Scientific Research Publishing

Open Journal of Leadership, 2025, 14(3), 403-418
https://www.scirp.org/journal/ojl
ISSN Online: 2167-7751
ISSN Print: 2167-7743

Leading through the Synthetic Media Era: Platform Governance to Curb AI-Generated Fake News, Protect the Public, and Preserve Trust

Prajkta Waditwar

Strategic Sourcing, Box, Inc (Independent Research), Redwood City, CA, USA
Email: prajkta.waditwar@gmail.com

How to cite this paper: Waditwar, P. (2025). Leading through the Synthetic Media Era: Platform Governance to Curb AI-Generated Fake News, Protect the Public, and Preserve Trust. *Open Journal of Leadership, 14,* 403-418.
https://doi.org/10.4236/ojl.2025.143020

Received: August 20, 2025
Accepted: September 19, 2025
Published: September 22, 2025

Abstract

Generative AI has collapsed the cost of fabricating persuasive audio, images, and video; social platforms can then amplify these forgeries to millions within minutes. Ordinary users—not only public figures—now face wire-transfer fraud via deepfaked executives, non-consensual intimate imagery, cloned-voice "kidnapping" scams, celebrity-death hoaxes, health/finance misinformation, and market-moving fake photos. Beyond immediate losses, reputational abuse and automation-related job loss correlate with anxiety, depression, and, for some, suicidal ideation. We argue the central risk is not "AI" itself but the low-friction spread of forged sight/sound cues in engagement-optimized feeds. We propose a risk-tiered, authenticate-then-distribute regime anchored by a Pre-Publication Authenticity Verification (PPAV) pipeline that combines provenance (C2PA/Content Credentials), watermark signals, media forensics, similarity/history checks, and semantic claim verification. We add a governance blueprint (policy, operations, metrics), a quarter-by-quarter implementation roadmap, and a victim-first model with rapid takedowns and mental-health-aware UX.

Keywords

Artificial Intelligence, Synthetic Media, Deepfakes, Content Authenticity, C2PA, Watermarking, Perceptual Hashing, Online Safety, Mental Health, Risk Management

1. Introduction

Social platforms are the default layer for news, entertainment, and private com-

munication. Consumer-grade models now synthesize photorealistic faces, lip-synced video, and human-sounding voices from seconds of source material. Injected into ranking systems, falsehoods routinely outpace corrections, overwhelming users' ability to separate fact from fabrication. The harms are concrete for everyday people: finance staff authorize fraudulent wires after deepfaked meetings; parents receive cloned-voice ransom calls; individuals discover viral non-consensual intimate imagery; AI-styled memorial graphics falsely announce celebrity deaths; and highly realistic "breaking-news" images briefly move markets. These incidents cause financial loss, reputational injury, and psychological distress. Evidence links online victimization to anxiety/depression and links unemployment/financial stress to elevated suicide risk. Platforms therefore need safeguards that address both authenticity and human impact.

This paper: 1) Maps the threat landscape; 2) Formalizes a risk taxonomy; 3) Proposes authenticate-then-distribute with PPAV screening; 4) Details governance (policy/ops/metrics); 5) Provides an implementation roadmap; 6) Adds evaluation, legal, and consent frameworks.

Taxonomy provenance. Our R1-R5 harm taxonomy (identity/consent, financial, information, psychological, systemic) diverges from "process-first" frameworks (e.g., NIST AI RMF's cross-cutting risk characteristics and the EU DSA's "systemic risks") by prioritizing *harm class at content level* to drive *class-specific gates, SLAs, and metrics*. NIST AI RMF organizes risk management functions and characteristics (e.g., safety, privacy, explainability) (Raimondo et al., 2023), while the DSA mandates platform-wide assessment/mitigation of systemic risks (e.g., illegal content, disinformation, minors' safety, fundamental rights). Our taxonomy complements these by mapping concrete synthetic-media incidents to controls that can be executed at upload time.

2. The Social-Media Threat Landscape

Creation costs for convincing fakes are near zero: seconds of audio can clone a voice; a handful of photos can synthesize a face; off-the-shelf apps lip-sync or restyle video. Distribution is frictionless—upload flows accept realistic media; recommenders can propel a post to mass reach in minutes. Because virality precedes verification, time-to-harm is short; even fast removals cannot retract screenshots, downloads, and mirrors. Users are exposed because traditional trust cues are forgeable (familiar face/voice), context collapse places high-risk claims amid entertainment, and incentives are asymmetric (attackers need one win; targets must detect every attempt). Authenticity checks must therefore shift before distribution.

3. Real-World Damage Caselets

Across platforms, synthetic media has already produced concrete harms for ordinary people and organizations. In one widely reported case, a finance employee

joined what appeared to be a routine multi-participant video conference with senior executives and, acting on urgent instructions, authorized wire transfers totaling roughly US$25 million; only later did the firm discover the entire "boardroom" was a deepfake (Internet Crime Complaint Center, 2024). Similarly, an accounts-payable controller at another company sent about US$243,000 after receiving a phone call that flawlessly mimicked the CEO's voice, accent, and cadence (Internet Crime Complaint Center, 2024; Reshef, 2023). Beyond corporate fraud, non-consensual sexual deepfakes have targeted celebrities and private citizens—including schoolgirls—causing acute distress, bullying, and long-tail reputational damage (Internet Crime Complaint Center, 2024; Ortutay, 2023). Synthetic "breaking-news" images have briefly moved markets, as when a photorealistic picture of an explosion near a federal building spread widely before being debunked. (Wang et al., 2024). Deepfake advertising has also co-opted public figures' likenesses to pitch dubious products, such as fake "$2 phone" giveaways, deceiving consumers and tarnishing reputations (Reporter, 2023; Pringle, 2023). Families have been hit by kidnapping scams in which parents receive calls using a cloned voice of a loved one demanding immediate payment. Even when no money changes hands, AI-generated memorial graphics and auto-written obituaries have falsely announced the deaths of popular actors and TV hosts, spreading panic while driving monetizable clicks.

These incidents point to specific controls leaders should institute. High-value payments must never be approved within a single channel; require **out-of-band verification** (e.g., a call back to a known number or a second approver on a different medium) and liveness checks for urgent requests (Qi et al., 2020). Treat any **likeness-based advertising** as high risk and demand verifiable consent and asset provenance before an ad can run. Classify **celebrity-death claims** as high-risk information harm and hold such uploads until authenticity can be corroborated by trusted sources; suppress recommendations when provenance is absent. For **crisis-language P2P transfers** (e.g., ransom or emergency pleas), add friction such as cool-off timers and secondary confirmations. Finally, adopt **one-click privacy takedowns** for intimate imagery and **hash-blocking** to prevent re-uploads, coupled with fast SLAs so victims receive timely relief (Farid, 2021).

Where possible, we anchor impersonation, voice-cloning, and deceptive-ad risks in law-review and government material (keeping the original news reports as supplementary footnotes). The FBI has issued formal PSAs documenting AI-enabled impersonation and fraud trends and operational mitigations; recent law-review work examines deepfake exploitation and the right of publicity in ads and endorsements (Preminger & Kugler, 2024; Murray, 2025).

High-value payments require out-of-band verification and liveness; likeness-based advertising requires verifiable consent and provenance; celebrity-death claims are gated until corroborated; crisis-language P2P transfers gain cool-off timers; intimate imagery routes to one-click privacy takedown and hash-blocking (Controls detailed in **Table 1**).

4. Risk Taxonomy for Platforms

We group synthetic-media harms into five classes:

R1—Identity & consent harm covers any simulation of a real person without permission—most painfully, non-consensual sexual imagery and impersonations used in ads or outreach. The injury here is personal and immediate: dignity, privacy, safety, and livelihood can be damaged the moment a look-alike face or voice circulates.

R2—Financial harm captures direct monetary losses when synthetic media is used to deceive—wire-transfer fraud after a deepfaked "executive" meeting, voice-clone phone calls authorizing payments, or fake celebrity endorsements that trick users into buying scams or handing over credentials.

R3—Information harm arises when realistic fakes mislead at scale: fabricated health or finance claims, market-moving "breaking-news" images, or celebrity-death hoaxes that cause panic and poor decisions before corrections can catch up.

R4—Psychological harm focuses on the human aftermath—anxiety, depression, and in some cases suicidal ideation—following reputational attacks, image-based abuse, or mass harassment, with adolescents and other vulnerable groups at higher risk.

R5—Systemic harm reflects the broader erosion of trust in authentic user-generated content and the creator economy; when users cannot tell real from fake, engagement quality drops, advertisers pull back, and the platform's legitimacy suffers.

R1 - R5 is intentionally *incident-centric*: it optimizes for fast, class-specific actions at upload time, whereas NIST and the DSA are *system-centric* (governance, process, and systemic risk). The two layers interlock: platform-level duties (DSA/NIST) require class-level controls (R1 - R5) to be measurable and auditable.

This taxonomy matters because each risk prompts a different response: R1 needs consent gates, one-click takedowns, and hash-blocking; R2 requires out-of-band payment verification and fraud-aware friction; R3 calls for authenticate-then-distribute policies and provenance-aware ranking; R4 adds victim-first operations and mental-health signposting; and R5 demands transparent reporting and independent audits to rebuild trust. Controls should be class-specific rather than treating all synthetic media as uniform risk.

5. Principle: Authenticate-Then-Distribute

Under this model, distribution is earned—not assumed.

The baseline rule is simple: *no provenance, no promotion*. If an upload has no verifiable origin, the platform may still allow it to exist (e.g., on the uploader's profile or via direct link), but it does not enter recommendation systems and it must carry a clear context panel stating that its *origin or authenticity is unverified*.

For high-risk classes—identity/consent harms, financial harms, and information harms (R1 - R3)—the bar is higher: *no provenance, no posting to public*

feeds. Before such content can be publicly distributed, the uploader must satisfy one of two conditions.

Table 1. Risk taxonomy → controls, thresholds, and SLAs.

Risk	Examples	Mandatory Controls	Thresholds/Actions	Reviewer Queue & SLA
R1 Identity & Consent	Non-consensual intimate imagery; likeness-based ads/impersonation	Verifiable consent artifact; one-click privacy takedown; hash-blocking (PDQ/PhotoDNA; TMK+PDQF-style; audio FP)	High risk by default; quarantine unless consent proven; labels for allowed parody/satire	Privacy queue, P95 < 4 h takedown; re-upload block
R2 Financial	Deepfake payment requests; fake endorsements; investment hoaxes	Out-of-band payment verification; provenance-aware ranking; ad consent proof	No promotion without provenance; quarantine if claim involves payment instruction	Fraud queue, P95 < 24 h
R3 Information	Health/finance misinformation; breaking-news/celebrity-death hoaxes	Semantic corroboration against authoritative sources; labels; context panels	No promotion pending verification; reject if fabricated	News/Info queue, P95 < 24 h
R4 Psychological	Mass harassment; reputation attacks	Rate-limits on reposts; MH signposting; reporting tools	Rapid throttling of brigading; contextual prompts	Abuse queue, P95 < 24 h
R5 Systemic	Erosion of trust in UGC/creators	Transparency reports; audits; creator appeals	Publish metrics; annual audit	Policy review, quarterly reporting

Either the asset includes cryptographic provenance (e.g., C2PA/Content Credentials) that attests to its capture and edit chain, or the uploader passes the verified identity checks and marks the upload as synthetic with an explicit disclosure and prominent label.

Even then, the item is quarantined for review or reach-constrained until checks clear.

In practice, this shifts platforms from a publish-then-moderate posture to an assure-then-amplify posture, reducing the chance that high-risk fakes go viral before anyone can intervene. Items may be reach-limited or queued for rapid review. Rejections must be "reject with rationale + appeal link".

For contexts where pre-publication screening is technically limited (E2E messaging, low-latency live, or ephemeral "Stories"), apply recipient-side warnings, share-rate limits, forward limits, and contextual interstitials; allow optional "verify before send" for business accounts; and run PPAV asynchronously with retroactive throttling or takedown if risk thresholds are exceeded.

6. Governance Blueprint for Meta/Instagram/Youtube/Google

6.1. Technical & Product Controls

Make provenance the default. Every upload should be checked for a valid C2PA/Content Credentials manifest at the point of ingestion; when present, expose a "Content Credentials" panel so viewers can see how and where the asset was captured and edited. Pair this with a **detector ensemble** that fuses multiple signals—creator disclosures, watermark detectors from major vendors, classic and modern media forensics (e.g., CFA/demosaicing, resampling, temporal coherence), and robust perceptual hashing (Farid, 2021). If these signals disagree, the system should automatically quarantine the item for rapid review rather than allowing it to enter recommendations. Build **consent and identity gates** around simulations of a real person's face or voice: distribution only proceeds after verified consent is supplied; victims must have a one-click privacy takedown that triggers hash-blocking to prevent reuploads. Make ranking **provenance-aware** by downranking realistic media with unknown origin, boosting items with strong credentials, and adding share-time friction (confirmation prompts, send limits) during virality spikes. Finally, **harden the ad stack**: any likeness-based endorsement must carry both consent proof and provenance; otherwise the ad is automatically rejected and the advertiser reviewed.

Treat valid Content Credentials as *strong positive provenance*. Adoption continues to grow across the ecosystem (Adobe's Content Credentials, camera integrations, and new C2PA members), so provenance coverage is expected to rise; expose a public "Content Credentials" panel.

6.2. Policy & Enforcement

Publish bright-line rules that ban the highest-harm behaviors, including non-consensual sexual deepfakes and impersonation for financial gain, and enforce them uniformly across products. Operate with a **victim-first posture**: staff a 24/7 triage function, guarantee sub-four-hour service levels for intimate imagery, and provide a live victim dashboard that shows case status, actions taken, and reupload blocks in effect. Deter repeat abuse with **graduated penalties**—cross-product strikes that carry to sister apps, API throttling for abusive automation, and payments off-ramps that disable monetization and ad credits when an account is tied to synthetic-media harm.

6.3. People & Process

Stand up a dedicated **Synthetic Media Operations** team that brings safety engineering, policy, legal, and communications under one roof and runs from playbooks tailored to the R1-R3 risks (identity/consent, financial, and information harms). Treat preparedness like security: run regular **deepfake drills** so moderators, trust leads, and treasury/AP staff can practice voice-clone fraud responses, celebrity-death-hoax workflows, and privacy-takedown escalations. Each drill

should end with a blameless post-mortem and concrete fixes to policy text, reviewer tools, and on-call rotations.

6.4. Executive Metrics

Leadership should review a concise integrity dashboard each week. The core timings are **time-to-label** and **time-to-takedown**, tracked at median (P50) and tail (P95) so slow cases cannot hide. Measure **provenance coverage** as the percentage of recommended watchtime attributable to assets with valid Content Credentials; aim to grow this steadily. Track **reupload recidivism** after hash-blocking—healthy systems drive this toward zero—and monitor **victim resolution time** from first report to final suppression. For advertising, report **likeness-ad consent compliance** (what share of such ads shipped with verified consent and provenance). These metrics should inform quarterly integrity reports and tie directly to executive objectives so safety outcomes are owned at the top.

7. Pre-Publication Authenticity Verification (PPAV) Model

Goal. The PPAV model is designed to stop high-risk synthetic media from entering public feeds **before** authenticity and consent are established, while letting ordinary creative posts flow with minimal friction. It inserts a short, automated screening pipeline at upload, then routes only uncertain or sensitive items for fast human review.

7.1. Layered Architecture (Defense-in-Depth)

At ingest, the platform first performs **provenance verification**. If the file carries valid C2PA/Content Credentials, the system confirms the capture/edit chain and attaches a visible "Content Credentials" panel; in high-risk classes, malformed or forged manifests cause the upload to fail closed. In parallel, a **watermark scan** looks for vendor watermarks using cross-modal detectors; presence or absence is treated as a **signal**, not proof, about synthetic origin. The third layer runs **media forensics**: for images/video it examines demosaicing/CFA consistency, resampling and error-level artifacts, frequency spectra, diffusion/GAN fingerprints, lighting and eye-gaze coherence; for audio it checks spectral/phase patterns and prosody continuity; optional liveness cues such as remote pulse signals can be used for faces. Next, **similarity and history** checks use robust perceptual/DNN hashing to compare the asset to prior uploads and trusted archives, graphing suspicious re-use so that light edits or crops still match. Finally, a **semantic claim and source check** extracts any salient claim from the caption or overlays—e.g., a celebrity death, a miracle health cure, or a guaranteed financial return—links the entities to a knowledge base, and looks for corroboration from authoritative sources; missing or contradictory corroboration triggers escalation. Signals from all layers, along with contextual features such as uploader history, velocity, and geography, feed a **risk aggregation** function that produces a composite score and triggers policy actions.

In risk aggregation, "uploader trust" is a bounded variable combining: a) **Ac-**

count integrity (age, confirmed email/phone, 2FA, verified ID where applicable); b) **Policy history** (strike-free days, appeals upheld, prior privacy/takedown events); c) **Provenance history** (share of past posts with valid C2PA); d) **Behavioral signals** (automated posting patterns, coordinated-inauthentic-behavior matches). The composite is calibrated so policy decisions never rely *solely* on trust—high-risk claims still require provenance/corroboration.

Policy actions. Low-risk items publish immediately, remain eligible for recommendations, and display credentials when available. Medium-risk items still publish but are **not promoted** and carry a context panel noting that authenticity is unverified. High-risk items are **quarantined** for rapid human review; the uploader may be required to complete identity verification and apply explicit "synthetic" labels, or the item may be blocked entirely—especially when a real person's likeness is used without consent, in which case it routes to privacy takedown.

7.2. Example Flow: "Celebrity X Has Died" Upload

Suppose a memorial graphic is uploaded. The file arrives **without** a C2PA manifest, so provenance is missing. No watermark is detected, so the watermark layer is neutral. Forensics find diffusion-style fingerprints and inconsistent EXIF data, increasing risk. Similarity checks show a near-match to an old press photo, again raising risk. The semantic layer cannot find any authoritative obituary while the celebrity's official accounts remain active, pushing risk higher. The aggregated score crosses the high-risk threshold: the post is held for review (or rejected). If later allowed, it is published **without promotion** and with a conspicuous warning until verification is complete.

7.3. Engineering and Operations Considerations

To keep user experience fast, the first three layers—provenance, watermark, and similarity—should run cheaply **in parallel**; the heavier forensics and semantic retrieval can gate only when the content matches known high-risk classes or exhibits unusual velocity. Reviewers need a purpose-built console showing the provenance panel, forensic heatmaps, retrieval hits, and the prior-upload graph side-by-side to speed decisions. Quality management requires class-specific thresholds (stricter for R1 identity/consent harm than for R3 information harm), continuous tracking of false positives/negatives, and red-team corpora for regression testing. Respect privacy by minimizing PII retention, documenting model cards, and auditing subgroup error rates for any biometric or liveness cues. Finally, build adversarial resilience by rotating forensic features, ensembling multiple detectors, and incorporating behavioral signals such as sock-puppet networks or suspicious payment trails.

The cheapest PPAV layers (provenance header checks, perceptual hashing, and watermark probes) are CPU-friendly and already deployed at industry scale (e.g., PDQ/TMK + PDQF for image/video hashing). Benchmarks and vendor documentation indicate these systems are designed for high throughput on commodity hardware; modern evaluations compare algorithm accuracy/recall and describe

CPU-only deployments. This supports a design in which P/W/S layers run for *all* uploads, with heavier forensics/semantic checks triggered only for suspected high-risk content or velocity spikes.

7.4. PPAV Flow Diagram

P-Layer (Provenance). Validate C2PA/Content Credentials; display a public panel. If a manifest is malformed/forged on high-risk content, reject with rationale + appeal.

W-Layer (Watermarks). Detect vendor watermarks; treat as probabilistic signals, not gates.

F-Layer (Media Forensics). Image/video: demosaicing/CFA consistency, resampling/ELA, frequency spectra, diffusion/GAN fingerprints, lighting/eye-gaze/temporal coherence. Audio: spectral/phase anomalies, prosody continuity. Liveness checks (when a user opts in) may use remote PPG or blink/pose dynamics; results are privacy-protected and auditable.

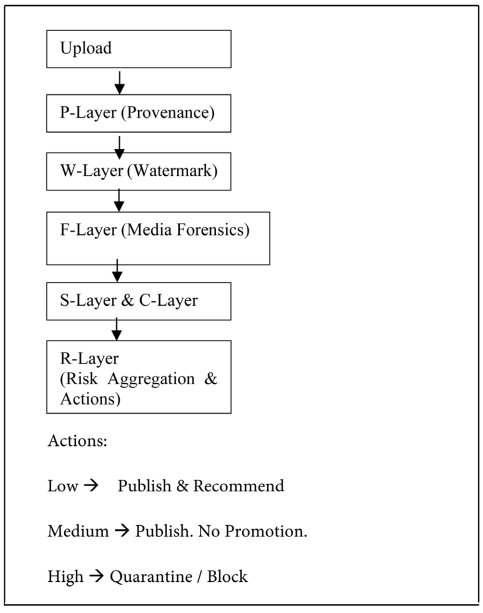
S-Layer (Similarity & History). Perceptual/DNN hashing against prior uploads and trusted archives; graph suspicious asset reuse.

C-Layer (Semantic Claims). Extract claims from captions/overlays (e.g., "<Person> has died", "miracle cure", "guaranteed 10× crypto") and seek corroboration from authoritative sources via a curated whitelist. When sources conflict or are missing, degrade gracefully (no promotion + context).

R-Layer (Risk Aggregation). Combine all signals with context (uploader trust, velocity, geography) to produce a composite risk and trigger actions.

Latency Budget (Typical): P/W/S in parallel ≤80 ms aggregate; F/C on suspected high-risk ≤500 ms; overall ≤1 s pass-through, with quarantine for the top risk decile.

Pre-Publication Authenticity Verification (PPAV) model:



7.5. Algorithm 1—Risk Aggregation (Pseudocode)

This code explains how the system looks at many clues about a post—like whether it has proof of origin, watermarks, forensic signs of editing, matches to old content, or bold claims—and then combines all that into one risk score. Based on that score, the system decides if the post can be shown normally, shown with limits and warnings, or held back for human review. High-risk cases, like using someone's face without consent, get blocked immediately. This way, platforms can stop dangerous fakes from spreading too fast while still letting safe content flow smoothly.

```
Inputs: x = {
  prov ∈ {0,1}, prov_quality ∈ [0,1],
  wm_score ∈ [0,1],
  forensics_score ∈ [0,1],
  sim_near_dupe ∈ ℕ, sim_reuse_graph ∈ [0,1],
  claim_class ∈ {none, death, health, finance, other}, claim_conf ∈ [0,1],
  uploader_trust ∈ [0,1], velocity ∈ [0,1], geo_risk ∈ [0,1]
}

# Feature normalization
z = [
  prov, prov_quality, wm_score, forensics_score,
  min(sim_near_dupe/5,1), sim_reuse_graph,
  onehots(claim_class), claim_conf,
  uploader_trust, velocity, geo_risk
]

# Class-specific weights (trained on labeled R1–R3 corpora)
R = sigmoid( w_class · z + b_class )

Policy:
if R < τ_low:        publish + eligible for recommendation
elif R < τ_high:     publish; no promotion; add "authenticity unverified" panel
else:                quarantine; require ID + explicit synthetic label or block
```

7.6. Consent, Identity, and Appeals

Verifiable consent artifact. Likeness-based uploads/ads require a cryptographically signed grant from the rightsholder (or authorized agent/estate), with selective disclosure (recipient platform, scope, duration, revocation endpoint).

Minors & vulnerable persons. Require parental/guardian consent; default to deny if provenance/consent are ambiguous.

Parody, satire, newsworthiness. Allow with explicit "synthetic" label, provenance (when available), and reach limits until human review; provide newsroom whitelisting with internal provenance.

Revocation & appeals. Consent grants must be revocable; when revoked, platforms de-list and hash-block. For uploader appeals, provide a 24 - 48 h SLA and a per-item "Why is my reach limited?" explainer with signal-level reasons (e.g., "no provenance; death claim lacked corroboration"). A transparency log records labels/limits.

For any biometric/liveness features (e.g., remote PPG, blink/pose), perform a documented Data Protection Impact Assessment: explicit purpose limitation, opt-in capture, derived-signal storage only with strict retention, subgroup error-rate

auditing, and external review for appeal fairness.

7.7. Legal and Human-Rights Considerations (Brief Matrix)

- **Biometric privacy (face/voice)**

  Comply with biometric laws (e.g., consent/retention limits); store only derived signals for limited time; publish model cards and subgroup error audits.

- **Speech & expression**

  Preserve lawful parody/satire/critique; use labels and reach-limiting rather than removal where feasible; keep an appeals channel.

- **Advertising rules**

  Likeness-based endorsements: require verifiable consent and ad transparency; maintain auditable records.

- **Cross-border data**

  Minimize PII, use regional storage, and document transfer bases.

- **Safe-harbor hosting**

  Maintain notice-and-takedown mechanisms and hashing to prevent re-uploads.

Right-of-publicity in ads. Likeness-based endorsements sit squarely in right-of-publicity doctrine; recent scholarship shows jurisdictions updating rules to address deepfake exploitation. Require auditable consent artifacts and provenance for all likeness ads; otherwise reject.

7.8. Metrics, Definitions, and Targets

Provenance Coverage (PC).

$$PC = \frac{\text{watch time of recommended items with valid C2PA}}{\text{total recommended watch time}}.$$

Target (H1): ≥40%; *H2:* ≥60%.

Time-to-Label (TTL) / Time-to-Takedown (TTD). Median (P50) and tail (P95) from upload/report to label/takedown.

Target: **P95 TTD <4 h** for R1 intimate imagery.

Re-upload Recidivism (RR).

$$RR = \frac{\text{re-upload attempts blocked by hashing within 30 days}}{\text{unique original items}}.$$

Target: ↓ to <1% by Q4.

- **Victim Resolution Time (VRT)**

  From first report to final suppression across mirrors; report P50/P95.

- **Metric Gaming Risks**

  Publish definitions; audit for "credential inflation" (low-value credentials to boost PC) and "takedown deferrals" that pad TTL/TTD. Include third-party audits annually.

Relative Risk Reduction (RRR) = 1 − (Prevalence of high-risk synthetic media in recommendations *after* PPAV ÷ prevalence *before* PPAV), measured on audited samples. Report RRR with CIs per class (R1 - R3).

Anti-Gaming. Publish metric definitions; audit for "credential inflation" (low-value provenance used to game ranking) and for "takedown deferrals" that mask time-to-label/takedown.

7.9. Evaluation Plan

A) Offline (Lab)

- Datasets: FaceForensics++, DFDC, and internal red-team corpora labeled by risk class (R1 - R3).
- Metrics: AUROC/PR for each layer and for the aggregate RRR; ablation to quantify marginal gains; calibration curves; class-conditioned error (e.g., R1 false negatives).
- Latency/cost: Measure per-layer runtime on production hardware; report 95th-percentile latency and estimated cost per million uploads.
- Use FaceForensics++, DFDC, and internal red-team corpora stratified by R1 - R3. Report **AUROC/PR per layer**, **class-conditioned error** (esp. R1 false negatives), **calibration curves**, and **P95 latency** by layer; include ablations for watermark/provenance signals and perceptual hashing variants (PDQ/TMK + PDQF).

B) Online (A/B or Geo-Sim)

- Design: Enable "no provenance, no promotion" in one region for 2 - 4 weeks.
- Primary outcome: Prevalence of mislabeled synthetic media in recommendations (audited sample).
- Secondary: Creator reach fairness (by creator segment/origin), appeals volume, reviewer load, user-reported harm.
- Guardrails: VRT for R1 victims must not regress; creator appeals resolved within 48 hours.
- Primary outcome = prevalence of mislabeled synthetic media in recommendations; secondary = creator-reach fairness, appeals volume, reviewer load, and VRT for R1 cases. Guardrails: no regression on R1 victim SLAs.
- **Threshold calibration & concept drift.** Calibrate per-class thresholds via periodic ROC analysis on fresh, labeled samples; use population stability indices and drift detectors to trigger retraining; apply conservative caps for R1 so that precision remains high with mandatory human review on boundary cases.

8. Human Impact: Mental Health & Social Stability

- **Online victimization → mental-health risk.** When a person's face or voice is weaponized—through non-consensual sexual deepfakes or reputation-destroying fabrications—the harm is not just reputational; it is clinical. Victims commonly report intense anxiety (Maurya et al., 2022), depressive symptoms, sleep disturbance, and hypervigilance that can mirror post-traumatic stress. For adolescents and other vulnerable groups, shame and persistent harassment can escalate to self-harm or suicidal ideation, especially when the content is amplified by recommendation systems and impossible to fully remove. Because the injury is immediate and compounding, platforms should treat identity-and-consent incidents (R1) as **high-severity safety cases**, prioritizing speed, privacy, and survivor control over generic content-policy workflows.
- **Unemployment and financial stress.** Automation and AI can displace tasks

faster than workers can re-skill, creating periods of income loss and status shock. A large clinical and economic literature links job loss and sustained financial strain to elevated risks of depression and suicide; in some regions and time periods, prolonged unemployment also correlates with increases in certain property crimes as households attempt to bridge basic needs. AI is not the sole driver of these outcomes, but adoption shocks can **amplify** known risk factors—particularly for contract workers, creators subject to abrupt monetization changes, and communities concentrated in a single industry.

Unemployment/financial stress: 2024-2025 research continues to show that job insecurity and unemployment elevate depression and suicide risk; recent public-health analyses and meta-analytic work reaffirm the association. Emerging 2025 work argues generative-AI-driven displacement could create tipping-point risks, warranting proactive safeguards.

AI labor shocks today: Fresh labor evidence (2025) shows entry-level workers in AI-exposed sectors are experiencing measurable employment declines, underscoring the need for employer and platform interventions.

- **Safeguards.** Platforms and employers can blunt these harms with a "people-first" posture. Operate **victim-first** response for R1 cases: sub-four-hour takedown service levels for intimate imagery, one-click privacy requests, and robust hash-blocking to prevent reuploads, paired with dedicated survivor support and clear status dashboards. Build **mental-health signposting** into products so that searches related to abuse, self-harm, or doxing trigger evidence-based help resources and de-escalation guidance, while repost attempts of flagged media prompt warnings and offer alternatives. Manage **workforce transitions** deliberately: publish roadmaps ahead of AI rollouts, fund re-skilling with guaranteed placement targets, provide access to counseling and peer groups, and avoid sudden policy swings that destabilize creator income—favor predictable monetization rules with grace periods and appeal channels. Together, these measures acknowledge that authenticity failures and economic shocks are not only trust problems; they are public-health and community-stability problems that demand rapid, compassionate, and measurable interventions.

9. Platform Policy Addenda

- **Celebrity-death protocol.** Treat any upload that asserts or strongly implies a public figure's death as **R3: information harm**. Distribution should be gated until authenticity is established through either cryptographic provenance (e.g., Content Credentials) or corroboration from trusted, authoritative sources. While a claim is under review, the post can exist only in a low-reach state with a prominent context panel that explains, in plain language, that the platform has not verified the report. Recommendation and notifications should remain **off** until verification succeeds, and if verification fails the item should be removed and downstream mirrors suppressed via hashing.

- **Payments and P2P safety.** Crisis-language requests ("urgent", "kidnapped", "medical emergency") are a common attack vector when paired with cloned voices. Platforms that enable peer-to-peer transfers should insert **cool-off timers** and require an additional confirmation step—such as a second factor, callback to a known contact, or a short delay—before funds move. In the product education and safety settings, encourage families to set up **household code-words** so recipients can quickly authenticate a caller or message without revealing sensitive information.
- **Transparency and accountability.** Publish a quarterly **Synthetic-Media Integrity Report** that gives users, researchers, and advertisers visibility into the problem and the platform's response. At minimum, report the prevalence of synthetic-media incidents, time-to-label and time-to-takedown (median and tail), the share of recommended watchtime carrying valid provenance, reupload recidivism after hash-blocking, and victim outcomes such as resolution time for privacy takedowns. Pair these metrics with explanations of policy changes and red-team findings so stakeholders can evaluate progress and hold leadership accountable.

10. Implementation Roadmap (12 - 18 Months)

- **Quarter 1.** Start by naming a single **AI Integrity Owner** with executive authority and budget. That leader should map every **high-risk surface** (uploads, live, ads, messaging, P2P payments) and document current safeguards and gaps. In parallel, pilot **C2PA/Content Credentials verification at upload** on one or two products, exposing a simple "Content Credentials" panel to viewers. Ship **baseline labels** for disclosed or detected synthetic media so teams can exercise the pipeline and measure latency and error rates.
- **Quarter 2.** Move from pilots to enforcement by turning on **no-provenance/no-promotion** in ranking systems across core feeds. Launch **consent and identity gates** for any upload that simulates a real person's face or voice: require verified consent before distribution, and record auditable proofs. Stand up a **victim dashboard** and 24/7 triage with sub-four-hour SLAs for intimate imagery, plus **hash-blocking** to prevent re-uploads. This quarter establishes the operational muscle: queues, on-call rotations, runbooks, and metrics.
- **Quarter 3.** Harden monetization flows. Require **likeness-based ad consent proofs** and provenance for all endorsements; auto-reject ads that fail checks and escalate repeat offenders. Run cross-functional **deepfake drills** (voice-clone fraud, death-hoax workflows, privacy takedowns) for moderators, policy, comms, and finance/AP teams, then fix tooling and policy gaps found during exercises. Publish the first **Synthetic-Media Integrity Report** with prevalence, time-to-label/takedown, provenance coverage, reupload recidivism, and victim outcomes to set a public baseline.
- **Quarter 4 and beyond.** Commission an **independent audit** against NIST/ISO AI-risk practices and C2PA conformance, and publish remediation plans. Ex-

pand provenance coverage by integrating **capture/edit SDKs** so popular creator tools emit Content Credentials by default. Use red-team corpora to tune **PPAV thresholds** for each risk class (stricter for identity/consent than information harm), track false positives/negatives, and iterate on reviewer UX. By the end of this phase, the platform operates in an **authenticate-then-distribute** posture with transparent metrics and third-party validation.

- **Economics & incentives note.** Each quarter, publish a lightweight **cost-benefit snapshot**: (a) compute spend from PPAV (per-million-uploads; attribute P/W/S vs. F/C); (b) reviewer hours; (c) advertiser compliance rate for likeness ads; (d) creator-reach impact by segment; (e) harm-reduction metrics (RRR; VRT). This creates budget discipline without compromising safety.

11. Limitations & Future Work

Watermarking and forensics are adversarially brittle; C2PA coverage is uneven and manifests can be stripped. PPAV mitigates these with ensembles and provenance-aware ranking, but new generative methods may outpace detectors. Future work includes open benchmarks for risk-class detection, standardized creator-impact metrics, and multi-platform shared hashing for synthetic harms.

Watermarking and forensics are adversarially brittle; C2PA coverage is uneven and manifests can be stripped. PPAV mitigates with ensembles and provenance-aware ranking, but new generative methods may outpace detectors. We also note open risks around cross-platform interoperability and provenance stripping; future work should include open benchmarks for *risk-class detection*, *standardized creator-impact metrics*, and *shared hashing for synthetic-harm signatures* across platforms.

12. Conclusion

The conclusion argues that the core problem isn't "AI itself", but how easily AI now lets anyone forge the signals people rely on—sight, sound, and familiar voices—and how quickly social platforms can amplify those forgeries. We are already seeing the costs in three dimensions: money lost to scams, reputations damaged by impersonations and deepfakes, and real mental-health harms to victims. The remedy is not to slow or ban creative uses of AI; it's to raise assurance exactly where harm is likely. That means switching from *publish-then-moderate* to authenticate-then-distribute for high-risk content: before a post can reach public feeds or recommendations, it passes a Pre-Publication Authenticity Verification (PPAV) gate that combines provenance (e.g., Content Credentials), forensic signals (media/watermark checks), and semantic checks (what the content claims) to decide whether to publish normally, publish with limited reach and context, or quarantine for review. It also means running victim-first operations—fast takedowns, hash-blocking of reuploads, and clear survivor support—so people get relief quickly when harm occurs. Finally, platforms must report transparent, auditable metrics (time-to-label, time-to-takedown, provenance coverage, reupload

recidivism) so progress is visible and accountable. Taken together, these steps protect users without stifling creativity, reassure advertisers that the environment is trustworthy, and preserve the long-term legitimacy of the social web.

Conflicts of Interest

The author declares no conflicts of interest regarding the publication of this paper.

References

Farid, H. (2021). An Overview of Perceptual Hashing. *Journal of Online Trust and Safety, 1.* https://doi.org/10.54501/jots.v1i1.24

Internet Crime Complaint Center (2024) FBI-IC3 2024 Annual Report and AI-Enabled Fraud PSAs (Voice-Clone and Impersonation Trends).

Maurya, C., Muhammad, T., Dhillon, P., & Maurya, P. (2022). The Effects of Cyberbullying Victimization on Depression and Suicidal Ideation among Adolescents and Young Adults: A Three Year Cohort Study from India. *BMC Psychiatry, 22,* Article No. 599. https://doi.org/10.1186/s12888-022-04238-x

Murray (2025). *Deceptive Exploitation: Deepfakes, the Rights of Publicity, and Trademark. IDEA Law Review*. Franklin Pierce Law School.

Ortutay, B. (2023). "Take It Down": A Tool for Teens to Remove Explicit Images. *AP News.* https://apnews.com/article/technology-social-media-cameras-business-0f31fbd7ab814f8cdb52ca2df6a46123

Preminger, A., & Kugler, M. B. (2024). The Right of Publicity Can Save Actors from Deepfake Exploitation. *Berkeley Technology Law Journal, 39,* 83-840. https://btlj.org/wp-content/uploads/2024/09/0003_39-2_Kugler.pdf?utm_source=chatgpt.com

Pringle, E. (2023). YouTube's Biggest Star, MrBeast, Seemed to Have Launched the "World's Largest iPhone Giveaway"—It Turns out That, like Tom Hanks, He Was the Face of an AI Scam. *Fortune*. https://fortune.com/2023/10/04/mrbeast-jimmy-donaldson-tom-hanks-subject-ai-scams-false-advertising-deepfakes/?utm_source=chatgpt.com

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*.

Raimondo, G. M., U.S. Department of Commerce, National Institute of Standards and Technology, & Locascio, L. E. (2023). *Artificial Intelligence Risk Management Framework (AI RMF 1.0)*. National Institute of Standards and Technology. https://nvlpubs.nist.gov/nistpubs/ai/nist.ai.100-1.pdf

Reporter, G. S. (2023). Tom Hanks Says AI Version of Him Used in Dental Plan ad without His Consent. *The Guardian*. https://www.theguardian.com/film/2023/oct/02/tom-hanks-dental-ad-ai-version-fake?utm_source=chatgpt.com

Reshef, E. (2023). *Kidnapping Scam Uses Artificial Intelligence to Clone Teen Girl's Voice, Mother Issues Warning*. ABC7 San Francisco. https://abc7news.com/post/ai-voice-generator-artificial-intelligence-kidnapping-scam-detector/13122645/?utm_source=chatgpt.com

Wang, T., Liao, X., Chow, K. P., Lin, X., & Wang, Y. (2024). Deepfake Detection: A Comprehensive Survey from the Reliability Perspective. *ACM Computing Surveys, 57,* 1-35. https://doi.org/10.1145/3699710