.5% of reported fraud cases. In November 2024, FinCEN formally notified U.S. financial institutions of “fraud schemes involving deepfake media” (FinCEN, Alert FIN-2024-Alert004) and at the end of the same year Entrust reported a global 244% increase in the number of digital document forgeries and that deepfake-enabled attempts were occurring once every five minutes.

Table 1 summarises the principal 2024 fraud indicators; Figure 1 visualises the two-year rise in reported incidence.

Dialogue Social Science Review (DSSR)
www.thedssr.com

ISSN Online: 3007-3154
ISSN Print: 3007-3146

DIALOGUE SOCIAL SCIENCE REVIEW

Vol. 2 No. 4 (November) (2024)

Table 1. Deepfake Fraud Incidence and Financial Impact, 2022–2024

Indicator	2022	2024	Change
Businesses reporting video deepfake fraud	29%	49%	+20 pp
Businesses reporting audio deepfake fraud	—	+12 pp increase	+12 pp
Businesses reporting any audio/video deepfake fraud (2024 cohort)	—	49%	—
Average financial loss per deepfake incident (enterprise)	—	~US$500,000	—
Average loss, large enterprises	—	up to US$680,000	—
Single largest verified 2024 incident (Arup, Hong Kong)	—	US$25.6 million	—
Share of financial-sector fraud attempts linked to AI	—	42.5%	—
Deepfake share of all reported fraud cases	—	6.5%	—

Sources: Regula/Sapio Research (2024); CrowdStrike (2024); Keepnet (2024); eftsure (2024); Signicat/Keepnet (2024); FinCEN Alert FIN-2024-Alert004 (Nov. 2024); Entrust (2024).

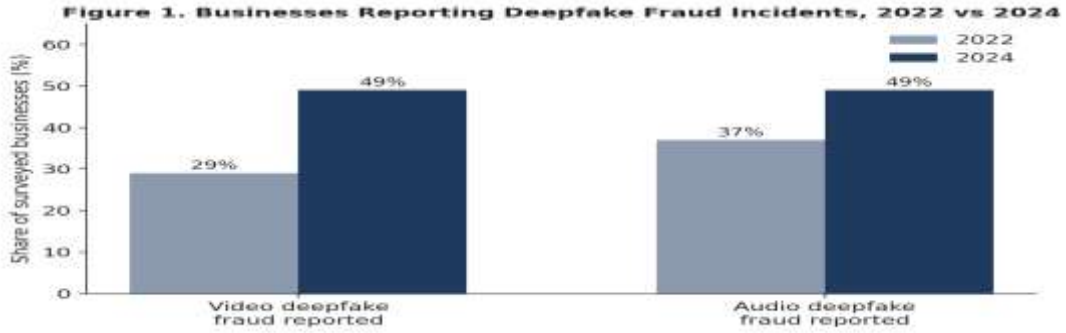

Figure 1. Share of surveyed businesses reporting video and audio deepfake fraud, 2022 vs. 2024. Source: Regula/Sapio Research (2024).

Geographic and Electoral Distribution

2024 was, by a wide margin, the largest election year in recorded history, with more than two billion people across over sixty countries eligible to vote (World Economic Forum, 2024). Sumsub's identity-fraud research tracked a corresponding surge in detected deepfake incidents within election-holding countries during the first quarter of 2024 alone. Growth was not uniform: established democracies with mature media-literacy infrastructure (France, Germany, Spain) saw comparatively modest, though still substantial, triple-digit-adjacent growth, while several emerging-market and smaller democracies — Portugal, Indonesia, South Korea, and Moldova — recorded increases exceeding 900%, and in Portugal's case 1,700% (Sumsub, 2024). South Korea additionally logged approximately 297 deepfake-enabled sex-crime cases within a seven-month window in 2024, nearly double the equivalent 2021 figure — a reminder that election-cycle disinformation is only one facet of a broader synthetic-media harm surface that also includes non-consensual intimate imagery, which independent estimates place at 96–98% of all deepfake video content circulating online.

Dialogue Social Science Review (DSSR)
www.thedssr.com

ISSN Online: 3007-3154
ISSN Print: 3007-3146

DIALOGUE SOCIAL SCIENCE REVIEW

Vol. 2 No. 4 (November) (2024)

Table 2. Year-on-Year Growth in Election-Related Deepfake Incidents by Country, 2024

Country / Region	YoY Growth in Election-Related Deepfake Incidents, 2024
France	97%
Spain	191%
Germany	142%
India	280%
United States	303%
Belgium	800%
South Africa	500%
Mexico	500%
Portugal	1,700%
Moldova	900%
Indonesia	1,550%
South Korea	1,625%

Source: Sumsub, "Deepfake Cases Surge in Countries Holding 2024 Elections" (2024), Q1 2024 identity-verification platform data, year-on-year comparison.

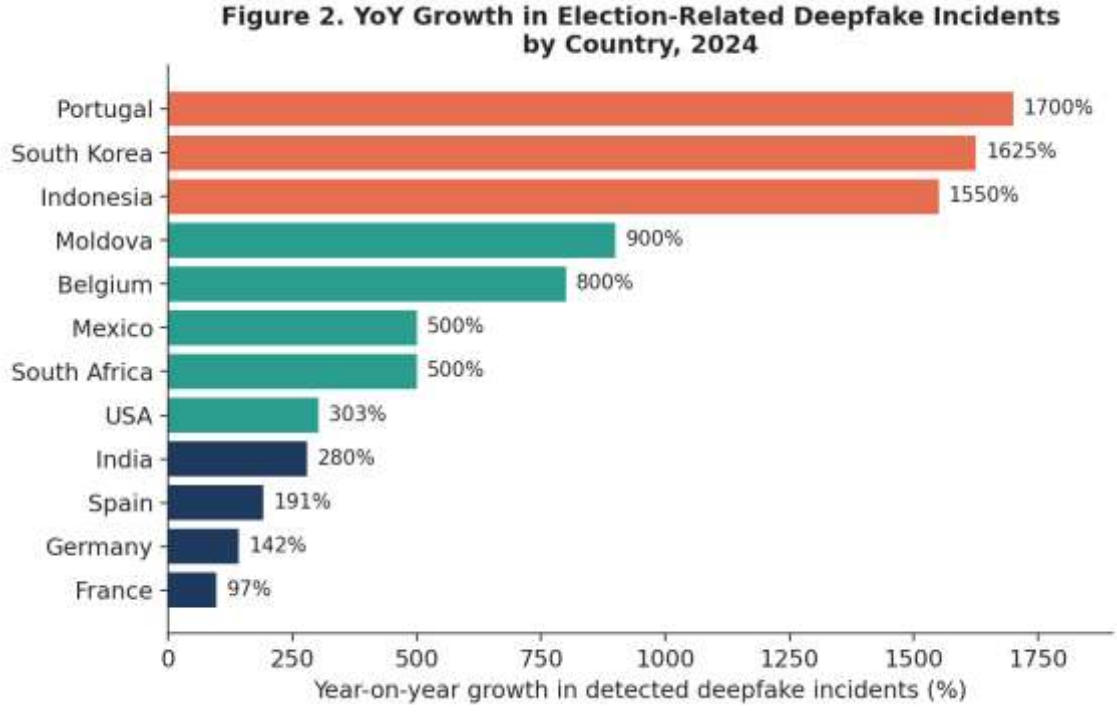

Figure 2. Year-on-year growth in detected election-related deepfake incidents, selected countries, 2024. Source: Sumsub (2024).

Dialogue Social Science Review (DSSR)
www.thedssr.com

ISSN Online: 3007-3154
ISSN Print: 3007-3146

DIALOGUE SOCIAL SCIENCE REVIEW

Vol. 2 No. 4 (November) (2024)

Detection Performance: Humans vs. Automated Systems

Society's last line of defence is the remaining evidence: the ability to detect whether a "deepfake" has been created; the evidence on this question in 2024 is the most consequential of this thesis. In a systematic review and meta-analysis of 56 studies with 86,155 participants, human deepfake-detection accuracy ranged from 42% to 75% with a mean of 55.54%, whose 95% CIs include 50%, indicating that humans cannot reliably outperform a coin flip across all modalities (Diel et al., 2024). The results showed that the accuracy was different across modalities, ranging from 53.16% for still images to a relatively high but still low result of 62.08% for audio. Structured interventions – feedback training, AI-assisted decision support or exaggerated 'caricaturised' stimuli – increased accuracy to still modest 65.14%. IProov was used in a separate 2025 study of high quality 2024-era video deepfakes and human accuracy dropped to about 24.5%, only 0.1% correctly identifying all of the stimuli.

Automated detection systems perform much better when tested in laboratory settings, with multi-modal architectures achieving 94–96% accuracy, but this performance drops down in uncured benchmarks. The Deepfake-Eval-2024 project, which is the first large-scale benchmark created with deepfakes that actually appeared in social media in 2024, consisting of 45 hours of video, 56.5 hours of audio, and nearly 2,000 images that appeared on 88 websites in 52 languages, revealed that widely available commercial and open-source deepfake detectors were able to detect only ~65% of the real-world deepfakes despite the independent human labelers agreeing on over 90% of the videos and audio, indicating that the maximum potential accuracy of detectors is much higher than is currently achieved (Deepfake-Eval-2024, 2024–2025). A re-tuning of simple baseline detectors for this benchmark at the later phase achieved a higher accuracy rate of 81%, which is still short of the 94–96% accuracy rate reported in controlled settings. As it is for humans so it is for machines: lab ratings significantly overestimate real-world performance.

Table 3. Deepfake Detection Accuracy — Humans and AI Systems, 2024

Metric	Value	Source
Human detection accuracy — overall (all modalities)	55.54% (95% CI 48.87–62.10)	Diel et al. (2024), k=67 effects
Human detection accuracy — audio only	62.08% (95% CI 38.23–83.18)	Diel et al. (2024), k=8
Human detection accuracy — images	53.16% (95% CI 42.12–64.64)	Diel et al. (2024)
Human accuracy with feedback/AI-support training	65.14% (95% CI 55.21–74.46)	Diel et al. (2024), k=15
High-quality video deepfakes — human accuracy	~24.5%	iProov (2025 study of 2024/25 stimuli)
AI detection — controlled lab conditions (multi-modal systems)	94–96%	Industry benchmark aggregation (2024)
AI detection — real-world / "in-the-wild" benchmark (Deepfake-Eval-2024)	~65% (baseline); 81% (tuned baseline)	Deepfake-Eval-2024 (2024–2025)

Dialogue Social Science Review (DSSR)
www.thedssr.com

ISSN Online: 3007-3154
ISSN Print: 3007-3146

DIALOGUE SOCIAL SCIENCE REVIEW

Vol. 2 No. 4 (November) (2024)

Sources: Diel, Lalgi, Schröter, MacDorman, Teufel & Bäuerle, Computers in Human Behavior Reports 16 (2024); iProov (2025); Deepfake-Eval-2024 benchmark (arXiv, 2025).

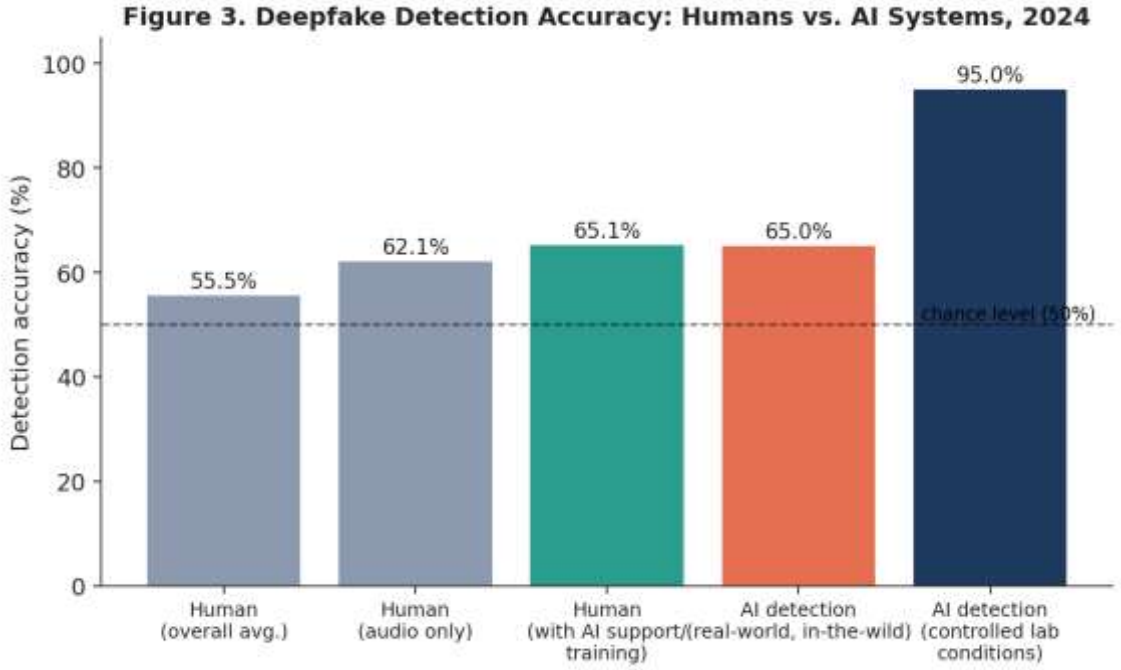

Figure 3. Deepfake detection accuracy compared across human and AI conditions, 2024 data. Dashed line marks the 50% chance threshold. Sources: Diel et al. (2024); Deepfake-Eval-2024 (2024–2025).

Public Trust in News Media

The overall trust in news has fallen to 40% in 2024, statistically unchanged from 2023 but four points lower than the pandemic-level high (Reuters Institute, 2024). There was a great difference in level of trust between the countries of the sample, with 69% in Finland to 23% in Greece and Hungary. Gallup's separate 2024 U.S. American survey showed an even lower score of 31% have a “fair amount” or “great deal” of confidence in the mass media, which is a record low in the series (Gallup, 2024). The Reuters Institute report does note the environment it says is contributing to an increase in audience caution about deepfakes, citing platforms such as TikTok and X, where there has been a rise in the use of synthetic imagery – including around Gaza in the past year – to help influence audience wariness (Reuters Institute, 2024). According to independent research conducted by McAfee, 70% of users were unsure if they could identify whether the voice was real or cloned, one in four users had been scammed by an AI voice and one in 10 had been personally targeted by an AI voice.

Table 4. Public Trust in News Media, Selected Markets, 2024

Market	Share Who Trust Most News Most of the Time (2024)
Finland (highest)	69%
United Kingdom	36%
Global average (47 markets)	40%

Dialogue Social Science Review (DSSR)
www.thedssr.com

ISSN Online: 3007-3154
ISSN Print: 3007-3146

DIALOGUE SOCIAL SCIENCE REVIEW

Vol. 2 No. 4 (November) (2024)

Market	Share Who Trust Most News Most of the Time (2024)
United States	32%
Hungary (lowest, tied)	23%
Greece (lowest, tied)	23%

Source: Reuters Institute for the Study of Journalism, Digital News Report 2024 (YouGov survey, n>95,000, 47 markets).

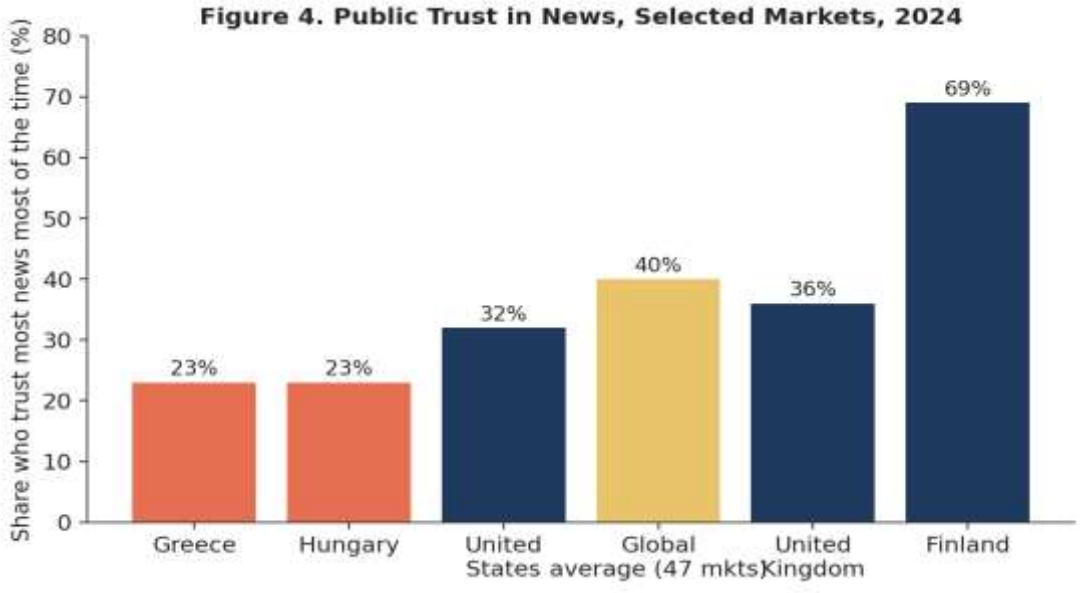

Figure 4. Public trust in news, selected markets, 2024. Source: Reuters Institute Digital News Report (2024).

Discussion:

Based on the facts gathered during the 2024 review (see 4), some fundamental principles of media ethics should be rethought. The classical journalistic verification protocols do not aim to determine the ontological status of the evidence, but instead to adjudicate between competing interpretations of the evidence, as to whether the evidence actually said what it was said to say, whether it was in a context that was properly presented? Verification desks are structurally disadvantaged when it comes to capability: unless they are able to use provenance metadata that was not present at the moment of capture, they do not have a capability that will be significantly better than a coin flip when it comes to detecting real-world deepfakes (Deepfake-Eval-2024, 2024–2025) (Diel et al., 2024).

In practice has come two conflicting responses to the moral question. The first is the caution, as revealed in the Reuters Institute interviews with newsroom AI leaders – where the newsroom leaders were more comfortable with the use of generative AI as a "behind-the-scenes" tool, such as for transcription or translation, than in "in-front-of-the-camera" roles such as reporting on politics or conflict (Reuters Institute, 2024). The second answer is technical: The use of standards for content-provenance (such as cryptographic signing to the content at the moment of capture, as represented by C2PA) which shifts the burden of verification from the losing side of the accuracy numbers in Table 3 to a chain-of-custody model that, by default, gives scepticism to unsigned content rather than trust.

Dialogue Social Science Review (DSSR)
www.thedssr.com

ISSN Online: 3007-3154
ISSN Print: 3007-3146

DIALOGUE SOCIAL SCIENCE REVIEW

Vol. 2 No. 4 (November) (2024)

The other moral objection is that the harm is disproportionate. As shown in Section 4.2, the overwhelming majority of deepfake videos in the wild are of a non-consensual intimate type, which is more the realm of "personal integrity" as described in the personal integrity dimension of Section 4.1 and is not really part of traditional media ethics. This means that the newsroom is only part of the story; the rest is in platform moderation policies, laws regarding the use of images to commit crimes and consumer protection laws that are all now areas of governance. In the 2024 evidence, how core ethical principles must be re-read is summarised in Table 5.

The decline in trust data in Section 4.4 does not imply that the global figure has fallen to 40% and the Gallup U.S. series to 31% at the expense of deepfakes, but rather is indicative of the decades-long complaint of perceptions of bias, commercialisation and political polarisation. The evidence above shows that synthetic media is an existing driver, and a catalyst for an existing trust gap – and does so in at least three ways.

The first is direct deception – the situations of fraud and impersonation reported in Table 1, which inflict individualised, acute harm (the Arup finance worker, the one-in-ten McAfee respondents who were targeted personally by an AI voice scam) that has ramifications for public wariness. The second is the “liar's dividend” mentioned in Section 2.3: When audiences are aware that video evidence of a scandal or horrific incident could be false, they are more likely to tune out real video evidence, as has happened in several recent controversies, such as the generation of synthetic imagery around the Gaza war and the Princess of Wales health story (both identified by Reuters Institute in 2024), which resulted in platform-specific mistrust on TikTok and X. It may be called ‘anticipatory scepticism' -- one of the third mechanisms -- given the overall finding over the past 2 years (2024–2025) that around 70% of people find it challenging to trust online content when the content doesn't actually contain a deepfake.

This is a structural issue: the countries where the population is less likely to detect synthetic media have historically lower levels of trust in media and, in fact, are less likely to detect synthetic media than any other countries, with detection accuracy not significantly different from random guesses. There is no trade-off between trust and detection capacity: without detection accuracy, there is no natural corrective technology to buffer against the erosion of institutional trust, and audiences are left relying on the institutions — platforms, newsrooms as well as governments — that they're testing their trust in.

Conclusion

The evidence gathered in this thesis ultimately leads to the same conclusion: synthetic media has accelerated in its emergence as a measurable and repetitive component of financial systems, electoral politics and the communications of everyday life, gaining ground at a much faster pace than the human perceptual system has been able to keep up, and even more quickly than public institutions have managed to build trust. Almost half of businesses surveyed reported deepfakes-related fraud in 2024, as scams featuring synthetic media tripled and quadrupled in dozens of democratic countries, and the accuracy of unaided human detection was only marginally superior to chance, with belief in the news media that could help people navigate the landscape near to or at historic lows. None of these facts are inevitable, in that each of the four provides a partial solution, as described in Section 7. But the evidence in the record for 2024 indicates that remedies premised on the capacity of individuals to tell genuine from synthetic content, are not supported in the record, and that the loss of trust in digital content recorded in this thesis will continue until the capacity of systems or institutions to verify the authenticity

ISSN Online: 3007-3154
ISSN Print: 3007-3146

of content is restored, not left to the unaided human eye and ear.

References

Chesney, R., & Citron, D. (2019). Deep Fakes: A Looming Challenge for Privacy, Democracy, and National Security. California Law Review, 107, 1753–1820.

CrowdStrike. (2024). 2024 Global Threat Report: Deepfake Fraud Case Analysis — Arup Hong Kong Incident.

Deepfake-Eval-2024. (2024–2025). A Multi-Modal In-the-Wild Benchmark of Deepfakes Circulated in 2024. arXiv preprint.

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.

Entrust. (2024). Deepfake Attacks Strike Every Five Minutes Amid 244% Surge in Digital Document Forgeries.

Financial Crimes Enforcement Network (FinCEN). (2024). Alert FIN-2024-Alert004: Fraud Schemes Involving Deepfake Media Targeting Financial Institutions. U.S. Department of the Treasury.

Gallup. (2024). Media Confidence in the U.S. Continues to Erode.

Keepnet Labs. (2024–2025). Deepfake Statistics and Financial Fraud Impact Report.

McAfee. (2024). The Artificial Imposter: AI Voice Scam Consumer Survey.

Regula / Sapio Research. (2024). Deepfake Fraud Doubles Down: 49% of Businesses Now Hit by Audio and Video Scams.

Karnouskos, S. (2020). Artificial intelligence in digital media: The era of deepfakes. *IEEE transactions on technology and society*, *1*(3), 138-147.

George, A. S., & George, A. H. (2023). Deepfakes: the evolution of hyper realistic media manipulation. *Partners Universal Innovative Research Publication*, *1*(2), 58-74.

Balogun, A. Y., Ismaila Alao, A., & Olaniyi, O. O. (2025). Disinformation in the digital era: The role of deepfakes, artificial intelligence, and open-source intelligence in shaping public trust and policy responses. *Oluwaseun Oladeji, Disinformation in the digital era: The role of deepfakes, artificial intelligence, and open-source intelligence in shaping public trust and policy responses (March 04, 2023)*.

van Doorn, M., Duivestein, S., & Pepping, T. (2021). *Real fake: Playing with reality in the age of AI, deepfakes and the metaverse*. Uitgeverij kleine Uil.

Reuters Institute for the Study of Journalism. (2024). Digital News Report 2024. University of Oxford.

Sumsub. (2024). Deepfake Cases Surge in Countries Holding 2024 Elections — Sumsub Research.

World Economic Forum. (2024). Building Trustworthy Media Ecosystems in the Age of AI and Declining Trust