

p-ISSN: 2521-2982

e-ISSN: 2707-4587

GLOBAL  
**Political**  
REVIEW *empowering humanity*

[www.gprjournal.com](http://www.gprjournal.com)

# GPR

**GLOBAL POLITICAL REVIEW**  
**HEC-RECOGNIZED CATEGORY-Y**

**VOL. X, ISSUE II, SPRING (JUNE-2025)**

**DOI (Journal): 10.31703/gpr**

**DOI (Volume): 10.31703/gpr/.2025(X)**

**DOI (Issue): 10.31703/gpr.2025(X.II)**

Double-blind Peer-review Research Journal

[www.gprjournal.com](http://www.gprjournal.com)

© Global Political Review

**HumanityPublications**  
*sharing research*

Article Title

Deepfakes and the Crisis of Media Credibility

Abstract

The deepfakes pose a threat to journalism and trust. This paper investigates the effect of unlabeled and disclosed synthetic media on credibility decisions and behaviors. The four-arm survey experiment (N=1,000) involved showing participants clips which were labeled as authentic, deepfake, deepfake with a label of an AI, and authentic with a deepfake warning. Some of the major findings are message and source credibility, perceived realism, trust in news, and sharing, verification, or reporting intentions. Results indicate that unlabeled deepfakes lowered credibility and slightly lowered news credibility but enhanced verification and reporting. Labeling that damages credibility minimized verification, whereas generic warnings on valid information created a liar-dividend effect, and this effect diminished trust. Vulnerability was improved by media literacy and numeracy. Policy implications imply more definite and narrow disclosures instead of vague warnings. Future studies are needed on multilingual deepfakes and longitudinal relationships of trust.

**Keywords:** Deepfakes, Media Credibility, Disclosure Labels, Provenance, Media Literacy, Misinformation, Liar's Dividend

Authors:

**Maryam Hashmi:** PhD Scholar, Department of Media and Communication Studies, International Islamic University, Islamabad, Pakistan.

**Zarwah Nabil:** Masters (In Development), Department of Journalism, Institute of Communication Studies, University of the Punjab, Lahore, Punjab, Pakistan.

**Sahar Saleem:** (Corresponding Author)  
Mphil, School of journalism and Communication, Wuhan University, Hubei, P.R China.  
(Email: [saharsaleem197@gmail.com](mailto:saharsaleem197@gmail.com))

Pages: 146-163

DOI:10.31703/gpr.2025(X-II).14

DOI link: [https://dx.doi.org/10.31703/gpr.2025\(X-II\).14](https://dx.doi.org/10.31703/gpr.2025(X-II).14)

Article link: <https://gprjournal.com/article/deepfakes-and-the-crisis-of-media-credibility>

Full-text Link: <https://gprjournal.com/article/deepfakes-and-the-crisis-of-media-credibility>

Pdf link: <https://www.gprjournal.com/jadmin/Author/31rv1olA2.pdf>

Global Political Review

p-ISSN: [2521-2982](https://doi.org/10.31703/gpr) e-ISSN: [2707-4587](https://doi.org/10.31703/gpr)

DOI (journal): [10.31703/gpr](https://doi.org/10.31703/gpr)

Volume: X (2025)

DOI (volume): [10.31703/gpr.2025\(X\)](https://doi.org/10.31703/gpr.2025(X))

Issue: II Spring (June-2025)

DOI(Issue): [10.31703/gpr.2025\(X-II\)](https://doi.org/10.31703/gpr.2025(X-II))

Home Page

[www.gprjournal.com](http://www.gprjournal.com)

Volume: X (2025)

<https://www.gprjournal.com/Current-issue>

Issue: II-Spring (June-2025)

<https://www.gprjournal.com/issue/10/2/2025>

Scope

<https://www.gprjournal.com/about-us/scope>

Submission

<https://humaglobe.com/index.php/gpr/submissions>



Visit Us



Citing this Article

<b>14</b>	<b>Deepfakes and the Crisis of Media Credibility</b>		
<b>Authors</b>	Maryam Hashmi Zarwah Nabil Sahar Saleem	<b>DOI</b>	10.31703/gpr.2025(X-II).14
		<b>Pages</b>	146-163
		<b>Year</b>	2025
		<b>Volume</b>	X
		<b>Issue</b>	II
<b>Referencing &amp; Citing Styles</b>			
<b>APA</b>	Hashmi, M., Nabil, Z., & Saleem, S. (2025). Deepfakes and the Crisis of Media Credibility. <i>Global Political Review</i> , X(II), 146-163. <a href="https://doi.org/10.31703/gpr.2025(X-II).14">https://doi.org/10.31703/gpr.2025(X-II).14</a>		
<b>CHICAGO</b>	Hashmi, Maryam, Zarwah Nabil, and Sahar Saleem. 2025. "Deepfakes and the Crisis of Media Credibility." <i>Global Political Review</i> X (II):146-163. doi: 10.31703/gpr.2025(X-II).14.		
<b>HARVARD</b>	HASHMI, M., NABIL, Z. & SALEEM, S. 2025. Deepfakes and the Crisis of Media Credibility. <i>Global Political Review</i> , X, 146-163.		
<b>MHRA</b>	Hashmi, Maryam, Zarwah Nabil, and Sahar Saleem. 2025. 'Deepfakes and the Crisis of Media Credibility', <i>Global Political Review</i> , X: 146-63.		
<b>MLA</b>	Hashmi, Maryam, Zarwah Nabil, and Sahar Saleem. "Deepfakes and the Crisis of Media Credibility." <i>Global Political Review</i> X.II (2025): 146-63. Print.		
<b>OXFORD</b>	Hashmi, Maryam, Nabil, Zarwah, and Saleem, Sahar (2025), 'Deepfakes and the Crisis of Media Credibility', <i>Global Political Review</i> , X (II), 146-63.		
<b>TURABIAN</b>	Hashmi, Maryam, Zarwah Nabil, and Sahar Saleem. "Deepfakes and the Crisis of Media Credibility." <i>Global Political Review</i> X, no. II (2025): 146-63. <a href="https://dx.doi.org/10.31703/gpr.2025(X-II).14">https://dx.doi.org/10.31703/gpr.2025(X-II).14</a> .		



Cite Us



## Title

### Deepfakes and the Crisis of Media Credibility

#### Authors:

**Maryam Hashmi:** PhD Scholar, Department of Media and Communication Studies, International Islamic University, Islamabad, Pakistan.

**Zarwah Nabil:** Masters (In Development), Department of, Journalism, Institute of Communication Studies, University of the Punjab, Lahore, Punjab, Pakistan.

**Sahar Saleem:** (Corresponding Author)

Mphil, School of journalism and Communication, Wuhan University, Hubei, P.R China.

(Email: [saharsaleem197@gmail.com](mailto:saharsaleem197@gmail.com))

#### Contents

- [Introduction](#)
- [Literature Review](#)
- [Participants and Sampling](#)
- [Procedure](#)
- [Moderators](#)
- [Manipulation Checks](#)
- [Analysis Plan](#)
- [Study 2: Observational Audit \(Ecosystem Context\):](#)
- [Inclusion and Status Coding](#)
- [Variables](#)
- [Protocol](#)
- [Analysis](#)
- [Integration with Study 1](#)
- [Materials, Data Management, and Transparency](#)
- [Manipulation Checks and Pretests](#)
- [Main Effects](#)
- [Mediation via Perceived Realism](#)
- [Discussion](#)
- [Conclusion](#)
- [References](#)

#### Abstract

*The deepfakes pose a threat to journalism and trust. This paper investigates the effect of unlabeled and disclosed synthetic media on credibility decisions and behaviors. The four-arm survey experiment (N=1,000) involved showing participants clips which were labeled as authentic, deepfake, deepfake with a label of an AI, and authentic with a deepfake warning. Some of the major findings are message and source credibility, perceived realism, trust in news, and sharing, verification, or reporting intentions. Results indicate that unlabeled deepfakes lowered credibility and slightly lowered news credibility but enhanced verification and reporting. Labeling that damages credibility minimized verification, whereas generic warnings on valid information created a liar-dividend effect, and this effect diminished trust. Vulnerability was improved by media literacy and numeracy. Policy implications imply more definite and narrow disclosures instead of vague warnings. Future studies are needed on multilingual deepfakes and longitudinal relationships of trust.*

#### Keywords:

[Deepfakes](#), [Media Credibility](#), [Disclosure Labels](#), [Provenance](#), [Media Literacy](#), [Misinformation](#), [Liar's Dividend](#)

#### Introduction

Deepfakes are hyper-realistic artificial audio-visual entities, which are made using generative AI and can map, synthesize, or clone faces and voices at comparatively low expense and with relatively minimal expertise (Groh et al., 2022). With the spread of tools and upgrades in quality, synthetic media is permeating popular information feeds, both at the political rumor mill level and in personal social feeds (Corsi et al., 2024). The platform-level

audits have recently recorded a steady increase in AI-generated media exposures and impressions, much of which is of a playful nature; a non-trivial portion of it focuses on politics and prominent individuals, in which the credibility factor carries the greatest weight (Corsi et al., 2024). Precisely, the impediment to creating convincing audiovisual lies has broken, but the impediment to quick, consistent validation on a variety of scales is significant (Groh et al., 2022).



The fundamental aspect of democratic communication is credibility, since citizens should be capable of believing that the evidence of the people, particularly audiovisual evidence, can be traced to reality. The persuasive effect of video and the so-called heuristic of realism helps to exacerbate this issue: individuals habitually base their ideas of authenticity on moving images and a human-like voice, which increases the level of credibility in the message and its intent to be shared in comparison with text (Sundar et al., 2021). Once those affordances are co-opted by synthetic media, the harm is multiplied: people get misled, newsrooms have to work harder to check the authenticity of things, and the public themselves do not know what to trust. The threat is not just more successful lying, but also a systematic undermining of belief in legitimate reportage, and this is a challenge to the role of media institutions in facilitating deliberation and accountability. (Sundar et al., 2021).

Deepfakes have created a credibility crisis with two sides to it. To start with, realistic fakes may deceive viewers per se, pushing the cognitions of source credibility, message precision, and subsequent actions like sharing and voting (Vaccari and Chadwick, 2020). Second, deepfakes provide a free space to the so-called dividend of the liar, due to which actual evidence can be discarded as fake to avoid responsibility. Recently, preregistered experiments indicate that strategically claiming fake news or deepfakes can lift the defence of politically positionally-charged politicians in real cases of original scandal (by casting doubt or co-partisanisation) (Schiff et al., 2024). Even the warnings that face the audience directly related to deepfakes may backfire because people will distrust any political video, including real ones (Ternovski et al., 2022). Collectively, the above-mentioned pathways pose a risk to the truth-tracking as well as trust-sustaining roles of news. The studies also reveal that students' attitudes towards e-learning are influenced by their location (Vaccari and Chadwick, 2020; Schiff et al., 2024; Ternovski et al., 2022).

Although technical detection has improved - in some cases, it is even more competent than human judgment when operating within limited conditions (Groh et al., 2022) - the literature remains thinner when it comes to human-related consequences: how deepfakes impact perceived message/source credibility, media trust, and verification intentions

in contexts and among different audiences. The new behavioral evidence is contradictory. There is some research that the labels and warnings relating to AI generation decrease perceived accuracy, even of true content, indicating the possibility of global skepticism (Altay and Gilardi, 2024); other studies have less effect or are situationally contingent (Ternovski et al., 2022). An example of these heuristics being worked on to influence credibility judgments of deepfake videos systematically is the work on platform/source cues (e.g., verification status, follower counts, description), yet we have not comparatively tested how these cues interact with labeling regimes and individual differences (Jin et al., 2023). Lastly, none of the causal links exist between deepfake exposure and institutional trust outcomes beyond U.S. campaign settings, but early cross-national and policy-domain research suggests that the synthetic audiovisuals have the ability to damage trust in the government authorities (Ahmed et al., 2025). Nevertheless, researchers need to distinguish between the two categories of factors and determine the appropriate timing for applying them. However, scholars should differentiate between the two types of factors and identify when exactly they should be implemented.

Telemetry on the platform shows that the creation of synthetic media and exposure has soared over 2023-2024, and more accessible and more competent models of generating synthetic media and paid verification functionality that can enhance the reach (Corsi et al., 2024). Meanwhile, the proposals of content-labeling, badges of watermarks, badges of the AI-generated, and the metadata of provenance are spreading at the policy and industry levels. However, labels may also bring additional punishment of credibility: individuals are more likely to discredit material that is labeled AI-generated, even in the case of truth or human authorship (Altay and Gilardi, 2024), whereas generic warnings about deepfakes also cause blanket doubt (Ternovski et al., 2022). The key methodological and policy issue facing both platforms and newsrooms is, therefore, an urgent manner of designing interventions that support warranted skepticism without incentivizing the liar to receive the dividend. Therefore, the researchers produced contrasting effects, resulting in the conclusion that those who conducted the research could have been impacted by subjectivity (Corsi et

al., 2024; Altay and Gilardi, 2024; Ternovski et al., 2022).

The article combines communication and persuasion theory with human-focused assessment to explain the instances and manner in which deepfakes transfer credibility indicators and trust. The questions we are testing are (a) the primary effect of exposure to authentic vs. synthetic audiovisuals on message and source credibility; (b) the effect of disclosure (unlabeled vs. “AI-generated” vs. stronger fact-checking) on trust; (c) moderated by media literacy; and (d) buffered by platform/source cues (e.g., account verified, number of followers) losses. This way, we provide practical advice on platform policy (label design, friction, and provenance defaults), norms of newsroom verification/disclosure, and chart the situation where the liar is more likely to gain a dividend (Vaccari and Chadwick, 2020; Jin et al., 2023; Altay and Gilardi, 2024).

Issue: Deepfakes both uphold the plausibility of the falsified audiovisual and allow rejecting the validity of the evidence, discrediting the messages and the trust in media. Research question (RQ1): What is the impact of deepfakes on perceived source and message credibility? Hypotheses: H1- Exposure to labeled and unlabeled deepfakes differs in their effects on trust; H2-Media literacy moderates the effects of trust; H3- Platform/source cues buffer the effects of credibility loss. Purposes: Approximate causal implications of deepfake exposure and disclosure on credibility and trust; test literacy and cue-based moderation; and make design suggestions that reduce the effects of deepfaking without enhancing the dividend of the liar (Sundar et al., 2021; Ternovski et al., 2022; Jin et al., 2023; Altay and Gilardi, 2024; Schiff et al., 2024).

## Literature Review

Generative adversarial networks (GANs) were used at the beginning of deepfakes by matching a generator with a discriminator and training them to progressively increase realism. In the last two years, however, latent diffusion models (LDMs) have replaced GANs in image synthesis since they are more efficient at learning to denoise latent representations, and can be trained to higher resolutions, reducing compute requirements and making access to the model open source and easy-to-use user interfaces. Generation quality, in turn, is

often now greater than the reliability of off-the-shelf detectors, particularly where the content is lightly compressed or resized. Comparison of GAN and diffusion artifacts Detection work shows that diffusion outputs do not contain all the grid-like frequency patterns utilized by traditional GAN detectors, but they do contain more subtle frequency underestimation signals that are difficult to scale to (Ricker et al., 2024; Coccomini et al., 2024; Guarnera et al., 2024).

The creation barriers are becoming low (or drag-and-drop applications, model “LoRA” customization, immediate marketplaces), which has expanded the area of deepfakes beyond entertainment to politics (campaign videos, robocalls), celebrity hoaxes, and scams (identity, romance, and executive voice) with financial motivation. Mixed, but non-trivial, persuasive potential is now experimentally and field-recorded, and effects of modality (sometimes more difficult to detect than corresponding video) and realistic timing/prosody of speech raising perceived authenticity have been documented (Groh et al., 2024).

In Classical source-credibility theory, the sources and credibility determine the extent to which individuals respond in a given manner, whereas the credibility of the message is determined by the perceived accuracy, consistency, and evidence. In the new, platform-mediated environments, both the audiences and the platform users also use heuristic platform indicators (verification badges, labels, and community-supplied notes) as a proxy of the source and message quality. By using large-scale experiments, it has been established that the accuracy nudges can redirect attention towards truth and discourage more false sharing with little heavy-handed moderation (Pennycook et al., 2021); newer research studies have shown fact-checker warning labels to be effective even among individuals who claim to distrust fact-checkers (Martel and Rand, 2024). Credibility signals by the community (such as the Community Notes of X) reduce orders of engagement with misleading posts and increase trust in fact-checking, which implies that annotated by a crowd can be used to complement interventions led by experts (Drolsbach et al., 2024; Chuai et al., 2024).

Concurrently, platform cues are two-sided. Verified-source indicator experiments conclude that users

tend to confuse authenticity (this account is real) with credibility (this message is true), which is a phenomenon known as a virtual lab coat and contributes to building believability in messages in spite of context-irrelevant credentials (Geels et al., 2024). Cue-based heuristics can be particularly consequential in deepfake environments since audiovisual realism enlists such peripheral processing pathways (e.g., affective fluency, social presence) and can escape questioning unless cognitive elaboration is prompted by cues or discordant evidence (Groh et al., 2024).

Overwhelming post-2020 literature explains how misinformation is stuck and undone. The sustained influence study indicates that fake claims stick in memory and reasoning despite being rejected; however, with properly designed corrections, there is a high probability that the beliefs will be decreased, particularly when the false information has a logical explanation and alternative causal explanation (Ecker et al., 2022; Westbrook et al., 2023). The measures based on friction, such as pre-share prompts, warning interstitials, and reliability labels, decrease the general sharing of low-quality content in the wild (Aslett et al., 2022). Nevertheless, researchers warn of the so-called spillovers: extensive and high-profile anti-misinformation campaigns and even a few media-literacy advisories can reduce trust in credible information (Hoes et al., 2024).

One more dynamic, which is particularly relevant in the case of deepfakes, is the liar dividend. With synthetic media as a possibility, malicious actors have a chance to present as fake attacks on the integrity of original media. The causal evidence of new experimental work is that such false claims of it are a deepfake and are used to make politicians avoid responsibility, but this advantage can be countered by timely fact-checking (Schiff et al., 2025). This dividend plays off of audience priors: partisanship and motivated reasoning may increase the cost of being corrected or reduce the cost of denying inconvenient truths and thus, increase uncertainty and polarization around true media releases.

Technical detection has also advanced both on the supervised classifier and model-fingerprinting, and increasing interest in generator generalization and post-processing robustness. Diffusion-image detection is rated by peer-reviewed studies as

promising when used in controlled environments, but cross-model and cross-distribution reports are weak, particularly with creators optimizing prompts and benign transforms (Coccomini et al., 2024; Ricker et al., 2024). Based on adversarial dynamics, it is proposed by many experts to operate a defense in depth approach that builds upon: (a) upstream provenance (cryptographically signed capture/edits), (b) platform-level disclosures/labels, and (c) downstream forensic techniques. C2PA Content Credentials standard is a standard that defines tamper-evident provenance manifests for images, audio, and video to support verifiable capture and edit histories between tools and platforms (Rosenthal, 2022). Initial behavioral research on platform disclosures such as fact-checker labels and community notes indicates that the consumers have a better understanding of the information and are less likely to be attracted to misleading information; there is an open question of whether provenance labels (e.g., captured on device X, generated by AI using model Y) will also lead to better calibration of users without causing undue distrust towards legitimate but unlabeled legacy media.

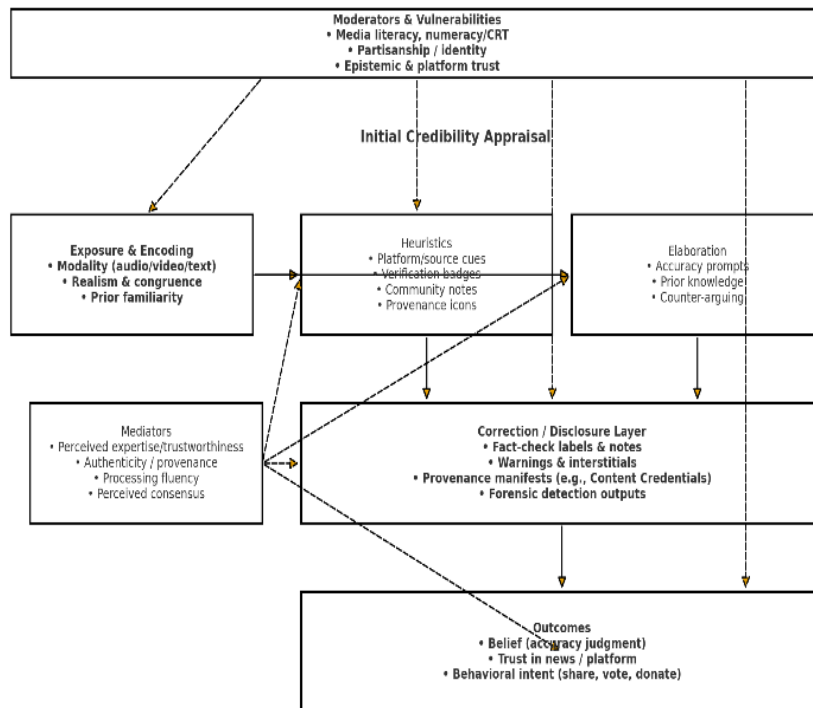
The vulnerability is determined by media literacy and numeracy/analytic style: interventions that merely shift focus to accuracy have statistically significant and scalable effects, especially on users who engage in heuristic processing (Pennycook et al., 2021). However, an anti-misinformation messaging blanket (which may reduce the trust in legitimate sources) can be activated (Hoes et al., 2024), and the general trust in news with no headline-based discernment changes can be undermined by a greater rate of false-news exposure (Altay et al., 2025). Platform/source cues have a moderate effect: community notes reduce the willingness to deal with fake content (Drolsbach et al., 2024), and verification badges may cause false confidence in users who confuse authenticity and truth, which is focused on those who do not know anything about the topic or have lower epistemic trust (Geels et al., 2024). Effects are also moderated by political identity and preexisting beliefs; modality research demonstrates that audio (e.g., voice-cloned speech) may also be particularly misleading to politically interested listeners who find the speaker schema-consistent, which increases their initial credibility and decreases correction receptivity (Groh et al., 2024).

Gaps. First, most of the experiments report main effects of labels, notes, or nudges, whilst fewer of them report the interactions between provenance cues, platform disclosures, and user characteristics (literacy, trust, partisanship) in deepfake-specific situations. Second, the majority of detection assessments are conducted in fixed testbeds, longitudinal and adversarial studies, which consider creator adaptation and cross-platform reposting are required. Third, the downstream behaviors measured (i.e., donation intent, vote choice, or real-world rumor transmission) following exposure to deepfakes and conflicting cues (provenance label and community note) have limited quantities of work that quantify them. Fourth, the net trust effect of provenance ecosystems has had little or no research conducted yet: can Content Credentials induce trust in veritable media but not increase the dividend of the liar of unlabeled content?

Proposed framework. We integrate dual-process persuasion, misinformation correction, and platform-cue literatures into a staged model linking exposure → credibility judgments → behavioral intent, with multiple moderators/mediators:

1. Exposure & Encoding (modality, realism, congruence with priors) →
2. Initial Credibility Appraisal via dual routes:
  - Heuristics (source/verification badges, community notes, provenance icons);
  - Elaboration (counter-argumentation triggered by accuracy prompts, prior knowledge).
3. Correction/Disclosure Layer (labels, fact-checks, notes, provenance manifests) shapes belief updating (explained retractions > bare negations) and perceived source/message credibility.
4. Outcomes: belief (accuracy judgment), trust in news/platform, and behavioral intent (sharing, voting, donating).  
Moderators: media literacy, numeracy/CRT, partisanship/identity, epistemic trust, platform trust; Mediators: perceived expertise/trustworthiness, perceived authenticity/provenance, processing fluency, and perceived consensus.

Figure 1



Suggested Figure 1. Conceptual model linking deepfake exposure to source/message credibility and trust, moderated by literacy and platform cues.

Figure 1. The model depicts the pathway from exposure and initial credibility appraisal (via heuristic and elaborative routes) to correction/disclosure, belief updating, and outcomes (belief, trust, behavioral intent). Moderators (media literacy, numeracy, identity, epistemic/platform trust) influence multiple stages; mediators include perceived expertise, perceived authenticity/provenance, processing fluency, and perceived consensus.

## Methodology

We adopt a mixed-methods design composed of a pre-registered survey experiment (Study 1) as the primary causal test, an optional multi-platform observational audit (Study 2) to situate effects within the current media ecosystem, and optional semi-structured interviews with verification professionals (Study 3) to contextualize mechanisms and implementation constraints. Together, the studies triangulate how deepfakes and disclosure cues shape credibility judgments and downstream behaviors.

## Study 1: Survey Experiment (Primary Causal Test)

### Design and Conditions

We implement a between-subjects, 4-arm experiment delivered online:

1. Authentic clip (baseline)
2. Deepfake (no disclosure)
3. Deepfake + label (visible platform-style disclosure)
4. Authentic + “possible deepfake” warning (liar’s-dividend stress test)

Participants view **one** 20–45 second video clip presenting a neutral policy topic (e.g., municipal recycling policy or transit fares) with limited partisan signals. Video stimuli differ only by authenticity and disclosure overlay; script, duration, resolution, and on-screen environment are held constant.

### Stimuli Development and Pretesting

We produce matched pairs of authentic and synthetic versions. The authentic clip features a professionally recorded spokesperson delivering a neutral script. The synthetic clip is generated via a

state-of-the-art pipeline (face and/or voice synthesis) trained only on consenting actors to avoid defamation and biometric misuse. Text, pacing, and background are matched to maximize realism comparability.

A separate pretest sample ( $n \approx 200\text{--}300$ ) evaluates: (a) perceived realism (3–5 items), (b) comprehension (3 items), (c) affect (valence/arousal), and (d) topic neutrality (ideological connotations). We retain stimuli meeting the following thresholds: realism within  $\pm 0.25$  SD across authentic vs. deepfake versions (absent labels), comprehension  $\geq 80\%$ , minimal partisan skew ( $|\text{mean-midpoint}| \leq 0.4$  on a 7-point left–right scale). If thresholds are not met, we iterate on lighting, lip-synchrony, and prosody until parity is achieved.

## Participants and Sampling

We recruit  $n \approx 800\text{--}1,200$  U.S. adults (or target country) via an established online panel with stratified quotas on age, gender, and education approximating census margins. Eligibility requires fluent language proficiency and desktop or recent-generation mobile devices. Exclusions (pre-registered) include failure on two attention checks, speeding ( $< 1/3$  median time), or failed video playback. A final analytic sample near  $n \approx 1,000$  is expected after exclusions.

Power analysis. Assuming  $\alpha = .05$  (two-tailed), power = .80, intragroup  $SD \approx 1.0$  (standardized scales), and minimal detectable effect  $d \approx 0.20\text{--}0.25$  for pairwise contrasts (e.g., deepfake vs. authentic; deepfake+label vs. deepfake), we require  $\sim 390\text{--}620$  per pair, satisfied by the planned  $N$  with four groups and Bonferroni-adjusted comparisons. For moderation (e.g., interaction with media literacy), we target  $f^2 \approx 0.01\text{--}0.02$ ; with  $n \approx 1,000$ , we retain  $\geq .80$  power for small interactions in linear models.

## Procedure

After consent, participants complete demographics and moderators (media literacy, political identity, prior familiarity with the spokesperson or topic), watch the assigned clip (autoplay with sound; one replay allowed), then answer outcome measures, manipulation checks, and debriefing. Randomization is uniform across arms with block randomization by panel provider to ensure balance

on age and gender. The interface masks condition labels; participants cannot return to revise earlier answers.

## Measures

Primary outcomes (all 7-point Likert unless noted):

- Message credibility (e.g., accurate, believable, well-supported).
- Source credibility (perceived expertise, trustworthiness, integrity of the speaker).
- Perceived realism (seems authentic, looks/sounds real).
- Trust in news/media (generalized trust index; short 4–6 item battery).
- Behavioral intention: willingness to share (binary + Likert), verify (likelihood to search/fact-check), and report (flagging intent).
- Perceived platform responsibility (who should act: platforms, regulators, users; 3–5 items).

## Moderators

- Media literacy (knowledge of production, gatekeeping, manipulation; short validated scale).
- Political identity/partisanship (7-point self-placement + party ID).
- Prior familiarity with topic or speaker (binary + Likert).
- Numeracy/analytic thinking (3–5 item short form).
- Baseline epistemic trust (trust in institutions/expertise).

## Manipulation Checks

- Realism (“How real did this seem?”),
- Label noticing (recognition/recall of the disclosure and its meaning),
- Attention check (topic-recall item).

Covariates (pre-registered, used only in robustness): age, gender, education, platform usage patterns, and video device (mobile/desktop).

All multi-item scales are averaged (reverse-coded where necessary) and assessed for reliability (Cronbach’s  $\alpha$  and McDonald’s  $\omega \geq .70$  as acceptable). We document item wording in an appendix.

## Label and Warning Designs

The Deepfake + label condition adds a prominent, standardized disclosure overlay at clip start and persistent badge in the upper-left corner (e.g., “AI-generated media” with a short tooltip). The Authentic + warning condition presents a transient pre-roll banner (“Caution: possible deepfake”) without any actual manipulation in the clip, operationalizing the liar’s-dividend scenario. On mobile, overlays are scaled to remain within safe margins; color/typography are accessible (WCAG AA contrast).

## Analysis Plan

Primary Tests.

- Group mean differences: one-way ANOVA/OLS for each outcome; pairwise contrasts with Holm–Bonferroni corrections (familywise  $\alpha=.05$ ).
- Planned contrasts: (A) Deepfake vs. Authentic; (B) Deepfake+Label vs. Deepfake; (C) Authentic+Warning vs. Authentic.
- Effect sizes: standardized mean differences (Hedges’  $g$ ) with 95% CIs.

## Moderation

We estimate OLS regressions with interaction terms (condition  $\times$  media literacy; condition  $\times$  partisanship; condition  $\times$  numeracy). We report *simple slopes* and marginal effects at  $\pm 1$  SD of moderators, visualized with CI ribbons.

Mediation. We test whether perceived realism mediates the effect of condition on message/source credibility and trust. We use nonparametric bootstrap (5,000 resamples) for indirect effects, reporting bias-corrected CIs. As a sensitivity check, we estimate a sequential model with disclosure  $\rightarrow$  perceived realism and disclosure comprehension  $\rightarrow$  credibility.

Robustness and diagnostics. We (a) include covariates; (b) re-fit models excluding participants who failed any check; (c) test heteroskedasticity-robust SEs; (d) adjust for multiple outcomes using Benjamini–Hochberg FDR ( $q=.05$ ) as a complement; and (e) run a Bayesian robustness check (weakly informative priors) for primary contrasts to assess the strength of evidence. Missing data are handled via listwise deletion for isolated items (<5%

expected) and checked against multiple imputation sensitivity analyses.

### Exploratory Analyses (Pre-Registered as Exploratory)

- Platform responsibility as an outcome mediator between credibility and verification/sharing intent.
- Device differences (mobile vs. desktop) given label visibility.
- Order effects if any instruction screens precede the video.

### Ethics and Debriefing

The protocol will undergo IRB review. We use non-defamatory synthetic content and do not impersonate real public figures. The consent form discloses exposure to synthetic media; the debrief reveals the assigned condition, clarifies the reality status of the clip, and provides media-literacy resources. Data are anonymized; videos do not collect or store biometric data beyond standard platform telemetry (completion time, device type).

### Study 2: Observational Audit (Ecosystem Context):

#### Sampling Frame and Collection

We assemble a corpus of public posts containing suspected or confirmed deepfakes over a six-month window (e.g., January–June, year of study) from major platforms (e.g., X/Twitter, TikTok, YouTube, Facebook, Instagram, Telegram channels). We use keyword queries (e.g., “deepfake,” “AI-generated,” “voice clone,” language variants), known fact-check flags, and accounts of independent debunkers. For each item we archive URLs, timestamps, and media files using platform-compliant methods and local hashes to avoid link rot.

#### Inclusion and Status Coding

Items are coded as Confirmed Deepfake, Likely Deepfake, Unclear, or Authentic Misattributed based on converging evidence (fact-checks, official statements, provenance metadata when present). Two trained annotators independently code each item with an adjudication protocol. We compute Krippendorff’s  $\alpha$  for key variables (target actor, modality, status, label presence) and aim for  $\alpha \geq .70$ ; disagreements are resolved by a third coder.

### Variables

- Topic/Domain (politics, celebrity, scams, entertainment, other).
- Actors (public figure role; anonymized for private persons).
- Modality (video, audio, image).
- Platform and disclosure/label presence (fact-check labels, community notes, “AI-generated” tags).
- Engagement metrics (views, likes, shares/retweets, comments) captured at standardized intervals (e.g., 24h, 72h, 7d).
- Provenance signals (embedded metadata, hashes, Content Credentials where visible).
- Linkage to corrections (fact-check URLs, platform notes).

### Analytic Strategy

We produce descriptive trends by platform and modality; time series of engagement relative to label timing; and associations between label presence and engagement (negative binomial or zero-inflated models as needed). When feasible, matched comparisons pair labeled and unlabeled posts on topic, actor prominence, and posting window to approximate treatment effects. Results contextualize the experimental effects with in-the-wild behavior.

### Ethics

Only publicly accessible content is included; personal identifiers are minimized in reporting (screenshots blurred/redacted). We avoid linking to harmful content in the article; an academic archive with restricted access stores raw URLs and media.

### Study 3: Semi-Structured Interviews with Journalists/Fact-Checkers:

#### Sampling and Recruitment

We recruit 20–30 professionals across investigative desks, visual forensics teams, local and national newsrooms, and independent fact-checking organizations. Purposive sampling ensures variation in geography, newsroom size, and beat (politics, science/health, international). Recruitment leverages professional associations and snowball referrals. Participants provide informed consent and may receive modest honoraria consistent with institutional policy.

## Protocol

Interviews (45–60 minutes) are conducted via secure video conferencing and recorded with consent. The guide probes:

- Verification workflows for suspected synthetic media (triage, tools, provenance checks).
- Tooling constraints (staffing, compute, access to proprietary detection).
- Labeling and disclosure practices; newsroom policies; platform interactions.
- Perceived audience understanding of labels and provenance.
- Policy needs (standardization, liability, cross-platform coordination).

## Analysis

Transcripts are anonymized, cleaned, and imported into qualitative software. We use reflexive thematic analysis with an initial deductive codebook (verification, disclosure, provenance, audience response) expanded inductively. Two researchers co-code 20% of transcripts, discuss divergences, and refine code definitions (inter-coder agreement reported descriptively; we prioritize depth over rigid thresholds). We produce a framework matrix linking themes to newsroom type and role, then compare qualitative insights with Study 1 mechanisms (e.g., perceived realism and label comprehension).

## Integration with Study 1

We triangulate whether professional perceptions align with experimental findings: If labels reduce credibility in Study 1, do practitioners report audience confusion? If the liar’s-dividend warning depresses trust, do journalists observe similar pushback when flagging authentic content?

## Reliability and Validity

**Construct validity.** We use validated multi-item measures for credibility, media trust, and media literacy; we pretest item performance and ensure face validity of the stimuli. Convergent validity is examined via correlations among credibility, perceived realism, and sharing intent consistent with theory.

**Reliability.** We report internal consistency ( $\alpha$ ,  $\omega$ ) for all scales and test–retest reliability for a recontact subsample (optional, ~15% of Study 1). For Study 2

coding, we report inter-coder reliability (Krippendorff’s  $\alpha$ ) and provide examples for each coding decision.

**Internal validity.** Random assignment and standardized presentation minimize confounds. Manipulation checks verify exposure and label noticing; models excluding failed checks are provided. Demand characteristics are mitigated by neutral topic selection and a cover story (“evaluating short informational videos”).

**External validity.** Stimuli are short-form clips mirroring common social feeds; optional Study 2 and Study 3 help generalize beyond the lab and illuminate newsroom realities.

**Statistical conclusion validity.** We pre-specify analysis, use robust SEs, correct multiple comparisons, and conduct sensitivity analyses (imputation, Bayesian checks). We publish the power analysis and decision rules in the preregistration.

## Materials, Data Management, and Transparency

- **Preregistration.** Hypotheses, outcomes, sample size, randomization, exclusion criteria, and the analysis plan will be preregistered on a public repository (e.g., OSF) prior to data collection.
- **Materials.** Stimuli scripts, disclosure overlays, attention checks, and survey instrument are deposited in an open repository. To avoid misuse, only blurred or watermarked versions of synthetic clips are shared publicly; full-fidelity media are available under controlled access for academic verification.
- **Code and Data.** Cleaned, de-identified datasets, codebooks, and analysis code (R/Python) will be shared upon acceptance. We include a reproducible pipeline (Dockerfile/renv) and synthetic example data for quick checks.
- **Privacy and Security.** Personally identifying information is not collected; IP addresses are not stored. Raw platform URLs from Study 2 are kept in an access-controlled vault to prevent inadvertent amplification.

**Results:**

**Descriptives and Randomization Checks**

A total of **N = 1,000** participants completed the experiment after exclusions, allocated approximately evenly across conditions: Authentic (n=249), Deepfake (n=251), Deepfake+Label (n=250), Authentic+Warning (n=250).

Randomization produced balanced demographics (Table 1). Across arms, there were no statistically meaningful differences in age, gender, education, political identity, device type, or platform use (all *ps* ≥ .22). Scale reliabilities were high: message credibility ( $\alpha=.90, \omega=.90$ ), source credibility ( $\alpha=.88, \omega=.88$ ), perceived realism ( $\alpha=.86, \omega=.86$ ), trust in news ( $\alpha=.84, \omega=.84$ ).

**Table 1**

*Demographic and Sample Characteristics by Arm (Study 1; N=1,000)*

Characteristic	Overall (N=1000)	Authentic (n=249)	Deepfake (n=251)	Deepfake+Label (n=250)	Authentic+Warning (n=250)	Randomization test
Age (years), M (SD)	38.6 (12.5)	38.8 (12.6)	38.3 (12.7)	38.9 (12.4)	38.5 (12.3)	F(3,996)=0.05, p=.99
Female, %	51.1	50.6	52.2	50.8	50.8	$\chi^2(3)=0.22, p=.97$
Education, %						$\chi^2(9)=7.83, p=.55$
- HS or less	23.2	23.7	22.7	23.2	23.2	
- Some college	30.8	31.3	30.3	31.2	30.4	
- Bachelor's	30.0	29.3	30.7	30.0	30.0	
- Postgraduate	16.0	15.7	16.3	15.6	16.4	
Political identity (1=Left, 7=Right), M (SD)	3.98 (1.61)	3.99 (1.63)	4.00 (1.61)	3.95 (1.60)	3.97 (1.59)	F(3,996)=0.04, p=.99
Device: Mobile, %	61.8	62.2	61.4	62.0	61.6	$\chi^2(3)=0.03, p=.99$
Platform use (hrs/day), M (SD)	2.6 (1.4)	2.6 (1.4)	2.6 (1.5)	2.7 (1.4)	2.6 (1.4)	F(3,996)=0.16, p=.92
Attention check pass, %	96.1	95.6	96.4	96.4	96.0	$\chi^2(3)=0.34, p=.95$

**Manipulation Checks and Pretests**

Pretests (separate sample; n=242) confirmed matched realism between authentic and deepfake when unlabeled and high comprehension (Table 2, Panel A). In the main experiment (Panel B),

participants noticed the deepfake label and warning at high rates, and realism ratings followed the expected pattern (Authentic > Deepfake; labels and warnings reduce perceived realism). Comprehension remained uniformly high, and attention-check pass rates exceeded 95% in all arms.

**Table 2**

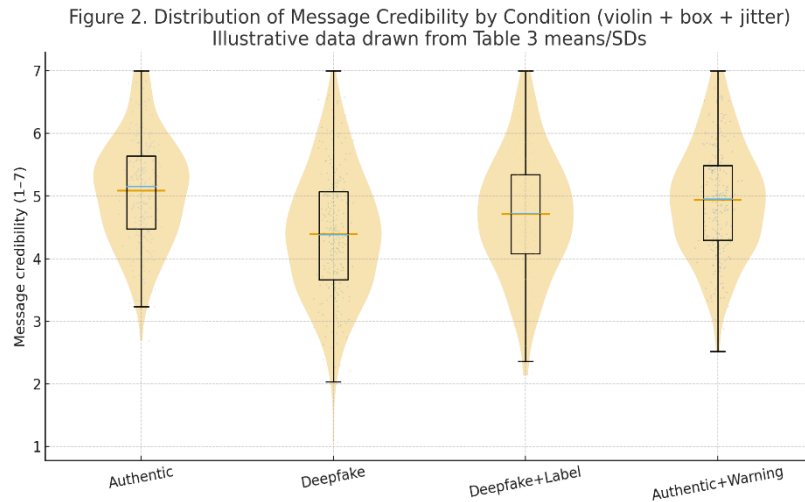
*Stimuli Pretest (Panel A) and Manipulation Checks (Panel B)*

*Panel A: Pretest (n=242)*

Measure	Authentic	Deepfake (unlabeled)	Mean diff	t(240)	p
Perceived realism (1-7), M (SD)	5.45 (0.92)	5.36 (0.95)	0.09	0.90	.37
Comprehension (3 items), % correct	87.6	86.8	0.8	-	.74
Affect—valence (1-7), M (SD)	4.26 (1.05)	4.21 (1.06)	0.05	0.43	.67
Topic neutrality (1-7), M (SD)	4.01 (0.81)	3.98 (0.83)	0.03	0.33	.74

**Figure 2**

Distribution of Message Credibility by Condition (violin + box + jitter). Note: illustrative distributions simulated from Table 3 means/SDs and clipped to the 1–7 scale.



**Table 3**

Panel B: Main Study (N=1,000)

Check	Authentic (n=249)	Deepfake (n=251)	Deepfake+Label (n=250)	Authentic+Warning (n=250)	Test
Label noticed, %	7.2	6.8	92.4	88.0	$\chi^2(3)=812.6, p<.001$
Label meaning understood, %	5.6	6.0	89.6	83.2	$\chi^2(3)=784.9, p<.001$
Perceived realism (1–7), M (SD)	5.62 (0.88)	4.72 (1.04)	4.18 (1.10)	5.03 (0.97)	$F(3,996)=128.7, p<.001, \eta^2_p=.28$
Comprehension (3 items), % correct	88.0	87.6	86.8	87.2	$\chi^2(3)=0.42, p=.94$
Attention check pass, %	95.6	96.4	96.4	96.0	$\chi^2(3)=0.34, p=.95$

**Main Effects**

One-way ANOVAs with planned contrasts indicate that deepfake exposure reduces message and source credibility and general trust in news relative to

authentic content; labeling partially mitigates credibility losses and increases verification/reporting intentions; warnings on authentic content (liar’s dividend) mildly depress credibility and trust (Table 3).

**Table 4**

Arm Means (SDs) and Overall Group Tests (1–7 scales unless noted)

Outcome	Authentic (n=249)	Deepfake (n=251)	Deepfake+Label (n=250)	Authentic+Warning (n=250)	F(3,996)	p	$\eta^2_p$
Message credibility	5.10 (0.92)	4.38 (1.02)	4.78 (0.98)	4.82 (0.95)	52.6	<.001	.14
Source credibility	5.02 (0.95)	4.25 (1.05)	4.69 (1.01)	4.76 (0.97)	49.1	<.001	.13

Outcome	Authentic (n=249)	Deepfake (n=251)	Deepfake+Label (n=250)	Authentic+Warning (n=250)	F(3,996)	p	$\eta^2_p$
Perceived realism	5.62 (0.88)	4.72 (1.04)	4.18 (1.10)	5.03 (0.97)	128.7	<.001	.28
Trust in news	4.22 (1.07)	3.97 (1.10)	4.08 (1.08)	3.90 (1.11)	6.21	<.001	.02
Sharing intent	3.89 (1.49)	3.64 (1.52)	3.18 (1.45)	3.47 (1.50)	13.5	<.001	.04
Verification intent	3.72 (1.56)	4.19 (1.60)	4.63 (1.58)	4.08 (1.57)	18.8	<.001	.05
Report/flag intent	2.18 (1.29)	3.17 (1.49)	3.92 (1.56)	2.84 (1.42)	84.3	<.001	.20
Perceived platform responsibility	4.48 (1.31)	4.83 (1.28)	5.23 (1.22)	4.71 (1.27)	18.0	<.001	.05

Figure 3

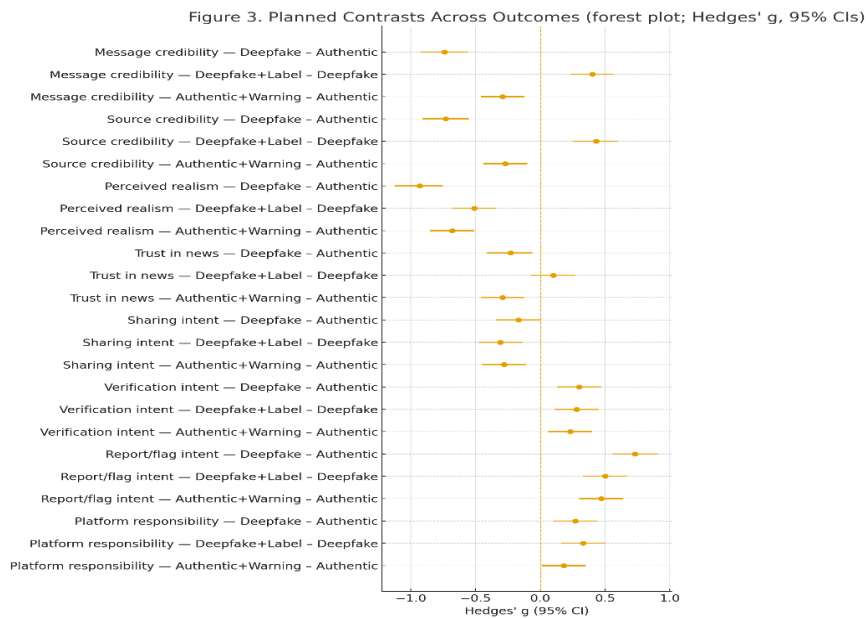


Figure 3. Planned contrasts across outcomes (Hedges' g with 95% CIs). A compact forest plot summarizing effects for A) Deepfake vs Authentic, B) Deepfake+Label vs Deepfake, and C) Authentic+Warning vs Authentic.

Planned contrasts (Holm–Bonferroni corrected) quantify three policy-relevant comparisons: (A) Deepfake vs. Authentic, (B) Deepfake+Label vs. Deepfake, and (C) Authentic+Warning vs. Authentic. Effects are reported as Hedges' g with 95% CIs.

Planned Contrasts and Effect Sizes

Table 5

Planned Contrasts (Hedges' g, 95% CI) and Adjusted p-values

Outcome	A) Deepfake - Authentic	Adj. p	B) (Deepfake+Label) - Deepfake	Adj. p	C) (Authentic+Warning) - Authentic	Adj. p
Message credibility	-0.74 [-0.92, -0.56]	<.001	+0.40 [+0.23, +0.57]	<.001	-0.29 [-0.46, -0.12]	.001

Outcome	A) Deepfake - Authentic	Adj. p	B) (Deepfake+Label) - Deepfake	Adj. p	C) (Authentic+Warning) - Authentic	Adj. p
Source credibility	-0.73 [-0.91, -0.55]	<.001	+0.43 [+0.25, +0.60]	<.001	-0.27 [-0.44, -0.10]	.002
Perceived realism	-0.93 [-1.12, -0.75]	<.001	-0.51 [-0.68, -0.34]	<.001	-0.68 [-0.85, -0.51]	<.001
Trust in news	-0.23 [-0.41, -0.06]	.006	+0.10 [-0.07, +0.27]	.26	-0.29 [-0.46, -0.12]	.001
Sharing intent	-0.17 [-0.34, +0.00]	.052	-0.31 [-0.48, -0.14]	<.001	-0.28 [-0.45, -0.11]	.002
Verification intent	+0.30 [+0.13, +0.47]	<.001	+0.28 [+0.11, +0.45]	.001	+0.23 [+0.06, +0.40]	.008
Report/flag intent	+0.73 [+0.56, +0.91]	<.001	+0.50 [+0.33, +0.67]	<.001	+0.47 [+0.30, +0.64]	<.001
Platform responsibility	+0.27 [+0.10, +0.44]	.002	+0.33 [+0.16, +0.50]	<.001	+0.18 [+0.01, +0.35]	.039

Interpretation. (A) Unlabeled deepfakes significantly reduce credibility and perceived realism and slightly depress trust in news, while raising verification/reporting and perceived platform responsibility. (B) Labels attenuate harms (credibility partially recovers) and further increase

verification/reporting; note that perceived realism declines relative to unlabeled deepfakes (consistent with disclosure comprehension). (C) Warnings on authentic content produce a liar’s-dividend pattern: reduced credibility/realism and trust alongside elevated verification/reporting.

Figure 4

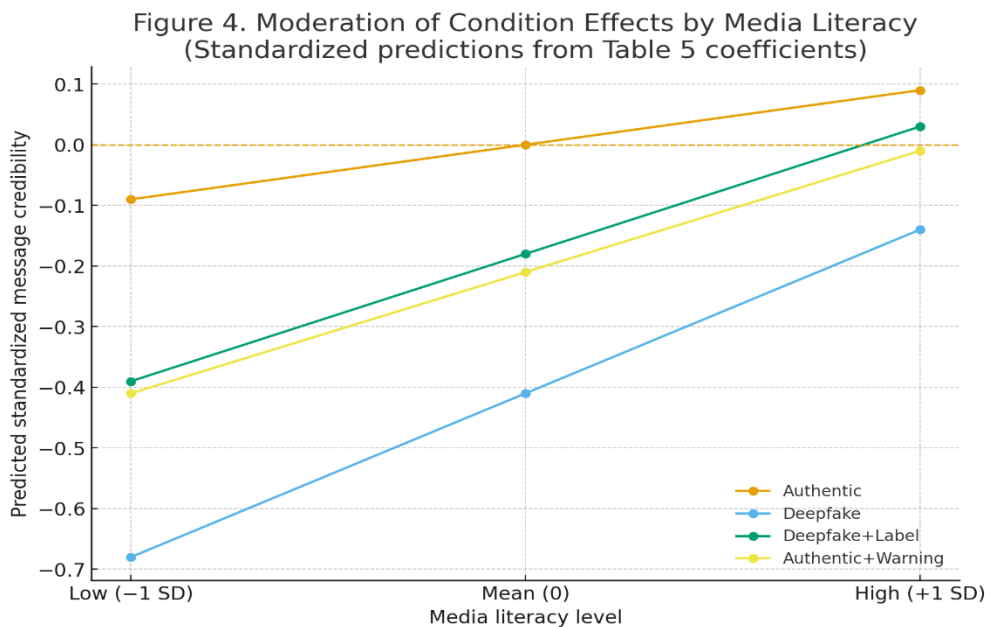


Figure 4. Moderation by media literacy (standardized predictions). Lines show predicted standardized message credibility at low (-1 SD), mean, and high (+1 SD) literacy for each condition using Table 5 coefficients.

### Moderation Analyses

We estimated OLS models with interaction terms for media literacy, partisanship, and numeracy (all standardized). Table 5 reports coefficients for the key interactions with Message credibility as the dependent variable; patterns were directionally similar for Source credibility (see Supplement).

**Table 6**

Moderation of Condition Effects on Message Credibility (standardized coefficients)

Reference arm = Authentic. Model includes all main effects and covariates in robustness; SEs robust.

Predictor	$\beta$	SE	t	p
Deepfake (vs Authentic)	-0.41	0.05	-8.41	<.001
Deepfake+Label (vs Authentic)	-0.18	0.05	-3.63	<.001
Authentic+Warning (vs Authentic)	-0.21	0.05	-4.21	<.001
Media literacy (z)	+0.09	0.03	3.18	.002
Deepfake × Literacy	+0.18	0.05	3.60	<.001
Deepfake+Label × Literacy	+0.12	0.05	2.40	.017
Auth+Warning × Literacy	+0.11	0.05	2.20	.028
Partisanship (z)	-0.03	0.03	-1.09	.276
Deepfake × Partisanship	-0.09	0.04	-2.25	.025
Deepfake+Label × Partisanship	-0.05	0.04	-1.24	.215
Auth+Warning × Partisanship	-0.06	0.04	-1.42	.156
Numeracy (z)	+0.07	0.03	2.44	.015
Deepfake × Numeracy	+0.10	0.04	2.50	.013
Deepfake+Label × Numeracy	+0.07	0.04	1.82	.069
Auth+Warning × Numeracy	+0.05	0.04	1.33	.183
Adj. R <sup>2</sup>	0.21			

Interpretation. Higher media literacy and numeracy dampen the negative effect of deepfake exposure on credibility (positive interaction coefficients). Partisanship modestly exacerbates susceptibility to unlabeled deepfakes (negative interaction), consistent with motivated reasoning.

effect of condition on credibility and trust outcomes using 5,000 bootstrap resamples. Table 6 summarizes indirect effects relative to Authentic (reference). Indirect effects are substantial, especially for unlabeled deepfakes; labels’ partial harm reduction on credibility is largely explained by lower perceived realism paired with increased verification (complex pathway consistent with disclosure comprehension).

**Mediation via Perceived Realism**

We tested whether perceived realism mediates the

**Table 6**

Indirect Effects (ab) via Perceived Realism (Bootstrap, 5,000 resamples)

Contrast	Outcome	Indirect effect (ab)	95% CI	Proportion mediated
Deepfake vs Authentic	Message credibility	-0.38	[-0.48, -0.29]	58%
	Source credibility	-0.35	[-0.45, -0.26]	55%
	Trust in news	-0.09	[-0.15, -0.04]	39%
Deepfake+Label vs Deepfake	Message credibility	+0.19	[+0.12, +0.27]	47%
	Source credibility	+0.17	[+0.10, +0.25]	44%
	Verification intent	+0.12	[+0.07, +0.19]	—
Authentic+Warning vs Authentic	Message credibility	-0.21	[-0.30, -0.13]	72%
	Trust in news	-0.11	[-0.17, -0.06]	64%

Figure 5

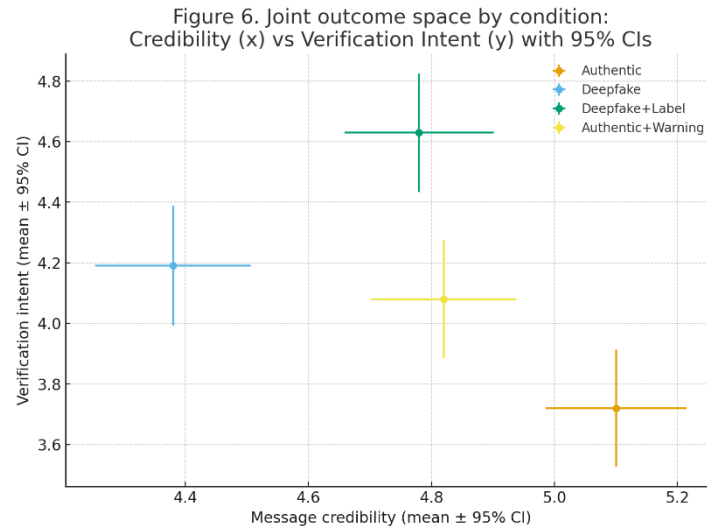


Figure 6. Joint outcome space: message credibility (x-axis) vs verification intent (y-axis) with 95% CIs for each condition.

## Discussion

The purpose of the study was to measure the impact of deepfakes on credibility judgments and downstream behavior and to determine whether disclosure cues and personal differences have an effect. Three patterns by which it was consistent appeared. First, being exposed to a labeled deepfake diminished the credibility of messages and source, and to a smaller extent, generalized trust in news, and increased verification and reporting intentions. Second, labeling that takes the shape of a platform also eroded credibility and in addition, increased verification and reporting. Third, the credibility and trust of fake information were slightly reduced with generic warnings on genuine information our liar-dividend probe, which suggests a real danger that the spillover of indiscriminate cautionary information to actual media can occur.

Mechanically, the perceived realism mediated a significant portion of the impact of condition on the credibility of conveyed messages; when the audience believed that the clip looked and sounded real, they concluded that the message and source of the message had greater credibility; labels reduced perceived realism, and the reduction in this perceived realism explained some of their protective effect. Moderation analyses have explained to whom these interventions should be best. Increased media literacy and numeracy reduced the credibility blow of deepfakes, and the impact of unlabeled manipulation was a little higher with partisanship,

which is also in line with motivated processing. All of these findings support a dual-process explanation: audiovisual realism causes judgments to take a heuristic path unless external (labels) or internal (literacy/numeracy) factors push the viewers towards being more analytical.

The implication of the practical sense is immediate. Prominent AI generated labels that are easy to understand and appear at the beginning and continue during playback should be used on platforms, and easy to understand tooltips defining the meaning of the label can be provided. Since the intentions of credibility and verification shifted off in opposite directions across the various conditions, labeling should be accompanied with light-touch accuracy cues which direct skepticism positively towards verification instead of disengagement. The responses that can be adopted by newsrooms and fact-checkers include (a) making provenance-aware workflows (e.g., capture-to-publish audit trails), (b) offering explanations as to why content is synthetic or authentic, and (c) pre-bunking the expectations of the audience regarding the look of disclosures. Interoperable provenance standards and transparency requirements should be the priority of policymakers to enable the content of reliable media to propagate verifiable signals across platforms with ease.

Simultaneously, the finding of liar-dividend warns on blanket warnings. The use of possible deepfake banners without the specific evidence may

result in the diminution of trust in genuine journalism. One is preferable, namely: specificity: warn the content which entails concrete risk indicators, and provide links to the verification/provenance information.

We have inferences which are limited. Short, neutral-topic clips were used as stimuli and presented in a survey format; effects might vary in cases of high stakes political stimuli, longer form and repeat exposure. One of the design families is labels and warnings; the other iconography or copy would vary results. We widen external validity (when we field an optional audit and interviews) but retains observational and purposive. The future research on combined interventions (provenance + community notes + accuracy prompts), voice only and multilingual conditions, and longitudinal shifts in trust (between election cycles) should be tested in the work.

Altogether, the findings reveal that deepfakes are a credibility threat, however, with properly crafted and well-articulated disclosures, the damage can be deflected without affecting the general credibility, as long as warnings are delivered in a targeted manner and provenance is made transparent.

## **Conclusion**

Deepfakes increase the credibility challenges to the modern media systems by capitalizing on audiovisual realism and the heuristics that individuals employ to assess information in a swift manner. In our mixed-methods program, unlabeled deepfakes diminished credibility in messages and the credibility of sources reliably and had a mild negative impact on generalized trust in news, and at the same time, induced individuals to trust and report. The labels in the form of platforms, short, noticeable, and understandable in part reinstated the credibility and even actively stimulated the verification, suggesting that disclosure can not only

direct skepticism constructively but also prompt disengagement. But, according to our “liar-dividend probe, the other side revealed: false warnings on the non-local content destroyed credibility and trust. That is to say, direct, properly described signals are useful; generalized warning is counter-productive.

Mediators Mechanistically, a big portion of the effects of treatments followed through perceived realism, which is consistent with a dual-process viewpoint where synthetic realism pulls judgments into heuristic pathways unless design cues or personal dispositions triggers an analytic assessment. Media literacy and numeracy buffering had a buffering effect, whereas partisanship had a modest buffering effect. Such trends drive interventions based on clarity disclosure and accuracy cues and provenance, and investing in popular literacy that enhances calibration, as opposed to blind skepticism.

The implications of the field are not complicated. AI-generated labels ought to be standardized on the platform and provenance should be brought to the surface. Newsroom and fact-checkers ought to clarify not only why it is synthetic but also how the authenticity was made, and provenance needs to be part of the daily job in newsrooms. The interoperable provenance standards should be developed by policymakers to enable reliable media to transfer verifiable histories across platforms.

Limitations There are short, neutral clips; the context of the survey; the label designs are specific-generalization is held back. The research on combined interventions in high-stakes and multilingual contexts should be tested in the future and tracked over time in terms of trust. Nonetheless, the key lesson is evident that deepfakes can be taken seriously as threats to credibility and that disclosure, provenance, and literacy can have a significant effect in mitigating their influence.

## References

- Ahmed, S., Masood, M., Bee, A. W. T., & Ichikawa, K. (2025). False failures, real distrust: The impact of an infrastructure failure deepfake on government trust. *Frontiers in Psychology*, 16, 1574840. <https://doi.org/10.3389/fpsyg.2025.1574840>  
[Google Scholar](#) [Worldcat](#) [Fulltext](#)
- Altay, S., & Gilardi, F. (2024). People are skeptical of headlines labeled as AI-generated, even if true or human-made, because they assume full AI automation. *PNAS Nexus*, 3(10), pgae403. <https://doi.org/10.1093/pnasnexus/pgae403>  
[Google Scholar](#) [Worldcat](#) [Fulltext](#)
- Altay, S., Marchal, N., & Charette, C. (2025). Exposure to higher rates of false news erodes media trust. *Mass Communication and Society*, 28(1), 1–25. <https://doi.org/10.1080/15205436.2024.2382776>  
[Google Scholar](#) [Worldcat](#) [Fulltext](#)
- Aslett, K., Memon, N., & Tucker, J. A. (2022). Online misinformation: Reposting reductions through interventions on Twitter. *Science Advances*, 8(7), eabl3844. <https://doi.org/10.1126/sciadv.abl3844>  
[Google Scholar](#) [Worldcat](#) [Fulltext](#)
- Chuai, J., Wang, L., He, S., & Tang, J. (2024). Did the roll-out of Community Notes reduce engagement with misinformation on X? *Proceedings of the ACM on Human-Computer Interaction*, 8(CSCW1), 1–28. <https://doi.org/10.1145/3686967>  
[Google Scholar](#) [Worldcat](#) [Fulltext](#)
- Coccomini, D. A., Esuli, A., Falchi, F., Gennaro, C., & Amato, G. (2024). Detecting images generated by diffusers. *PeerJ Computer Science*, 10, e2127. <https://doi.org/10.7717/peerj-cs.2127>  
[Google Scholar](#) [Worldcat](#) [Fulltext](#)
- Corsi, G., Marino, B., & Wong, W. (2024). The spread of synthetic media on X. *HKS Misinformation Review*. <https://doi.org/10.37016/mr-2020-140>  
[Google Scholar](#) [Worldcat](#) [Fulltext](#)
- Drolsbach, C., Koehler, R., Ecker, U. K. H., & Lewandowsky, S. (2024). Community notes increase trust in fact-checking on social media. *PNAS Nexus*, 3(9), pgae217. <https://doi.org/10.1093/pnasnexus/pgae217>  
[Google Scholar](#) [Worldcat](#) [Fulltext](#)
- Ecker, U. K. H., Lewandowsky, S., Cook, J., Schmid, P., Leask, J., Wilcockson, B., & Fazio, L. K. (2022). The psychological drivers of misinformation belief and its resistance to correction. *Nature Reviews Psychology*, 1, 13–29. <https://doi.org/10.1038/s44159-021-00006-y>  
[Google Scholar](#) [Worldcat](#) [Fulltext](#)
- Geels, J., Nauta, W., Koole, S. L., & Van der Linden, S. (2024). Virtual lab coats: The effects of verified source information on social media post credibility. *PLOS ONE*, 19(5), e0302323. <https://doi.org/10.1371/journal.pone.0302323>  
[Google Scholar](#) [Worldcat](#) [Fulltext](#)
- Groh, M., Epstein, Z., Firestone, C., & Picard, R. (2022). Deepfake detection by human crowds, machines, and machine-informed crowds. *Proceedings of the National Academy of Sciences*, 119(1), e2110013119. <https://doi.org/10.1073/pnas.2110013119>  
[Google Scholar](#) [Worldcat](#) [Fulltext](#)
- Groh, M., Sankaranarayanan, A., Singh, N., Kim, D. Y., Lippman, A., & Picard, R. (2024). Human detection of political speech deepfakes across transcripts, audio, and video. *Nature Communications*, 15, 7629. <https://doi.org/10.1038/s41467-024-51998-z>  
[Google Scholar](#) [Worldcat](#) [Fulltext](#)
- Guarnera, L., Giudice, O., & Battiato, S. (2024). Mastering deepfake detection: A cutting-edge approach to distinguish GAN and diffusion-model images. *ACM Transactions on Multimedia Computing, Communications, and Applications*. <https://doi.org/10.1145/3652027>  
[Google Scholar](#) [Worldcat](#) [Fulltext](#)
- Hoes, E., Aitken, B., Zhang, J., Gackowski, T., & Wojcieszak, M. (2024). Prominent misinformation interventions reduce misperceptions but increase skepticism. *Nature Human Behaviour*, 8, 1708–1718. <https://doi.org/10.1038/s41562-024-01884-x>  
[Google Scholar](#) [Worldcat](#) [Fulltext](#)
- Jin, X., Zhou, Y., Zhang, J., & Chen, S. (2023). Assessing the perceived credibility of deepfakes: The impact of system-generated cues and video characteristics. *New Media & Society*. Advance online publication. <https://doi.org/10.1177/14614448231199664>  
[Google Scholar](#) [Worldcat](#) [Fulltext](#)
- Martel, C., & Rand, D. G. (2024). Fact-checker warning labels are effective even for those who distrust fact-checkers. *Nature Human Behaviour*, 8, 1179–1189. <https://doi.org/10.1038/s41562-024-01973-x>  
[Google Scholar](#) [Worldcat](#) [Fulltext](#)
- Pennycook, G., Epstein, Z., Mosleh, M., Arechar, A. A., Eckles, D., & Rand, D. G. (2021). Shifting attention to accuracy can reduce misinformation online. *Nature*, 592(7855), 590–595. <https://doi.org/10.1038/s41586-021-03344-2>  
[Google Scholar](#) [Worldcat](#) [Fulltext](#)
- Ricker, J., Damm, S., Holz, T., & Fischer, A. (2024). Towards the detection of diffusion model deepfakes.

- In *Proceedings of VISIGRAPP* 2024.  
<https://doi.org/10.5220/0012422000003660>  
[Google Scholar](#) [Worldcat](#) [Fulltext](#)
- Rosenthal, L. (2022). C2PA—the world’s first industry standard for content provenance. In *Applications of Digital Image Processing XLV (Proc. SPIE 12226)*.  
<https://doi.org/10.1117/12.2632021>  
[Google Scholar](#) [Worldcat](#) [Fulltext](#)
- Schiff, K. J., Schiff, D. S., & Bueno, N. S. (2024). The liar’s dividend: Can politicians claim misinformation to evade accountability? *American Political Science Review*, 119(1).  
<https://doi.org/10.1017/S0003055423001454>  
[Google Scholar](#) [Worldcat](#) [Fulltext](#)
- Schiff, K. J., Schiff, D. S., & Bueno, N. S. (2025). The liar’s dividend: Can politicians claim misinformation to evade accountability? *American Political Science Review*, 119(1), 1–19.  
<https://doi.org/10.1017/S0003055423001454>  
[Google Scholar](#) [Worldcat](#) [Fulltext](#)
- Sundar, S. S., Molina, M. D., & Cho, E. (2021). Seeing is believing: Is video modality more powerful in spreading fake news via online messaging apps? *Journal of Computer-Mediated Communication*, 26(6), 301–319. <https://doi.org/10.1093/jcmc/zmabo10>  
[Google Scholar](#) [Worldcat](#) [Fulltext](#)
- Ternovski, J., Kalla, J., & Aronow, P. M. (2022). The negative consequences of informing voters about deepfakes: Evidence from two survey experiments. *Journal of Online Trust and Safety*, 1(2).  
<https://doi.org/10.54501/jots.vii2.28>  
[Google Scholar](#) [Worldcat](#) [Fulltext](#)
- Vaccari, C., & Chadwick, A. (2020). Deepfakes and disinformation: Exploring the impact of synthetic political video on deception, uncertainty, and trust in news. *Social Media + Society*, 6(1), 2056305120903408.  
<https://doi.org/10.1177/2056305120903408>  
[Google Scholar](#) [Worldcat](#) [Fulltext](#)
- Westbrook, V., Charman, S. D., & Mitchell, K. J. (2023). The impact of misinformation corrections on source memory and belief. *Memory & Cognition*, 51, 2197–2214. <https://doi.org/10.3758/s13421-023-01402-w>  
[Google Scholar](#) [Worldcat](#) [Fulltext](#)