WEP Warwick Evans Publishing

Transactions on Social Science, Education and Humanities Research
ISSN: 2960-1770, eISSN: 2960-2262 | Volume 17
CEPC 2026

Deepfakes and the Liar's Dividend: From Deception to Denial in the Erosion of Public Trust

Xinmeng Li *

School of Arts Languages and Cultures, The University of Manchester, Manchester, M13 9PL, UK

* Corresponding Author Email: xinmeng.li-4@student.manchester.ac.uk.

Abstract. Generative artificial intelligence can now generate fake audio and video at a relatively low cost, and it is very difficult to tell whether it is real; thus, public trust in what they see and hear is starting to erode. Communication and political science research have shown that, in the context of public and political life, the most serious trust-related harm caused by deepfakes is not widespread deception of the audience but rather the "liar's dividend": as the public becomes more aware that any recording may be fake, political and other parties can dismiss genuine evidence as false. Exposure to deepfakes more frequently results in uncertainty rather than a fixed false belief, and this uncertainty erodes trust in the news and in the evidentiary value of recordings. An increasing number of studies in law, communication and political science support the above model, although their extent is still unknown. Three documented cases in the United States are relevant to this issue: an AI-cloned political robocall, the rejection of an authentic campaign photograph as "AI", and a "deepfake defence" presented in court. The liar's dividend is empirically real but limited; it has been more reliably employed against textual evidence than against video evidence, and the main damage has been to public trust in all evidence.

Keywords: Deepfakes; liar's dividend; public trust; misinformation; political communication; synthetic media.

1. Introduction

Most of the photographic era presumed the authenticity of audiovisual records; therefore, videos and sound clips were generally considered strong evidence that something had occurred. Generative AI has disturbed this assumption. Tools that can create realistic faces and voices are now inexpensive, fast and widely available, and the resulting "deepfakes" are spreading on the same platforms as legitimate news. The risk itself is not only a technical problem. Trust in shared evidence is required for democratic discussion, journalism and the operation of all institutions; if the public can no longer be sure that a recording is authentic, then the common factual basis of public life will begin to erode.

The first warnings of deepfakes focused on deception: that is to say, an extremely realistic fake might deceive many people into thinking it is real. According to recent studies, the more severe problems in reverse are emerging. Chesney and Citron [1] put forward the "liar's dividend": since people are now aware that recordings can be forged, they are more likely to doubt the authenticity of genuine content, and dishonest parties can use this doubt to present real evidence as fake. The benefit, according to the authors, increases with the public's awareness of deepfakes and will be felt most acutely by people who are already suspicious of the media and their political adversaries. The two recent US episodes have built up to this. In January 2024, an AI-cloned voice of President Biden told New Hampshire voters to stay home, but it was a forgery. Seven months later, Donald Trump called a genuine photo of a large Harris campaign rally AI-generated and rejected it as fake. Exposure to synthetic political videos is likely to generate uncertainty rather than a lasting false belief, and this uncertainty will consequently erode trust in the news [2]; the denial move takes advantage of exactly this loss of trust. The degree of corrosion is undetermined.

In public and political communication, the main trust-related harm of deepfakes is not the large-scale deception of the audience, but the liar's dividend and the general erosion of trust in evidence that moves with it. It is a real harm that has limits. A large pre-registered study of more than 15,000 adults

has found that dismissing a real scandal as fabrication reliably increases a politician's support when the underlying report is in text form, but is much less effective in video format, and does not cause a general loss of trust in the media [3]. Given that deepfakes are a real problem for evidence at present, and there is no basis for the strong assertion that synthetic media have already caused an "information apocalypse", the current record does not support such a claim.

2. Conceptual Foundations

A deepfake is audio, image or video that has been generated or altered by machine-learning methods to appear genuine. The category generally refers to face-swapped videos, but it also includes voice clones and fabricated audio; these are relatively inexpensive to produce and have been subject to weaker platform moderation in the past compared with videos. Two qualifications build any account of trust. The first is a problem of technology. The most serious manipulations are not always the most complex: Paris and Donovan [4] distinguish between deepfakes and "cheap fakes" – misleading media made using conventional tools such as selective editing, relabeling, or slowing down footage – and show that cheap fakes are still more prevalent and, in practice, at least as damaging because they are relatively easy to produce in large numbers and spread rapidly. Attributing trust erosion solely to artificial intelligence is to overlook the age of this problem and to place the source of harm in a single technological method; rather, social divisions are exploited by all forms of altered media.

The liar's dividend names the second-order damage; it does not require the success of a fake. Chesney and Citron [1] have observed that educating the public about deepfakes may have an adverse effect: a population being taught that anything can be faked will become more sceptical and thus reluctant to believe real recordings of actual behaviour. Since a dividend only needs to be reasonably fabricated, it is available to people who have never made a deepfake at all. It has the strongest effect among those for whom there is already a narrative of distrust towards journalists or political opponents; it is familiar from research on motivated reasoning. Wardle and Derakhshan [5] have placed the harm in the wider information environment and distinguish between misinformation, false content shared without the intention of causing harm, disinformation, false content deployed to cause harm, and malinformation, genuine content circulated to cause harm. The liar's dividend includes the following: the denial is itself disinformation, but the basic materials are genuine evidence that has been reinterpreted to mislead. This hybrid nature is why it cannot be effectively used to check facts; it is an instrument for rectifying false claims rather than for defending the true ones, and thus it shifts the burden of proof onto the reporter of the truth.

3. Theoretical Lens

Communication research considers the mechanism of trust erosion to be more doubt than lack of belief. In a survey experiment using a British sample, Vaccari and Chadwick [2] showed that when exposed to a political deepfake by viewers, most were more uncertain than deceived: roughly one in six were misled, about a third were unsure what to believe, and the rest were not fooled. Crucially, the resulting uncertainty also reduced trust in the news spread via social media, although there were no outright lies. Deception is therefore not an index of harm. A deepfake that fools almost no one can still harm the information environment by raising a general sense of suspicion that any clip might be fake, thus creating a "liar's dividend". The authors themselves point out that widespread uncertainty may allow dishonest people to deny the existence of wrongdoing by claiming there is no proof and foster a general sense of disillusionment.

The reasons for the specific vulnerability of audiovisual evidence in epistemology are more apparent. Rini [6] believes that recordings have served as an "epistemic backstop" for testimony for a long time; there is a standing possibility to verify whether a spoken statement matches a picture or a tape later on, and this discipline what people are willing to speak about in their daily lives and maintains the trust in each other's words. Audiences tend to give more weight to video and audio than to text because recordings have historically been more difficult to fake convincingly; they have an intuition that

seeing is believing. Deepfakes attack this backstop at its root. With a reduction in the cost of producing high-fidelity recordings, the evidentiary weight given to video evidence is diminishing, and the original disparity between written and filmed data may also be lessening. The damage is not specific to a particular case: even a debunked fake will leave a corrosive knowledge that the next recording might be synthetic, and thus trust, once damaged, is difficult to repair by addressing a single falsehood.

4. State of the Evidence: Consensus and Contestation

Several conclusions are now widely accepted. Exposure to deepfakes is more likely to cause uncertainty than lasting deception; the loss of trust in news spread on social media and reduced confidence in the evidence provided by recordings are recurring problems, and the liar-dividend mechanism is real and observable rather than just hypothetical [1, 2]. Schiff, Schiff and Bueno [3] have performed five pre-registered survey experiments in the United States with over 15,000 adults to provide the strongest single-test evidence for this mechanism. Politicians who falsely labelled a real scandal as misinformation gained support across the aisle by doing two things, according to the authors: raising the information uncertainty of the report and rallying their core supporters against unfriendly media. The pooled effect was approximately one-fifth of the standard deviation for text-based scandals, and these false reports were more serious than silence or an apology.

The same study sets the limit of discipline for the general fear. Although the scandal was mentioned in the text, the dividend remained unchanged; however, according to the authors, it failed to reduce the general trust of the public in the media [3]. Video has been harder to deny than text so far, but whether this imbalance will continue depends on how convincing the generated videos become and how mistrustful the audience is of the news organisations already. This qualification matters for framing. The evidence supports a concern about strategic denial but does not license the claim that all evidence is equally contestable.

Evidence from real discourse shows that the mechanism is outside the laboratory. Twomey and others have analyzed more than 1,000 tweets about deepfakes during the 2022 Russian invasion of Ukraine [7] and found that users labelled authentic footage as fake far more frequently than they correctly identified actual deepfakes; furthermore, fear of fabrication damaged trust in genuine conflict imagery to such an extent that some users said they could no longer believe any such footage. The reported damage came less from the falsity of the synthetic media and more from the way people used deepfakes to discredit actual videos, a phenomenon known as the liar's dividend in practice.

The principal dispute concerns magnitude rather than existence. Habgood-Coote [8] suggests that catastrophic "epistemic apocalypse" framing is an exaggeration and historically naive; manipulated images predate machine learning by a long margin, social norms for trust in recordings have been more enduring than alarm suggests, and proposed remedies have relied too heavily on technology while ignoring these norms. The most stringent experiments support this restriction and have reported trust effects that are real but small, and for general media trust, they are absent [2, 3]. Case-level analysis also indicates the same. De Nadal and Jančárik [9] studied the 2023 Slovak audio deepfake released a few days before the national election and cautioned that the result could not be ascribed to synthetic media; rather, deep-seated institutional distrust served as the reason. They also pointed out that the more dangerous risk was the "liar's dividend", when politicians could use a deepfake-aware public to reject all genuine evidence.

A second problem is intervention, and it worsens the contradiction at the heart of the concept. Warning the audience of a deepfake may have the opposite effect; Ternovski, Kalla and Aronow [10] found that in two experiments, informing participants about deepfakes reduced their trust in subsequent authentic political videos without improving their ability to detect fakes. Media-literacy efforts that only raise alarm will thus create the very suspicion the liar's dividend thrives on, as Chesney and Citron anticipated in their perverse feedback, and now show direct experimental support.

5. Case Analysis

Three documented United States cases have traced the development from deception to denial, based on public reports and official documents. The three differ in artifact, channel and verification response: the first is a deepfake used for deception; the second and third are instances of the liar's dividend, where genuine evidence has been labelled as fake.

5.1. An AI-Cloned Robocall

Two days before the January 2024 New Hampshire presidential primary, thousands of voters received a robocall in which an AI-generated clone of President Biden's voice urged them not to vote and told them to "save your vote" for the November election [11]. Steve Kramer, the political consultant, made the call; within a day, it was determined to be a synthetic audio and spread on social platforms. Audio was used as the carrier for the attack; voice clones are inexpensive to generate, lack the visual defects that reveal weak videos, and travel via telephone, a channel that avoids the content-moderation systems of the platforms. The fake was crude in ambition, only a single suppressive instruction, yet serious enough to provoke a federal response.

Thus, the existing laws do not apply uniformly to synthetic fraud. The Federal Communications Commission determined that the AI-generated voice constituted an illegal robocall under the current rules of telecommunications and imposed a six-million-dollar fine on the organizer of the calls, as well as separate enforcement measures against the carrier that sent out these calls [11]. The civil action attached to the form of the violation was for an unauthorised spoofed robocall, not for synthetic impersonation itself. The criminal case collapsed: Kramer was charged with voter-suppression and candidate-impersonation counts but was acquitted by a New Hampshire jury in 2025 of all of them after the defence argued that the primary had been unofficial [12]. Enforcement reached the delivery mechanism but not the deception, and the harm most characteristic of a deepfake - an artificial voice impersonating a candidate to suppress voter turnout - slipped through the cracks of the regime designed for older crimes.

The episode also shows feedback that makes deepfakes harmful beyond a single fake. A well-known imitation shows people that a familiar voice can be created, thus increasing the environment of doubt that aids the liar's dividend. A deepfake that aims to deceive is, in fact, made more harmful by exposing it and thus leading people to ignore genuine recordings as fake; this is the very mechanism Chesney and Citron identified and has been occurring in the news cycle since then [1].

5.2. Dismissing an Authentic Photograph as “AI”

In August 2024, Donald Trump claimed that a picture of a large group of people at a Harris campaign rally near Detroit had been created by artificial intelligence and said that the crowd "did not exist" [13]. It is a real picture. Roughly 15,000 people had attended, multiple independent videos confirmed the scene, and a well-known researcher in deepfake detection found no signs of manipulation [13]. Where the robocall produced a false item, this episode produced nothing; rather, the deception was entirely in denial. It is the liar's dividend in its purest form, and it does not require deepfakes at all; only the now-plausible suggestion of one's existence is needed.

The change is a shift in the burden of proof. A genuine photograph is no longer self-authenticating; instead, journalists and their opponents have to undertake the defensive task of proving that "real" is still real by marshaling attendance figures, corroborating footage and forensic analysis to refute the claim that cost nothing to produce. This is the attack on the evidential safeguard that Rini describes: once recordings no longer have the presumption of authenticity by default, the labour of verification falls on those who use them, and the asymmetry of effort favours the denier [6]. The target is not a single fact, but rather the public’s confidence that the documents can resolve a dispute; this serves as the foundation for holding the campaign accountable.

The case further complicates the experimental results indicating a weak liar's dividend in videos. Schiff and others have found that branding a fabricated scandal reliably helps politicians in text-based

reports but fails to have the same effect against video evidence [3]; in this case, the denied evidence was visual, and despite this, it still found an audience. The tension is resolved through the mechanism Chesney and Citron emphasize. Survey experiments estimate average effects among the general population, whereas a denial like this one is directed at a partisan audience already inclined to believe the narratives of mistrust in the press, and thus motivated reasoning lowers their threshold for rejecting an inconvenient image [1]. The protection that video evidence is supposed to have is precisely at the point of a steep change in polarization, and this case adds to that experimental record.

5.3. The “Deepfake Defense” in Court

The dynamic has entered the courtroom. In Sz Huang v. Tesla, a wrongful-death lawsuit in the Superior Court of California, County of Santa Clara, after an Autopilot crash in 2018 that resulted in the death of Apple engineer Walter Huang, the attorneys for Tesla resisted questioning Elon Musk about his recorded 2016 statements promoting the safety of the system, claiming that such recordings were likely deepfakes and thus could not be authenticated [14]. The claim turned into a political tactic for a venue meant to hear evidence and thus refused to refute the statement by questioning its factual basis.

Judge Evette Pennypacker rejected this argument and ordered the deposition, writing that its logic was "deeply troubling" because it would allow prominent figures to "avoid taking responsibility for what they actually said and did" merely by claiming that any recording of them might be false [14]. The ruling names a systemic danger directly. If fame gave a presumption that one's recorded words were fabricated, then the more public a person was, the less accountable they would be; this inversion would destroy the function of recordings as evidence. A courtroom is the most institutionalised environment for the operation of audiovisual evidence as the epistemic support that Rini mentions, and deepfake defence attacks target this support, the failure of which would be most severe [6].

With the first two cases, it has been shown that there are no exceptions to the rule of the liar's dividend. The criminal check of the robocall failed at trial, and the dismissed-photograph claim received no formal adjudication whatsoever, as political speech is not subject to gatekeeping; on the other hand, the courtroom provided an adversarial process and a judge with the authority to rule against this maneuver. But it's still the same. A litigant loses little by making an appeal; each appeal presses on the presumption that a recording shows what it appears to show, a presumption that the spread of synthetic media steadily weakens. What Chesney and Citron have identified as the legal centre of the concept has now been used in court more often than in academic papers [1].

6. Discussion

The three cases are in fact sequential rather than individual. Audio fakes are relatively inexpensive, lack the visual cues that would indicate weak video, and in this instance, were spread via telephone rather than through moderate social media; thus, they were both easy to produce and serious enough to warrant federal sanctions. However, the more severe the movement from deployment to denial. Once the public believes that any recording can be artificial, then that awareness itself becomes a reason to disregard genuine evidence, such as a campaign photo or a deposition exhibit. The three links in the harm are: a general uncertainty that reduces trust, strategic deniability by which actors can avoid responsibility, and the verification burden that now falls on anyone who needs to prove that the genuine is genuine.

Governance has had little effect. The robocall case shows that the enforcement tools are applied unevenly; a civil penalty has been imposed, but criminal liability has not, indicating that the statutes written for older types of harm do not transfer well to synthetic deception. Detection technology is not highly reliable because the liar's dividend does not require a convincing fake; only a reasonable possibility of such a fake needs to be established, so even excellent detectors cannot counter a denial based on genuine documents. Even good-faith literacy campaigns can worsen the problem if they teach suspicion without also teaching verification [10]. A promising direction is to strengthen

provenance and authentication, that is, to implement technical and institutional measures to verify the authenticity of content, and to construct verification systems within newsrooms, courts and election authorities that foster media literacy and calibrated doubt.

Journalism and political communication now aim to defend the truth rather than exposing fake news. Conventional fact-checking corrects false information; the liar's dividend turns around and targets true information by demanding a swift, credible authentication of contested but genuine evidence and professional standards that raise the social cost of casually dismissing "deepfakes". There is an asymmetry in consequence: a single unanswered denial can inflict greater damage on shared trust than the fake itself would have caused.

7. Limitations and Future Research

The evidence is still restricted in a few ways. The most powerful experiments are based on survey designs and primarily use Western, often British or United States, samples, limiting causal and cross-national generalisation. Long-run claims about a creeping "reality apathy" are still largely theoretical and have not been empirically verified, and the text-versus-video asymmetry that currently limits the dividend may diminish as synthetic video becomes more convincing; this constitutes a falsifiable prediction worth monitoring. A lack of attention has been given to the most common type of harm, as shown in Deeptrace's 2019 map of online deepfakes: 96% of the detected deepfake videos were pornographic and exclusively involved women [15]. The gendered attribute is also one of the electoral and evidence problems discussed here, as intimate-image abuse undermines public trust in the platform and authorities.

8. Conclusion

The most serious threat deepfakes pose to the public's trust is not that many people will be deceived, but that no one will need to be: if any recording can be dismissed as fake, the liar's dividend will allow authentic evidence to be ignored, and shared trust in evidence itself will be destroyed. The mechanism is empirically real; it has been observed in survey experiments, in wartime discourse, and in US politics and courts, but it is limited, weaker against video than against text, and not, based on current evidence, a general loss of media trust. Both findings together reduce the risk of alarmism but do not disregard real dangers; instead, they suggest responses based on provenance, verification and calibrated public skepticism rather than mere detection. As synthetic videos have improved, the protections provided by text-video asymmetry are becoming less effective, and the defence of authentic evidence now needs to be strengthened.

References

[1] Chesney R, Citron D K. Deep Fakes: A Looming Challenge for Privacy, Democracy, and National Security. California Law Review, 2019, 107: 1753-1819. https://doi.org/10.15779/Z38RV0D15J.

[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. https ://doi.org/10.1177/2056305120903408.

[3] Schiff K J, Schiff D S, Bueno N S. The Liar's Dividend: Can Politicians Claim Misinformation to Evade Accountability? American Political Science Review, 2025, 119(1): 71-90. https: // doi.org/10.1017/S0003055423001454.

[4] Paris B, Donovan J. Deepfakes and Cheap Fakes: The Manipulation of Audio and Visual Evidence. New York: Data & Society Research Institute, 2019. https://datasociety.net/library/deepfakes-and-cheap-fakes/.

[5] Wardle C, Derakhshan H. Information Disorder: Toward an Interdisciplinary Framework for Research and Policy Making. Strasbourg: Council of Europe, 2017. https://rm.coe.int/information-disorder-report-version-august-2018/16808c9c77.

[6] Rini R. Deepfakes and the Epistemic Backstop. Philosophers' Imprint, 2020, 20(24): 1-16. https://philpapers.org/rec/RINDAT.

[7] Twomey J, Ching D, Aylett M P, 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. https://doi.org/10.1371/journal.pone.0291668.

[8] Habgood-Coote J. Deepfakes and the Epistemic Apocalypse. Synthese, 2023, 201(3): 103. https://doi.org/10.1007/s11229-023-04097-3.

[9] de Nadal L, Jančárik P. Beyond the Deepfake Hype: AI, Democracy, and "the Slovak Case." Harvard Kennedy School Misinformation Review, 2024. https://doi.org/10.37016/mr-2020-153.

[10] Ternovski J, Kalla J, Aronow P M. The Negative Consequences of Informing Voters about Deepfakes: Evidence from Two Survey Experiments. Journal of Online Trust and Safety, 2022, 1(2). https://doi.org/10.54501/jots.v1i2.28.

[11] Bond S. a Political Consultant Faces Charges and Fines for Biden Deepfake Robocalls. NPR, May 23, 2024. https: //www.npr.org/2024/05/23/nx-s1-4977582/fcc-ai-deepfake-robocall-biden-new-hampshire-political-operative.

[12] WBUR. N.H. Jury Acquits Consultant Behind AI Robocalls Mimicking Biden on All Charges. WBUR, June 16, 2025. https://www.wbur.org/news/2025/06/16/biden-ai-robocall-new-hampshire-steven-kramer-not-guilty.

[13] Doan L, Delzer E. Trump Falsely Claims Harris Campaign Used AI to Fake Crowd in Detroit. CBS News, August 12, 2024. https://www.cbsnews.com/news/trump-harris-campaign-photo-crowd-size-detroit/.

[14] Sz Huang et al. v. Tesla, Inc., No. 19CV346663, Tentative Ruling re Plaintiffs' Motion to Compel Discovery and Elon Musk's Deposition. Superior Court of California, County of Santa Clara, April 27, 2023. https://cdn.arstechnica.net/wp-content/uploads/2023/04/musk-deepfake-ruling.pdf.

[15] Ajder H, Patrini G, Cavalli F, Cullen L. The State of Deepfakes: Landscape, Threats, and Impact. Amsterdam: Deeptrace, 2019. https://regmedia.co.uk/2019/10/08/deepfake_report.pdf.