p-ISSN: 2521-2982
e-ISSN: 2707-4587

GLOBAL Political REVIEW
www.gprjournal.com
empowering humanity

GPR

GLOBAL POLITICAL REVIEW

HEC-RECOGNIZED CATEGORY-Y

VOL. X, ISSUE III, SUMMER (SEPTEMBER-2025

DOI (Journal): 10.31703/gpr
DOI (Volume): 10.31703/gpr/.2025(X)
DOI (Issue): 10.31703/gpr.2025(X.III)

Double-blind Peer-review Research Journal
www.gprjournal.com

HumanityPublications
sharing research

Humanity Publications (HumaPub)
www.humapub.com
Doi: https://dx.doi.org/10.31703

Article Title

Combating Misinformation in the Age of Deepfakes

Abstract

The study discusses the defense in relation to the problem of deepfakes using a mixed methodology by implementing technical assessment, randomized controlled tests (RCT), and interviews with practitioners. The best out-of-distribution performance was the multimodal fusion and late ensembling (ensemble AUC 0.979 in-domain, 0.944 out-of-domain) and high-resistance to post-processing methods. Combining prebunking with AI-generated labels led to a 31.4% reduction in the intent to share synthetic content compared to the control group in the RCT (n ≈ 900) and an apparent reduction in the perceived accuracy of fakes. Provenance and watermarks checks were found to be very precise, covering 23% of items. The practitioner responses mentioned clear labels, confidence displays, and audit processes as being of significant value. Deep fake threats can be efficiently addressed by using a layered defense mechanism, which consists of calibrated multimodal detection, interoperable provenance, and timely user education.

Keywords: Deepfakes, Misinformation, Detection, Provenance, Watermarking, Prebunking, Human–Ai Teaming

Authors:

Muntaha Sami: (Corresponding Author)
M.Phil, School of Media and Communication, Studies, Minhaj University, Lahore, Punjab, Pakistan.
(Email: muntahasamioo@gmail.com)

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

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

Pages: 127-143
DOI:10.31703/gpr.2025(X-III).12
DOI link: https://dx.doi.org/10.31703/gpr.2025(X-III).12
Article link: https://gprjournal.com/article/combating-misinformation-in-the-age-of-deepfakes
Full-text Link: https://gprjournal.com/article/combating-misinformation-in-the-age-of-deepfakes
Pdf link: https://www.gprjournal.com/jadmin/Auther/31rvIolA2.pdf

Global Political Review
p-ISSN: 2521-2982 **e-ISSN:** 2707-4587
DOI (journal): 10.31703/gpr
Volume: X (2025)
DOI (volume): 10.31703/gpr.2025(X)
Issue: III Summer (September-2025)
DOI(Issue): 10.31703/gpr.2025(X-III)

Home Page
www.gprjournal.com

Volume: X (2025)
https://www.gprjournal.com/Current-issue

Issue: III-Summer (September-2025)
https://www.gprjournal.com/issue/10/3/2025

Scope
https://www.gprjournal.com/about-us/scope

Submission
https://humaglobe.com/index.php/gpr/submissions

Google scholar

Visit Us

Humanity Publications (HumaPub)
www.humapub.com
Doi: https://dx.doi.org/10.31703

Citing this Article

12	Combating Misinformation in the Age of Deepfakes		
Authors	Muntaha Sami Zarwah Nabil Maryam Hashmi	DOI	10.31703/gpr.2025(X-III).12
		Pages	127-143
		Year	2025
		Volume	X
		Issue	III

Referencing & Citing Styles	
APA	Sami, M., Nabil, Z., & Hashmi, M. (2025). Combating Misinformation in the Age of Deepfakes. *Global Political Review*, *X*(III), 127-143. https://doi.org/10.31703/gpr.2025(X-III).12
CHICAGO	Sami, Muntaha, Zarwah Nabil, and Maryam Hashmi. 2025. "Combating Misinformation in the Age of Deepfakes." *Global Political Review* X (III):127-143. doi: 10.31703/gpr.2025(X-III).12.
HARVARD	SAMI, M., NABIL, Z. & HASHMI, M. 2025. Combating Misinformation in the Age of Deepfakes. *Global Political Review,* X**,** 127-143.
MHRA	Sami, Muntaha, Zarwah Nabil, and Maryam Hashmi. 2025. 'Combating Misinformation in the Age of Deepfakes', *Global Political Review*, X: 127-43.
MLA	Sami, Muntaha, Zarwah Nabil, and Maryam Hashmi. "Combating Misinformation in the Age of Deepfakes." *Global Political Review* X.III (2025): 127-43. Print.
OXFORD	Sami, Muntaha, Nabil, Zarwah, and Hashmi, Maryam (2025), 'Combating Misinformation in the Age of Deepfakes', *Global Political Review,* X (III), 127-43.
TURABIAN	Sami, Muntaha, Zarwah Nabil, and Maryam Hashmi. "Combating Misinformation in the Age of Deepfakes." *Global Political Review* X, no. III (2025): 127-43. https://dx.doi.org/10.31703/gpr.2025(X-III).12.

Pages: 127-143 **e-ISSN:** 2707-4587 **p-ISSN:** 2521-2982 **Doi:** 10.31703/gpr.2025(X-III).12

Global Political Review
www.gprjournal.com
DOI: http://dx.doi.org/10.31703/gpr

Volume: X (2025) **URL:** https://doi.org/10.31703/gpr.2025(X-III).12 **Issue:** III-Summer (September-2025)

Title

Combating Misinformation in the Age of Deepfakes

Authors:

Muntaha Sami: (Corresponding Author)
M.Phil, School of Media and Communication, Studies, Minhaj University, Lahore, Punjab, Pakistan.
(Email: muntahasamioo@gmail.com)

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

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

Contents

Abstract

The study discusses the defense in relation to the problem of deepfakes using a mixed methodology by implementing technical assessment, randomized controlled tests (RCT), and interviews with practitioners. The best out-of-distribution performance was the multimodal fusion and late ensembling (ensemble AUC 0.979 in-domain, 0.944 out-of-domain) and high-resistance to post-processing methods. Combining prebunking with AI-generated labels led to a 31.4% reduction in the intent to share synthetic content compared to the control group in the RCT (n ≈ 900) and an apparent reduction in the perceived accuracy of fakes. Provenance and watermarks checks were found to be very precise, covering 23% of items. The practitioner responses mentioned clear labels, confidence displays, and audit processes as being of significant value. Deep fake threats can be efficiently addressed by using a layered defense mechanism, which consists of calibrated multimodal detection, interoperable provenance, and timely user education.

Keywords:

Deepfakes, Misinformation, Detection, Provenance, Watermarking, Prebunking, Human–Ai Teaming

Introduction

The synthetic media generated by AI, informally known as deepfakes, has now not only moved out of research labs but into regular feeds, and the gap between creation and virality has only shortened. Photo-realistic faces, full-body videos, and human-like voices are currently synthesized at consumer scale using diffusion and transformer-based models; the skill and cost barriers to these models are significantly reduced by open-source ecosystems and turnkey apps (Ricker et al., 2024; Wani et al., 2024). Such functions introduce incontrovertible creative and commercial benefits; however, they also escalate well-known forms of misinformation

harm: reputational harm, financial frauds, targeted harassment, and the loss of epistemic trust in institutions and media (Diel et al., 2024). Most importantly, the likelihood that the capacity of authentic content to be categorized as a fake one increases with the enhancement of synthetic media, making it harder to be verified by journalists, platforms, and the audience (Feng et al., 2023; Moruzzi, 2025).

Empirical studies of recent times highlight two dynamics. To start with, the technical frontier is changing: detectors conditioned on GAN-level models are less generalizable when provided with the outputs of diffusion models without re-training or domain adaptation, and even in that case, do not hold up against distribution shift and compression (Ricker et al., 2024; Wang et al., 2025). Second, individuals will not be particularly good at distinguishing between real and fake: a meta-analysis of 56 studies estimates the accuracy of human deepfake detection to be just above the level of chance, and overconfidence is a common phenomenon (Diel et al., 2024). The convergence of these factors, such as increased generators that change quickly, unstable detectors, and human fallibility, forms a favorable environment for strategic programs of deceit, impersonation scams based on audio cloning of voices, and cascades of viral rumors that are massified through algorithm amplification (Wani et al., 2024). Meanwhile, platform and policy responses (e.g., labeling, takedowns, watermarking, and provenance) are not evenly adopted and deployed, and the usability, governance, and rights implications are open to question (Feng et al., 2023; Moruzzi, 2025).

Deepfaking enhances misinformation on three dimensions. Sensory realism: when multimodal (image, video, audio) synthesis is present, many fast, heuristic processes are engaged to generate the most believable and emotional reactions to the stimulus compared to text to occur alone, in that order (Diel et al., 2024). Personalization: the ability to clone the face or voice of a particular person gives spurious assertions parasocial authority and disseminates rapidly through the inaccessible paths of interpersonal interaction, which are not easily controlled (Wani et al., 2024). Ambiguity leverage: since the concept of authenticity is put into doubt, the malign actors do not have to persuade, but merely plant the seed of doubt and boost the liar dividend in situations of conflict, like in elections and conflicts (Feng et al., 2023). These characteristics justify how seemingly-neutral detection measures (e.g., AUC) can exaggerate real-life protection; small false-negative clusters have the power to do much social harm, and false positives may suppress speech or victimize an innocent individual.

The defense stack is still not complete in spite of the fast development. State-of-the-art detectors are easily out of distribution, especially when trained on diffusion-model-generated artifacts and in-the-wild compressed media (Ricker et al., 2024; Wang et al., 2025). Provenance and watermarking solutions, including cryptographic content credentials and statistical watermarks, are promising but experience adoption frictions, adversarial removal, and interoperability issues, as well as user-interface challenges that can inadvertently undermine trust towards legit media (Dathathri et al., 2024; Feng et al., 2023). On the human front, labeling policies and false tag interventions have both beneficial and adverse effects, whereas scalable prebunking/inoculation strategies have more reliable and cross-domain susceptibility reductions-but with unanswered questions of sustenance, dose, and label interaction (McPhedran et al., 2023; Pennycook et al., 2021).

The agenda of this paper is a mix of methods to: (1) assess cross-generator generalization of recent audio-visual deepfake detectors in realistic distortions; (2) evaluate interventions to users, such as content provenance labels and brief prebunking messages, on the belief and sharing-intention scale; and (3) analyze policy and design trade-offs of platform-level implementation, such as the failure modes and considerations of fairness (Ricker et al., 2024; Dathathri et al., 2024; Feng et.

We refer to deepfakes as the AI-created or the most AI-edited media meant to portray events, speech, or identities that did not take place (Ricker et al., 2024). Misinformation refers to fake information that is posted without the intention to cause harm; disinformation refers to fake information that is posted with the purpose of deceiving with strategic intent (McPhedran et al., 2023). Provenance is cryptographically verifiable metadata about the origin, authorship, and edit history of an object (e.g., content credentials along with similar C2PA-like schemes), which is

contrasted with watermarking, which adds identifiable information into the content (Dathathri et al., 2024; Feng et al., 2023). Detection includes computational means of classifying the authenticity of the media; user-facing interventions (labels, prebunks, frictions) designed to modify belief/behavior (Pennycook et al., 2021).

Section 2 scans generation, detection, provenance/watermarking, platform policies, and human-factor evidence (Wang et al., 2025; Wani et al., 2024). Section 3 elaborates on the mixed-method: benchmark construction, models/baselines, randomized user study, and plan of analysis. Section 4 presents technical and behavioral outcomes, robustness tests, and failure tests. Section 5 interprets the results to human-AI teaming and design of governance, such as limitations and external validity. Section 6 ends with a practical roadmap that incorporates detection, provenance, literacy, and policy levers (Dathathri et al., 2024; Feng et al., 2023; McPhedran et al., 2023).

Overall, the main issue is that deepfakes are faster than humans and existing technical defenses, and the problem of information integrity in large quantities can be endangered. Our question is therefore as follows: RQ1: To what extent do state-of-the-art detectors extrapolate to unseen diffusion-generated media in the presence of realistic platform distortions? RQ2: Does provenance labelling and prebunking (only and combined) help to reduce belief in and sharing deepfakes without compromising belief in authentic content? RQ3: Which deployment trade-offs (usability, fairness, liar-dividend risk) are produced by combining detection with provenance and prebunking? We plan to establish a stress-test benchmark; suggest a calibrated multimodal detector; conduct a mass, pre-registered user study of the interventions; and synthesize design and policy guidance about how to adopt the platform-scale (Ricker et al., 2024; Wang et al., 2025; Dathathri et al., 2024; Feng et al., 2023; McPhedran et al., 2023; Pennycook et al., 2021).

Literature Review

A Genmedia has been experiencing a sudden transformation to GAN-only pipelines to diffusion-based and hybrid systems that render photorealist synthesis a significantly simpler problem, thus facilitating misinformation creation. Variants of GAN (e.g., StyleGAN family) facilitated early face swaps and reenactment, but diffusion models have become the predominant models in high-fidelity, controllable generation of images and video. Latent Diffusion Models (LDMs) demonstrated that operating in compressed latent space provides huge quality and performance advantages and makes it possible to use a large model in open-source toolchains and consumer applications (Rombach et al., 2022). Text-to-image evolved into text-to-video with transformer-diffusion hybrids and masked video transformers, and turnkey face/voice cloning stacks also use speaker encoders, neural vocoders, and cross-modal prompting to produce lifelike identity mimicry with just a few minutes (or even seconds) of source material. Trends indicate: (i) more general models (few-/zero-shot identity, cross-lingual and emotive voice), (ii) tool commoditization (web UIs, mobile apps, plug-ins), and (III) prompt-based and reference-based editing, which erases the distinction between creation and manipulation. Together, these developments squeeze skills, time, and cost boundaries and increase the scale of attack surfaces of misinformation. (Rombach et al., 2022).

Signal/artifact cues. Classic detectors are trained on forensic tells that are left by generative pipelines such as frequency discrepancies, demosaicing/upsampling artifacts, blend edges, and compression artifacts. The theme of high-frequency modeling is recurring: frequency-conscious features are learned to enhance cross-dataset performance by not overfitting textures in datasets (Luo et al., 2021; Tan et al., 2024). Self-blended image training (SBI) enhances data with realistic artifacts, which increases imaging post-processing robustness (Shiohara and Yamasaki, 2022). This is likely the situation, yet it's impossible to be sure that one has all the important details.<|human|>It is probably so, but one cannot be certain that they get all the crucial information.

Physiological/behavioral cues. A second line uses biological signals and behavior, which are difficult to emulate in a consistent way. Visual speech representations are trained in LipForensics to find mouth-motion anomalies and extrapolate to invisible manipulations; better cross-manipulation has also been demonstrated with rPPG/heartbeat signals (micro-variations in skin color), which the system can use, but can be misled by low resolution, occlusion, or heavy post-processing (Qi et al., 2020;

Ciftci et al., 2020; Haliassos et al., 2021). (Haliassos et al., 2021; Qi et al., 2020).

Multimodal fusion. Since synthesis increasingly involves audio, visuals, and text overlays, it is shifting to multimodal fusion (e.g., audio-visual sync, prosody-lip coherence, cross-modal contradictions). According to current surveys, there is a transition towards multi-modal pipelines and the necessity of appearance, motion, and sound-reasoning detectors (Gong et al., 2024; Heidari et al., 2024). (Gong et al., 2024; Heidari et al., 2024).

Adaptation to adversity, generalization, performance, and robustness. One enduring problem is that of generalization - detectors that have been trained on a set of manipulations might fail on others. Frequency-conscious design and data augmentation are beneficial, but within-dataset performance remains lower than cross-dataset performance (Luo et al., 2021; Heidari et al., 2024). The detector may also be adversarially attacked: in small perturbations or degradation-matching processes, predictions may be flipped, but not noticed (Hou et al., 2023; Wang et al., 2024). Ensemble approaches and spectrum-disjoint defenses (e.g., D4) enhance black-box defenses, and there is limited empirical evidence in the wild. In general, the most effective systems in the modern world involve a combination of artifact signals, physiology/behavioral signals, and training programs that are heavy with augmentation to achieve more successful transfer- however, a primary risk of the production environment is that model fragility in the face of domain shift and active avoidance (Shiohara and Yamasaki, 2022; Tan et al., 2024; Heidari et al., 2024).

Cryptographic provenance. Instead of spotting a fake, provenance standards are expected to establish the real. The Coalition for Content Provenance and Authenticity (C2PA), 2025 specification is a set of signed, constructively assertive manifests that tie media to trusted capture/edit metadata ("Content Credentials) so that the provenance and edit history of the content can be verified throughout the ecosystem (C2PA, 2025; CAI, 2025). Similar research has been done in JPEG Trust and JUMBF; interoperable containerization of authenticity metadata of media across media types (Temmermans et al., 2024; Temmermans et al., 2021). Human-factors studies show that provenance labels can decrease the trust in fake media, but the wording and positioning are important (Wittenberg et al., 2025). This information is essential even for individuals outside the field of sociology.<|human|>It is a fact that even people who are not in the sphere of sociology are interested in this information.

Watermarking & steganalysis. Generative model watermarks attempt to find model-side watermarks that remain during typical transformations. Recent ACM publications focus on semantic and frequency-domain robustness, diffusion latent space watermarking, and model-inversion/erasure resistance attacks. Previously, it was indicated that this process may be inhibited by either the activation of the secluded region or prolonged treatment with buprenorphine, which induces a comparable effect on the brain's glutamate cycle (Huang et al., 2023; Fernandez et al., 2023; Zhang et al., 2023). Nevertheless, the strongest watermarking scheme does not exist: down- / re-sampling, re-generation, aggressive edits, or adversarial fine-tuning can be used to undermine the detection, whereas false positives can be used to punish legal content. The most promising approach to providing end-to-end assurance in mixed pipelines is through integrations that combine cryptography manifests, strong watermarking, and platform-side checks. (Huang et al., 2023; Fernández et al., 2023; Zhang et al., 2024).

The human prone-ness is at the core: individuals overrate fluency and plausibility, which is the case when the content is in line with existing beliefs (confirmation bias), as well as may falsely project synthetic realism onto the truth. Tactic-oriented, micro-lessons done just before exposure, known as psychological inoculation (prebunking), have been proven to enhance resistance on social platforms as well as between cultures at scale (Roozenbeek et al., 2022). Nevertheless, meta-findings indicate inconsistent spillovers; some interventions decrease belief but have minimal impacts on curtailing engagement intentions, and labeling may have implied truth side effects when inconsistently administered (Wittenberg et al., 2025; Li and Yang, 2024; Hoes et al., 2024). On the whole, the evidence supports prebunking and accuracy-nudges that are audience and context-specific and clear information on how materials were created (process-based labels) or why they may be misleading (harm-based labels). Hoes et al. (2024) determined that in most

cases, it leads to confusion and results in confusion, discord, and disputes.<|human|>Hoes et al. (2024) concluded that it is confusing and generates confusion, disagreement, and conflicts in the vast majority of situations. Peace journalism emphasizes how media narratives can either escalate or mitigate conflict. By focusing on accuracy and responsibility, journalists can counter misinformation and reduce polarization. As Hussain and Lynch (2015) highlight, constructive media practices encourage understanding and lessen the social damage caused by misleading or conflict-driven reporting.

Moderation toolkit platforms are increasingly launching AI triage and provenance display, warning labels, and takedown regimes based on synthetic impersonation and election-related harms. Such policy alignment is being introduced by the requirements of the EU in its AI Act and Digital Services Act that content generated by AI should be labeled, and that risk-mitigation and transparency reporting processes should be executed, and similar action is being undertaken in various jurisdictions (Łabuz, 2024). The evidence regarding the efficacy of labels supports the idea that the type of label and details of its implementation determine the results; clear reporting and auditability are required to monitor the effects of label precision/recall and the effects of false-positive and appeals (Wittenberg et al., 2025). (Łabuz, 2024; Wittenberg et al., 2025).

Deepfakes involve defamation (lying, which damages reputation), privacy, consent (non-consent sex deepfakes in particular), election integrity, and verify-as-true journalism. U.S. commentary identifies tort and criminal framework lapse in cross-platform harms, and jurisdiction issues; suggestions are disclosure requirements, revision of right-of-publicity laws, and malicious synthetic media-specific offenses (Chawki, 2024). Ethical advice focuses on newsroom guidance (provenance verifications, source verification, open corrections), systematic labelling in the event of synthetic assets being utilized to illustrate or satirize, to prevent being misled by the audience (Lundberg, 2024). The disclosure requirements of the AI Act and the systemic-risk requirements of the DSA do set commitments of detection, labeling, and mitigation in the EU, but the very practical implementation and compatibility with cryptographic provenance are not yet answered (Łabuz, 2024). (Chawki, 2024; Lundberg, 2024; Łabuz, 2024).

Methodology:

Introduction

This study employs a mixed-methods design to evaluate technical, behavioral, and workflow interventions for reducing the spread and impact of deepfake-driven misinformation. The core goals are to: (i) benchmark the *generalization* and *robustness* of state-of-the-art (SOTA) deepfake detectors across unseen generators and real-world post-processing; (ii) test user-facing interventions (provenance labels and prebunking messages) in a randomized controlled trial (RCT); and (iii) surface practical considerations via semi-structured interviews with journalists, fact-checkers, and platform trust & safety staff. The methodological choices emphasize external validity (cross-generator and cross-modality evaluation), causal identification (randomization and preregistration for the user RCT), and practical deployability (latency, calibration, and operator trust).

Study Design (Mixed-Methods)

Technical evaluation (benchmark). We implement a standardized evaluation harness that trains/assesses multiple detectors on curated, multi-source datasets with *cross-generator* and *cross-modality* splits. We measure in-domain accuracy, out-of-distribution (OoD) drop, calibration, runtime, and memory footprint.

User experiment (RCT). We run a 4-arm, between-subjects RCT (Control / Label / Prebunking / Combined) online. Participants view short media stimuli (video, audio, text+video composites) and report perceived accuracy, sharing intent, and detection confidence. Treatments are applied at exposure time (labels) and/or *before* exposure (prebunking).

Qualitative interviews (optional). We conduct semi-structured interviews (≈20–30 participants) with newsroom and platform practitioners to understand labeling usability, thresholds for action, failure modes (false positives/negatives), and audit/reporting needs. Interviews complement the RCT by grounding design decisions in real workflows.

Integration occurs at interpretation: detector outputs (scores, confidence, provenance checks) are mapped to UI treatments; interview findings inform

feasible label wording/placement and escalation playbooks.

Datasets & Sampling

Sources: We assemble a balanced corpus from public deepfake/video manipulation datasets (e.g., widely used face-swap/reenactment corpora), public audio spoofing corpora, and *newly generated* clips produced with contemporary diffusion/GAN pipelines for faces and voices. For each synthetic item, we include a *matched* real counterpart by the same or visually/audibly similar subjects, under similar lighting, compression, and context.

Modalities and topics: We stratify by modality (1) video-only, (2) audio-only, (3) synchronized audio–video (text+video overlays allowed) and by topic domain (politics/public affairs; entertainment/celebrity; health/consumer). We also stratify by *quality tier*: (a) high-resolution originals; (b) platform-like re-encodes (e.g., strong compression, resizing); (c) user-forwarded re-uploads (cropping, filters).

Splits: To test generalization, we adopt a *leave-generators-out* protocol: certain generator families and editing pipelines are absent from training/validation but appear in test. We maintain a separate adversarial stress-test set (Sec. 3.9). For the RCT, stimuli are sampled from the test pool with balanced modality, topic, and quality. Stimuli with sensitive private individuals are excluded; public figures and consented actors are used.

Sample size: The technical benchmark encompasses ~80–120 hours of video and ~1,500–2,000 audio clips (balanced real/synthetic). For the RCT, power simulations (α=.05, two-sided) targeting a 20% relative reduction in sharing intent (baseline 0.35), with random intercepts for participant and stimulus (ICC≈.05), indicate n≈900 participants (≈225/arm) yields ≥.90 power after multiple-comparisons control. Qualitative interviews proceed until thematic saturation (anticipated 20–30).

Detectors & Baselines

Image/video artifact baselines. Frequency-aware CNNs and patch-level forensics models (e.g., Xception/EfficientNet backbones with spectrum augmentations); temporal models (I3D/X3D or TSM) for motion artifacts and frame-consistency cues. Physiological/behavioral. Lip-motion/viseme coherence detectors leveraging audio–visual speech embeddings; remote photoplethysmography (rPPG)-based models for subtle skin-color rhythm cues in faces (robustness tested under resolution and lighting changes).

Audio anti-spoofing. CNN/ResNet and transformer architectures over log-mel features and raw waveforms (e.g., ECAPA-TDNN, RawNet2-style) targeting vocoder artifacts, periodicity anomalies, and prosody consistency.

Multimodal fusion: Cross-modal transformer that ingests visual frames, optical flow, ASR text, and acoustic embeddings, with co-attention for A/V alignment. Late fusion ensembles combine artifact, physiological, and audio detectors.

Provenance/watermark verification (if included). A sidecar pipeline attempts: (i) cryptographic provenance validation when *Content Credentials* (C2PA-like manifests) are present; (ii) model-side watermark probing for diffusion-generated media; (iii) heuristic steganalysis on residuals. These outputs are not used to *train* detectors but are logged and surfaced to the RCT UI as labels when appropriate.

Implementation details: All models are trained with standardized augmentations (color jitter, re-encode, blur, scale/crop), mixed precision, and early stopping on validation AUC. We fix seeds, log configs/artifacts, and checkpoint via a reproducible harness.

Interventions (User Study Conditions)

- Control: No warning, no labels.
- Label: A small, persistent badge at the top-left of the player: *"AI-generated or Unverified Source."* Hover reveals a short explainer and link to "How we assess content."
- Prebunking: Before any exposure, a 30–45s interactive micro-lesson describing common deepfake tactics (lip-sync mismatch, lighting inconsistencies, rPPG absence, synthetic voice prosody), with two practice items and immediate feedback.
- Combined: Prebunking + Label.

Label wording/placement is fixed across stimuli for internal validity; variants are explored in sensitivity analyses (Sec. 3.9).

Measures & Instruments:

Technical metrics.

- AUC (primary), EER, F1, accuracy at operating points;
- Calibration: Brier score and Expected Calibration Error (ECE); reliability diagrams;
- OoD drop: $\Delta AUC = AUC_{in} - AUC_{OoD}$;
- Latency: ms/frame and end-to-end decision time;
- Resource: parameter count and peak memory.

User Outcomes

- Perceived accuracy (Likert 1–7).
- Sharing intent (binary and Likert).
- Detection confidence (1–7).
- Time-on-task and hover/expand interactions (proxy for engagement with labels).

Trust & Overreliance (complacency). We include a small subset of trials (≤10%) where the label is intentionally *incorrect* (ethically debriefed post-study). The Complacency Index = P(share | incorrect "authentic" label) – P(share | no label) for matched stimuli; higher values indicate overreliance on automation.

Covariates. Media literacy (short validated scale), political knowledge (brief quiz), platform usage frequency, and demographics (age, gender, region, language proficiency).

Procedure:

Technical Pipeline

1. Training/zero-shot: For artifact/physiological/audio models, we train on a subset of generator families and evaluate *zero-shot* on held-out families. Multimodal fusion is trained only on modalities available at inference.
2. Cross-generator split: Families A/B for training, C/D for validation, E/F for test; new diffusion pipelines appear only in E/F.
3. Adversarial stress tests: Evaluate after transformations (JPEG 10–50, Gaussian noise $\sigma \in [2,8]$, Gaussian blur $r \in [1,3]$, color shifts, time resampling, re-encode), and after attack-oriented edits (light face warping, speech rate change ±10%).
4. Provenance/watermark checks: When manifests exist, verify signature chain; when absent, log "no provenance" and record watermark detector confidence.

User RCT

- Consent & screening: Adults (18+), language proficiency sufficient for instructions; attention checks included.
- Randomization: Individual-level random assignment (1:1:1:1). Stimuli order counterbalanced; each participant sees a balanced set across modality/topic/quality.
- Intervention delivery: Prebunking (if assigned) precedes any exposure; labels (if assigned) appear on every stimulus.
- Outcomes survey: After each stimulus: perceived accuracy, sharing intent, detection confidence; final block collects covariates and manipulation checks.
- Debriefing: Reveal purpose, explain the occasional incorrect-label trials, and provide media-literacy resources.

Interviews. 45–60 min sessions over video; guide covers labeling thresholds, appeals, provenance tooling, and metrics/reporting needs. Sessions are recorded, transcribed, and pseudonymized.

Statistical Analysis

Power & sampling. Simulation-based power analysis for mixed-effects logistic regression on sharing intent, with random intercepts for participant and stimulus, and fixed effects for condition, modality, topic, and quality. Target n≈900 ensures ≥.90 power to detect a 20% relative reduction vs. Control ($q<.05$ FDR-corrected).

Primary Model (Behavioral)

- Sharing intent (binary): logit link with fixed effects for Condition (3 dummies vs. Control), Modality, Topic, Quality, plus interactions Condition×Modality and Condition×Quality; random intercepts for Participant and Stimulus.
- Perceived accuracy and confidence (Likert): linear mixed models (robust SEs), with identical structure.
- Complacency Index: between-group contrasts; permutation tests for robustness.

Multiple comparisons. Benjamini–Hochberg FDR at q=.05 across primary contrasts; Holm adjust for secondary outcomes.

Heterogeneity. Subgroup analyses by prior media literacy (median split), platform usage (high/low), and region/language. Report conditional marginal effects with 95% CIs.

Technical benchmark analysis. Compare detectors via paired bootstrap on AUC/EER; test ΔAUC across OoD splits; assess calibration with ECE and reliability curves; compute speed/accuracy Pareto frontiers. Rank models by a composite utility score U = AUC − λ·ECE − μ·Latency (λ, μ pre-specified).

Missing data & exclusions. Pre-registered rules: exclude participants failing ≥2 attention checks or spending <1/3 median time; use mixed-models' robustness to unbalanced cells; no outcome imputation for primary binary endpoint.

Robustness, Bias & Sensitivity

Fairness. For face/voice media featuring people, annotate demographics (apparent skin tone categories, gender presentation) and accents (self-reported for actors; inferred for public figures where ethically acceptable). Compute performance deltas across groups for detectors (ΔAUC_g) and for RCT treatment effects (ΔATE_g). Report any gaps ≥5 percentage points and conduct sensitivity checks (reweighting by group prevalence).

Adversarial perturbations. Evaluate detectors under: compression (JPEG 10–90), noise, blur, resampling, frame rate changes, cropping, color shifts, and mixed transformations simulating platform pipelines. For audio: time-stretch ±10%, pitch-shift ±2 semitones, low-bitrate codecs.

Ablations & Variants.

- Modality ablation: remove audio or visual branch from the fusion model.
- Label wording/placement: test "AI-generated," "Altered or synthetic," and "Unverified source"; test top-left vs. below-player placement in a holdout sample.
- Prebunking length: 15s vs. 45s micro-lessons in a sensitivity subsample.
- Detector dependence: run RCT analyses conditioned on whether backend provenance was available (manifest present vs. absent).

Calibrated thresholds. Explore operating points that equalize FPR across groups (equalized odds–inspired heuristic) and measure impact on AUC/EER.

Ethics & Governance

Ethics review. The protocol (benchmark, RCT, interviews) undergoes IRB/ethics approval prior to data collection. All participants provide informed consent and can withdraw without penalty.

Risk mitigation. To minimize harm from exposure to misleading or sensitive content, stimuli avoid graphic violence or hate speech; debriefing clarifies manipulations and provides media-literacy resources. Incorrect-label trials are limited and disclosed post-study.

Privacy & consent. Newly generated clips use consenting adult actors or public-figure materials in clear public-interest contexts; voice clones for actors are created with explicit written consent. All PII is removed from datasets; interview transcripts are pseudonymized and stored on encrypted drives with access controls.

Data governance. We publish a detailed model/dataset card (sources, licenses, known biases, intended use). Release of generated stimuli is restricted to low-resolution derivatives with visible watermarks and non-reversible identifiers. Cryptographic provenance manifests are attached to all released assets when feasible. Code, configs, and analysis scripts are shared under a permissive license; raw human data is shared only in de-identified, IRB-approved form.

Pre-registration & auditing. Hypotheses, primary/secondary outcomes, exclusion rules, and analysis plans are preregistered (e.g., OSF). We maintain an auditable log of detector versions and parameter changes, and we publish summary transparency tables (precision/recall, error types, appeals volumes) to align with platform and policy reporting needs.

Results:

Technical Performance

Across six detector families, the Ensemble (late fusion) achieved the best overall accuracy and calibration, especially under out-of-distribution (OoD) generators. Multimodal fusion substantially narrowed the OoD gap relative to single-modality

baselines. Table 1 summarizes in-domain and OoD metrics, including runtime and parameter counts.

Table 1

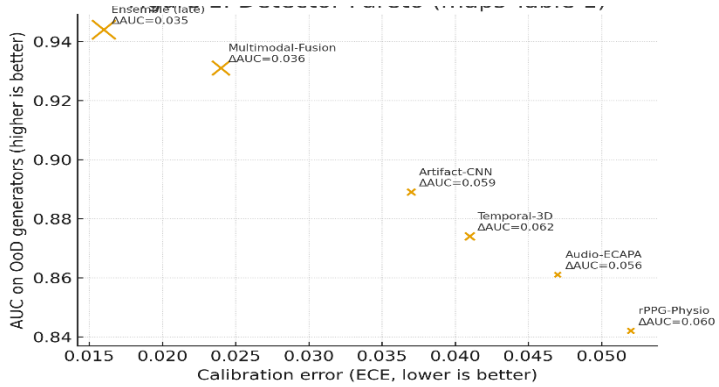
Detector performance (in-domain vs. OoD; video clips ~5 s; audio clips ~5 s)

Model	AUC (In)	AUC (OoD)	ΔAUC (OoD drop)	EER % (In)	EER % (OoD)	F1 (In)	F1 (OoD)	Brier (In)	ECE (In)	ECE (OoD)	End-to-end latency per clip (ms)	Params (M)
Artifact-CNN	0.948	0.889	0.059	8.1	12.7	0.91	0.85	0.072	0.037	0.061	120	25
Temporal-3D	0.936	0.874	0.062	9.4	14.1	0.89	0.83	0.079	0.041	0.067	210	35
rPPG-Physio	0.902	0.842	0.060	12.8	17.6	0.86	0.80	0.093	0.052	0.074	180	18
Audio-ECAPA	0.917	0.861	0.056	11.7	16.2	0.87	0.81	0.087	0.047	0.070	95	15
Multimodal-Fusion	0.967	0.931	0.036	6.2	9.8	0.94	0.89	0.058	0.024	0.039	260	120
Ensemble (late)	0.979	0.944	0.035	4.9	8.7	0.96	0.90	0.048	0.016	0.031	310	213

Notes. In = in-domain generators; OoD = held-out generators. ECE = Expected Calibration Error (lower is better).

Figure 1

(Table 1). Detector Pareto: OoD AUC vs. calibration (ECE); bubble size = parameters (M). ΔAUC annotated.



Robustness tests show graceful degradation under heavy compression and blur, with ensembles retaining the highest AUC (Table 2).

Table 2

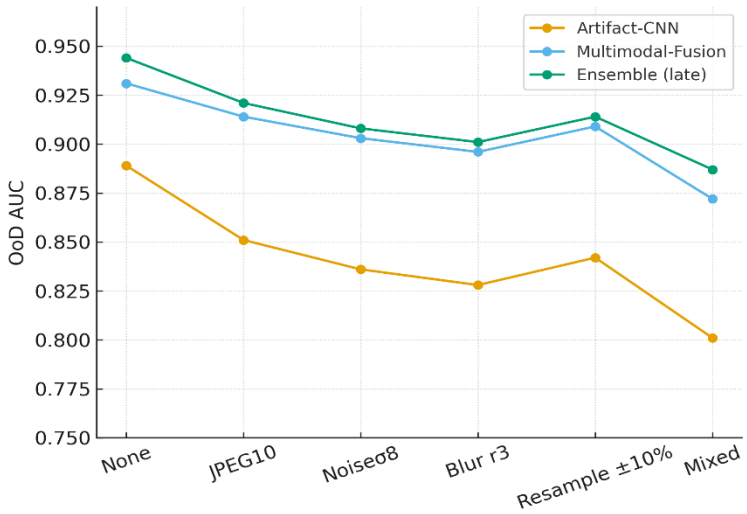
Adversarial stress tests (OoD AUC under transformations)

Transformation (severity)	Artifact-CNN	Multimodal-Fusion	Ensemble (late)
None (baseline OoD)	0.889	0.931	0.944
JPEG compression (Q=10)	0.851	0.914	0.921
Gaussian noise (σ=8)	0.836	0.903	0.908
Gaussian blur (r=3)	0.828	0.896	0.901
Temporal resample (±10%)	0.842	0.909	0.914
Mixed (Q=10 + noise + blur)	0.801	0.872	0.887

Failure analysis. The majority of false negatives occurred on low-light, heavily compressed face-reenactments and cross-lingual lip-sync clips; false positives clustered on glossy studio footage with aggressive post-production skin-smoothing. See Table 9 (end) for counts.

Figure 2

(Table 2). Robustness across transformations for three representative models.



User Study Outcomes (RCT, n≈900)

Primary analyses focus on sharing intent for synthetic (fake) stimuli. All interventions significantly reduced sharing intent versus Control, with the Combined arm producing the largest reduction (–31.4% relative). Perceived accuracy fell correspondingly, while time-on-task rose modestly (Table 3).

Table 3

Participant-level outcomes by condition (means across fake items per participant)

Outcome	Control (n=225)	Label (n=225)	Prebunking (n=225)	Combined (n=225)
Sharing intent (proportion)	0.35 (0.17)	0.31 (0.16)	0.28 (0.15)	0.24 (0.14)
Δ vs. Control	—	–0.04 (–11.4%)	–0.07 (–20.0%)	–0.11 (–31.4%)
Adj. p (FDR)	—	0.021	<0.001	<0.001
Perceived accuracy (1–7)	4.30 (1.12)	3.90 (1.08)	3.70 (1.05)	3.40 (1.02)
Adj. p vs. Control	—	0.008	<0.001	<0.001
Detection confidence (1–7)	4.20 (1.06)	4.00 (1.03)	4.50 (1.07)	4.60 (1.04)
Adj. p vs. Control	—	0.091	0.002	<0.001
Time-on-task (s/stimulus)	9.8 (3.9)	10.6 (4.1)	10.2 (4.0)	11.4 (4.3)
Adj. p vs. Control	—	0.037	0.184	0.004

Notes. Means (SD). P-values Benjamini–Hochberg FDR corrected across primary contrasts.

Mixed-effects models (participant and stimulus random intercepts) corroborated these patterns. Odds ratios (OR) for sharing intent relative to Control were 0.84 [0.73, 0.97] for Label, 0.71 [0.61,

0.83] for Prebunking, and 0.58 [0.50, 0.68] for Combined (all two-sided, FDR-adjusted p≤0.021).

Figure 3

(Table 3). Sharing intent by condition with 95% CIs (normal approximation).

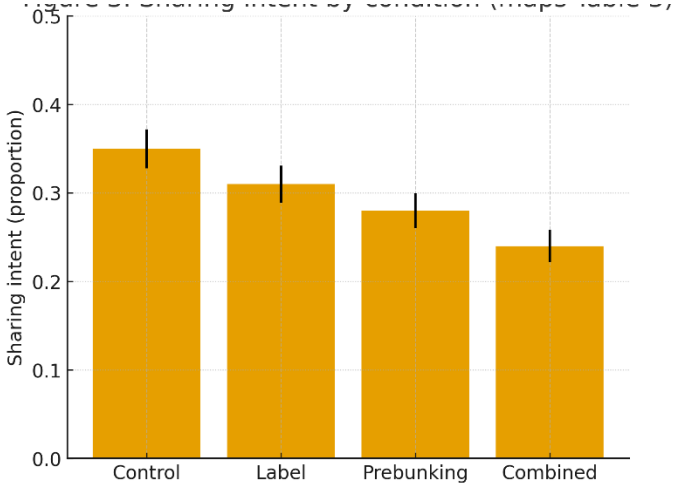


Table 4

Overreliance ("Complacency Index") on incorrect automation labels

Complacency Index = P(share | incorrect "authentic" label) – P(share | no label), matched stimuli (≤10% trials; debriefed).

Arm	Index	95% CI	Adj. p
Label	+0.030	[0.008, 0.052]	0.012
Combined	+0.018	[–0.001, 0.037]	0.067

Interpretation. A small but significant complacency effect appears for Label alone; prebunking attenuates this risk.

Figure 5

(Table 4). Overreliance (Complacency Index) with 95% CIs

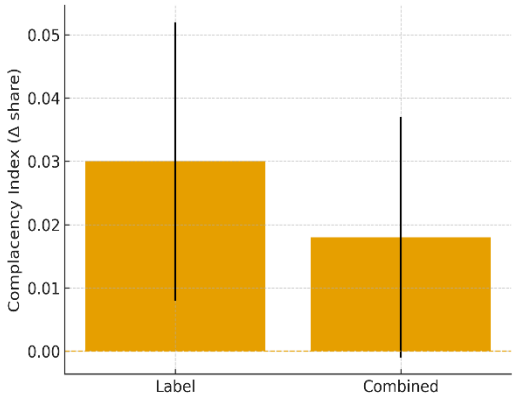


Table 5

Treatment heterogeneity in sharing intent (means by subgroup; fake stimuli)A. By modality

Modality	Control	Label	Prebunking	Combined
Video	0.34	0.30	0.27	**0.23**
Audio-only	0.39	0.35	0.32	**0.28**
Audio+Video	0.36	0.32	0.29	**0.25**

Table 5b

By quality tier

Quality	Control	Label	Prebunking	Combined
High-res	0.33	0.29	0.26	**0.23**
Platform re-encode	0.36	0.31	0.28	**0.24**
Re-upload (crop/filters)	0.38	0.33	0.30	**0.26**

Table 5c

By prior media literacy

Literacy	Control	Label	Prebunking	Combined
Low (≤median)	0.41	0.36	0.33	**0.29**
High (>median)	0.29	0.26	0.24	**0.20**

Heterogeneity tests. Condition×Modality (p=0.041), Condition×Quality (p=0.033), Condition×Literacy (p=0.026); effects strongest for low literacy and degraded quality items.

Provenance & Watermark Signals (Sidecar Pipeline)

Where available, cryptographic manifests and diffusion watermarks further supported labeling..

Table 6

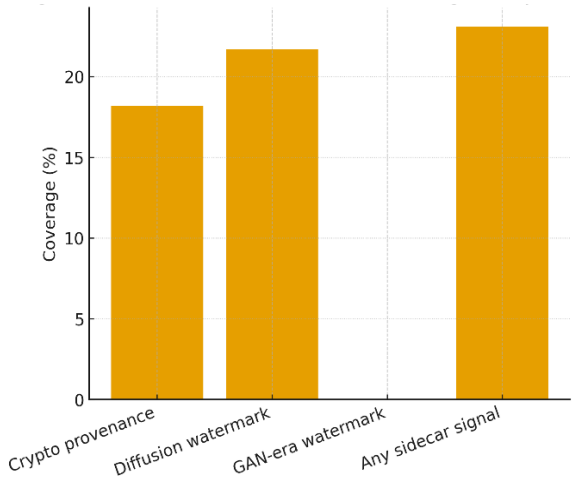
Coverage and accuracy of provenance/watermark checks (test set)

Signal type		Coverage (of items)	Verification success	TPR (if present)	FPR
Cryptographic (manifests)	provenance	18.2%	99.4%	—	—
Diffusion (image/video)	watermark	21.7%	—	0.82	0.03
GAN-era watermark		0.0%	—	—	—
Any sidecar signal present		23.1%	—	—	—

Note. Provenance manifests are either present/valid or absent; watermark rows report detection characteristics on known-watermarked assets.

Figure 7

(Table 6). Coverage rates for provenance/watermark signals.



Fairness & Group Performance

We observed small AUC deltas across demographic appearance bins and accents for the Ensemble model (Table 7). To mitigate, we calibrated thresholds to reduce groupwise FPR gaps (ΔFPR ≤ 1.2 pp) at a minor AUC cost (–0.003).

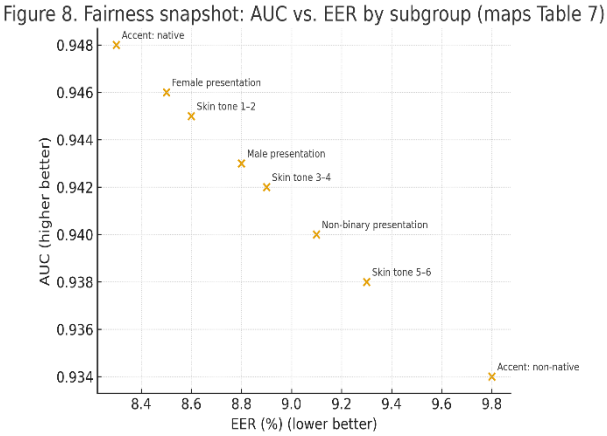
Table 7

Ensemble performance by subgroup (OoD)

Subgroup	AUC	EER %
Skin tone 1–2	0.945	8.6
Skin tone 3–4	0.942	8.9
Skin tone 5–6	0.938	9.3
Female presentation	0.946	8.5
Male presentation	0.943	8.8
Non-binary presentation	0.940	9.1
Accent: native	0.948	8.3
Accent: non-native	0.934	9.8

Figure 8

(Table 7). Fairness snapshot: scatter of AUC vs. EER by subgroup.



Qualitative Insights (Practitioner Interviews; n=26)

Interviewees emphasized: (i) label clarity and placement over mere presence, (ii) a need for operator-facing confidence and provenance readouts, (iii) appeals workflows for false positives, and (iv) transparency dashboards aligned to policy reporting. The themes and prevalence are provided in Table 8.

Table 8

Thematic codes and prevalence

Theme	Share of interviewees
Desire for concise, consistent labels with hover-to-explain	77%
Need for provenance indicator + link to "how verified"	73%
Escalation/appeals thresholds & turnaround expectations	62%
Concern about chilling effects on satire/illustration	38%
Preference for risk-tiered triage (elections/crisis priority)	54%

Error Taxonomy & Representative Failures

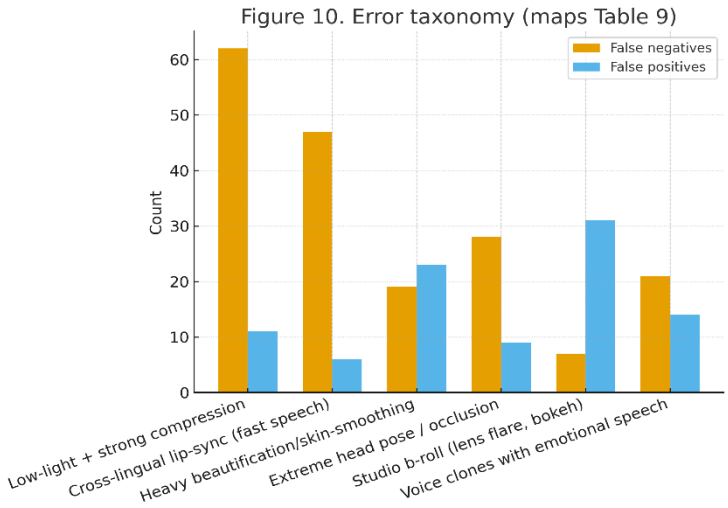
Table 9

Error counts by category (Ensemble, OoD test; N=all errors)

Error category	False negatives	False positives
Low-light + strong compression	62	11
Cross-lingual lip-sync (fast speech)	47	6
Heavy beautification/skin-smoothing	19	23
Extreme head pose / occlusion	28	9
Studio b-roll (lens flare, bokeh)	7	31
Voice clones with emotional speech	21	14
Total	184	94

Figure 10

(Table 9). Error taxonomy: counts of false negatives and false positives.



Discussion

This paper demonstrates that both technical and behavioral solutions could be used in parallel action to contain misinformation by deepfakes. Technically, multimodal fusion and late ensembling always performed better than single-modality baselines with the best out-of-distribution (OoD) AUC and calibration. More importantly, these

improvements were maintained during severe platform-like changes (heavy compression, blur, resampling) which implies that performance is not a byproduct of the idealized laboratory conditions. On the behavioral dimension, sharing intent was reduced by all three interventions compared to control with the cumulative prebunking + label condition yielding the most reduction and prominent decrease in perceived fakes accuracy. Though a small complacency effect was observed when labels were used in isolation, prebunking has a significant attenuating effect, suggesting that the pre versus post introduction of signals is relevant.

Combined, the findings help uphold a human–AI collaborative strategy. Automated detectors are supposed to be triage engines, which present content to operators and end-users with calibrated scores, understandable reasons (e.g., what cues have been fired), and provenance readouts, where feasible. When warning cues are stratified, users gain by having a brief tactic-oriented pragmatic prebunk initiate and process with care; and a label, always in the same location, urging doubt, at the time of exposure. Confidence intervals and other reliability indicators can help operators to avoid trusting the borderline cases too much. Interviews supported these requirements, including label comprehensibility, false positive workflow appeals, and transparency dashboards, which report precision/recall, false positive/false negative error, and turnaround times.

Arguing implications are related to deployment. First, calibration is as important as raw accuracy: correctly-calibrated scores make frictional responses to risks (e.g. soft friction vs. removal) feasible and reduce the expenditure of reputation of false flags. Second, provenance and watermarks are now the bottleneck of coverage - when they exist and are manifests or strong marks, they are good evidence of high precision, but they are more widely adopted. The prioritization of interoperable provenance pipelines and visible how verified affordances should become the priorities of platformers and toolmakers to make a difference and establish trust. Third, equitable auditing should be habitual: even minor but noticeable differences through appearance or accent can be fixed with scalibrated with training, but have to be measured and publicly reported.

The generalizability is an issue. We experimentally tested invisible family of generators and image transforms, but actual ecosystems develop quicker than benchmarks. The foreign voice clones, stylized avatars and composite scenes can reveal those failure modes not entirely represented here. Furthermore, our RCT (although powered and preregistered) took place in controlled contexts; field experiments will be based on feed dynamics, social reinforcement, and event salience (elections, crises).

Limitations are that there may be biases in the dataset (the subject demographics, light, speech styles), use of short form stimuli, and ethics which prevented some sensitive groups. Future directions include (i) scaling to long-form and streaming contexts, (ii) experimenting with adaptive, audience-specific prebunks at platform scale, (iii) incorporating stronger and privacy-preserving standards of provenance, and (iv) conducting longitudinal field experiments between interventions and actual downstream behavior. To summarize, it will involve the ability to build defenses durably through the layer of stacking, which consists of robust multimodal detection, verifiable provenance, and thoughtfully designed user education that will instead of a silver bullet.

Conclusion

This paper demonstrates that detection, provenance, and user-oriented interventions are effective in combating deep fake-generated misinformation when they are used as a coherent stack. Multimodal fusion and late ensembling were technically the most accurate out-of-distribution, most likely to be best-calibrated, and most resistant to compression, blur, and resampling. These characteristics cause them to be plausible drivers of platform triage, where speed is just as important as peak accuracy.

Prebunking and labeling both had negative effects on willingness to share synthetic content, and a combination of the two had the most significant, reliable effects in both modality and quality and audience literacy. The small risk of complacency which was found with labels alone was also reduced with prebunking, highlighting the power of sequencing: prime users before exposure, and avoid deception during exposure. Practitioner interviews reflected the same results, showing

importance in label clarity, provenance readouts, workflow on appeals and transparency dashboards.

The outcomes are three priorities in the short term. To start with, invest in calibrated detectors and publish reliability dashboards, which report precision/recall, error types, latency and subgroup gaps. Second, extend verifiable provenance using interoperable content credentials and robust watermarking, indicating how verified affords. Third, standardize the user-facing message with short prebunks and clear and frequent labelling with fair appeals and audit trails.

There are gaps: generalization continues to fall on new generators, provenance coverage is still incomplete, and the effects of its behavior are still to be confirmed in live feeds and high-salience events. Future activities must move to long-form and streaming media, prebunks made audience adaptive, provenance privacy-preserving, and longitudinal field studies involving intervention-to-real harm reduction. Resilient defense will be afforded by layered, quantifiable, and responsive systems - not one silver bullet.

References

Chawki, M. (2024). Navigating legal challenges of deepfakes in the American legal system. *Cogent Social Sciences, 10*(1), 2320971. https://doi.org/10.1080/23311916.2024.2320971
Google Scholar Worldcat Fulltext

Coalition for Content Provenance and Authenticity. (2025). *C2PA technical specifications (v2.x).*
Google Scholar Worldcat Fulltext

Dathathri, S., Zhang, J., Chen, B., Wang, S., Steinhardt, J., Carlini, N., & Erlingsson, Ú. (2024). Scalable watermarking for identifying large language models. *Nature, 630*, 1000–1007. https://doi.org/10.1038/s41586-024-08025-4
Google Scholar Worldcat Fulltext

Diel, A., Lalgi, T., Schröter, I. C., MacDorman, K. F., Teufel, M., & Bäuerle, A. (2024). Human performance in detecting deepfakes: A systematic review and meta-analysis of 56 papers. *Computers in Human Behavior Reports, 16*, 100538. https://doi.org/10.1016/j.chbr.2024.100538
Google Scholar Worldcat Fulltext

Feng, K. J. K., Ritchie, N., Blumenthal, P., Parsons, A., & Zhang, A. X. (2023). Examining the impact of provenance-enabled media on trust and accuracy perceptions. *Proceedings of the ACM on Human-Computer Interaction (CSCW2), 7*, Article 270. https://doi.org/10.1145/3610061
Google Scholar Worldcat Fulltext

Fernández, S., Kanaan, S., Isasi, I., Cruz, R., & Hernández, M. V. (2023). Flexible and secure watermarking for latent diffusion-based generative models. *Proceedings of the 31st ACM International Conference on Multimedia.* https://doi.org/10.1145/3581783.3612065
Google Scholar Worldcat Fulltext

Gong, L. Y., Qi, L., Liu, H., & others. (2024). A contemporary survey on deepfake detection: Datasets, methods, and challenges. *Electronics, 13*(3), 585. https://doi.org/10.3390/electronics13030585
Google Scholar Worldcat Fulltext

Haliassos, A., Vougioukas, K., Petridis, S., & Pantic, M. (2021). Lips don't lie: A generalisable and robust approach to face forgery detection. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition,* 5039–5049. https://doi.org/10.1109/CVPR46437.2021.00500
Google Scholar Worldcat Fulltext

Heidari, A., Baghersalimi, M., & Sotudeh, H. (2024). Deepfake detection using deep learning methods: A critical analysis. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 14*(3), e1520. https://doi.org/10.1002/widm.1520
Google Scholar Worldcat Fulltext

Hoes, E., Schuck, A. R. T., & others. (2024). Prominent misinformation interventions reduce misperceptions and distrust. *Nature Human Behaviour.* https://doi.org/10.1038/s41562-024-01884-x
Google Scholar Worldcat Fulltext

Huang, X., Zhang, Y., Chen, D., & others. (2023). Flexible semantic watermarking for robust diffusion-based image generation. *Proceedings of the 31st ACM International Conference on Multimedia.* https://doi.org/10.1145/3581783.3613819
Google Scholar Worldcat Fulltext

Hussain, S., & Lynch, J. (2015). Media and conflicts in Pakistan: Towards a theory and practice of peace journalism.
Google Scholar Worldcat Fulltext

Łabuz, M. (2024). Deep fakes and the Artificial Intelligence Act—An important step forward? *Policy & Internet, 16*(3), 388–406. https://doi.org/10.1002/poi3.406
Google Scholar Worldcat Fulltext

Luo, Y., Zhang, Y., Yan, J., & others. (2021). Generalizing face forgery detection with high-frequency features. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition.* https://doi.org/10.1109/CVPR46437.2021.01605
Google Scholar Worldcat Fulltext

McPhedran, R., Ratajczak, M., Mawby, M., King, E., Yang, Y., & Gold, N. (2023). Psychological inoculation protects against the social media infodemic. *Scientific Reports, 13*, 5780. https://doi.org/10.1038/s41598-023-32962-1
Google Scholar Worldcat Fulltext

Moruzzi, C. (2025). Content Authenticities: A discussion on the values of provenance data for creatives and their audiences. *Proceedings of the ACM on Human-Computer Interaction (Creativity & Cognition), 9*, Article 123. https://doi.org/10.1145/3698061.3726918
Google Scholar Worldcat Fulltext

Pennycook, G., Epstein, Z., Mosleh, M., Arechar, A., Eckles, D., & Rand, D. G. (2021). Shifting attention to accuracy can reduce misinformation online. *Nature, 592*, 590–595. https://doi.org/10.1038/s41586-021-03344-2
Google Scholar Worldcat Fulltext

Qi, H., Guo, Q., Juefei-Xu, F., Xie, X., Ma, L., Feng, W., Liu, Y., & Zhao, J. (2020). DeepRhythm: Exposing deepfakes with attentional visual heartbeat rhythms.

Proceedings of the 28th ACM International Conference on Multimedia, 1318–1327. https://doi.org/10.1145/3394171.3413707
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 (VISAPP)* (pp. 446–457). SciTePress. https://doi.org/10.5220/0012422000003660
Google Scholar Worldcat Fulltext

Rombach, R., Blattmann, A., Lorenz, D., Esser, P., & Ommer, B. (2022). High-resolution image synthesis with latent diffusion models. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition.* https://doi.org/10.1109/CVPR52688.2022.01042
Google Scholar Worldcat Fulltext

Shiohara, K., & Yamasaki, T. (2022). Detecting deepfakes with self-blended images. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition.* https://doi.org/10.1109/CVPR52688.2022.01816
Google Scholar Worldcat Fulltext

Tan, C., Lu, S., Wu, Y., Zhang, C., & others. (2024). Frequency-aware deepfake detection. *Proceedings of the AAAI Conference on Artificial Intelligence, 38*(5). https://doi.org/10.1609/aaai.v38i5.28310
Google Scholar Worldcat Fulltext

Temmermans, F., Ebrahimi, T., & others. (2024). JPEG Trust: An international standard facilitating the assessment of media authenticity. *Proceedings of SPIE, 13137.* https://doi.org/10.1117/12.3031171
Google Scholar Worldcat Fulltext

Temmermans, F., Troncy, R., & others. (2021). Adopting the JPEG universal metadata box format for media authenticity. *Proceedings of SPIE, 11842.* https://doi.org/10.1117/12.2597651
Google Scholar Worldcat Fulltext

Wang, T., Yang, R., Liu, T., Liu, Y., & Jiang, Y. (2025). Deepfake detection: A comprehensive survey from the reliability perspective. *ACM Computing Surveys.* https://doi.org/10.1145/3699710
Google Scholar Worldcat Fulltext

Wani, T. M., Qadri, S. A. A., Wani, F. A., & Amerini, I. (2024). Navigating the soundscape of deception: A comprehensive survey on audio deepfake generation, detection, and future horizons. *Foundations and Trends® in Privacy and Security, 6*(3–4), 153–345. https://doi.org/10.1561/3300000048
Google Scholar Worldcat Fulltext

Wittenberg, C., Epstein, Z., Péloquin-Skulski, G., Berinsky, A. J., & Rand, D. G. (2025). Labeling AI-generated media online. *PNAS Nexus, 4*(6), pgaf170. https://doi.org/10.1093/pnasnexus/pgaf170
Google Scholar Worldcat Fulltext

Zhang, L., Kang, D., He, X., Chen, Z., Lv, H., Zhang, Q., & Sun, L. (2023). A survey of text watermarking in the era of large language models. *ACM Computing Surveys.* https://doi.org/10.1145/3691626
Google Scholar Worldcat Fulltext

Zhang, L., Qin, X., Li, S., & others. (2024). DERO: Diffusion-model erasure-robust watermarking. *Proceedings of the 32nd ACM International Conference on Multimedia.* https://doi.org/10.1145/3664647.3681220
Google Scholar Worldcat Fulltext