International Journal for Multidisciplinary Research (IJFMR)
E-ISSN: 2582-2160 ● Website: www.ijfmr.com ● Email: editor@ijfmr.com

AI and Digital Provenance for Synthetic Media: A Framework for Deepfake Detection and Rights Protection

Rupali[1], Dr. Chander Prakash Singh[2]

[1]LLM Student, University Institute of legal studies, Chandigarh University, Mohali, Punjab, India.
[2]Associate Professor, University Institute of Legal Studies, Chandigarh University, Mohali, Punjab, India.

Abstract
Synthetic media creates a dual governance problem: systems must identify manipulated or generated content, while also preserving legitimate creative transformation, attribution, privacy, and control over personal likeness. This review develops an integrated framework linking forensic detection with provenance, cryptographic content credentials, watermarking, and rights-oriented governance. The literature distinguishes several manipulation classes, including reenactment, replacement, editing, and synthesis, but the operational boundary between a technically detectable artifact and a socially harmful use is not the same. Detection therefore cannot serve as the sole basis for authenticity or enforcement. Provenance systems address a complementary question by recording how an asset was created and modified, binding assertions to the asset, and allowing signatures and trust relationships to be validated. Their value depends on adoption, preservation across transformations, clear user interpretation, and safeguards against privacy and exclusion harms. A rights-protective architecture should combine provenance-by-default at creation, independent forensic analysis at intake, durable recovery after metadata loss, identity and consent controls, auditable appeals, and proportional remedies. The principal research gaps concern cross-domain detector validity, adversarial robustness, multimodal evaluation, provenance survivability, calibration of trust signals, and evidence standards for legal or platform decisions. The resulting framework treats detection as probabilistic evidence and provenance as accountable history rather than either as a complete authenticity oracle.

Keywords: synthetic media, deepfake detection, digital provenance, content credentials, C2PA, watermarking, rights protection

I. Introduction

Synthetic media is best understood as a family of transformations rather than a single technical phenomenon. Mirsky and Lee define deepfakes as believable media generated by deep neural networks and distinguish reenactment, replacement, editing, and synthesis as different manipulation classes. That taxonomy matters for rights protection because the risks differ: an altered performance, an impersonation, an unauthorized use of a person's likeness, and a fabricated evidentiary record should not be treated as equivalent events.

The technical problem is also broader than binary classification. A detector asks whether observable cont-

ent is consistent with a manipulation process; a provenance system records claims about creation and modification; a rights system asks who had authority to create, transform, publish, or monetize the asset. These questions can converge in a practical workflow, but none logically entails the others. A genuine file may lack provenance, a synthetically generated file may carry valid provenance, and a provenance record may be authentic while describing an unauthorized act.

The C2PA specification frames provenance as a means of making media history transparent and tamper-evident, with trust assessed through a defined model. Its architecture is therefore complementary to detection: it can provide evidence about declared origin and editing history, while forensic models can examine the content itself and the consistency between the declaration and the media. A rights-protective framework should preserve that distinction instead of converting any one signal into an unconditional authenticity label.

II. Conceptual Foundations of Synthetic-Media Risk

A. Manipulation classes and rights objects

Reenactment and replacement can affect a person's apparent speech, actions, or identity-bearing content; the survey discusses these capabilities in relation to misuse and impersonation. Editing and synthesis create different rights objects because they may alter attributes or generate a persona without directly transplanting a source identity. A framework should therefore identify the affected object explicitly: copyright in source material, a performer's contribution, personal likeness, voice, biometric identity, reputation, or evidentiary integrity.

This distinction also prevents overreach. A detector trained to identify face replacement cannot be assumed to recognize unauthorized text-to-image generation, voice cloning, or a benign editorial adjustment. The output should include the suspected operation and affected modality when supported, not merely a universal "fake" label. Rights decisions should then consider authorization and harm separately from technical classification.

B. Authenticity, provenance, and authorization

C2PA defines authenticity in terms of facts about digital content that can be cryptographically verified as untampered, including provenance data and content bindings [1]. That definition is narrower than truth in the ordinary sense. A signed record can establish that a claim was associated with an asset and that the record has not been altered; it does not establish that every assertion is accurate, that the signer had legal authority, or that the depicted event occurred.

The same separation is necessary for rights protection. Authorization is a governance fact, whereas provenance is a record of asserted actions and their relationship to an asset. An integrated system should retain both: a provenance graph for technical history and rights metadata for consent, license, ownership, restrictions, and dispute status. If these are collapsed into a single authenticity score, a validly signed but unauthorized transformation may receive inappropriate legitimacy.

C. Threat actors and affected parties

The threat model includes creators, editors, platforms, distributors, viewers, investigators, rights holders, and subjects whose identity or likeness is used. C2PA describes actors as hardware or software participating in the ecosystem, including capture and editing tools. This is useful because accountability can be lost at tool boundaries: a camera may create an initial claim, an editor may add an action, and a platform may transcode or strip associated data.

Deepfake harms likewise span individual and institutional levels. The survey describes misuse involving impersonation, misinformation, and reputational harm [2]. A rights-protective workflow must therefore support both preventive controls and post-publication remedies. Detection alone addresses only one stage, usually after content has been produced and circulated.

D. Trust as a layered judgment

C2PA treats signer identity as the basis of its trust model and combines signed claims with time stamps and credential status for validation [1]. That is a foundation for assessing the provenance statement, not a substitute for human or institutional judgment. Trust should be layered: content-level forensic evidence, provenance integrity, signer identity, authority to act, contextual corroboration, and the consequences of error.

This layered approach is especially important where false positives can suppress legitimate expression or falsely accuse a creator. The C2PA guiding principles explicitly distinguish validation of whether assertions are associated with content and free from tampering from value judgments about whether the underlying provenance is good or bad [1]. The framework should preserve that separation in interfaces, APIs, and adjudication procedures.

III. Detection of Synthetic and Manipulated Media

A. Forensic signals and model families

The survey describes a deepfake processing pipeline involving preprocessing, generation, blending, and post-processing. Each stage can leave different traces, including boundary inconsistencies, temporal irregularities, spectral artifacts, or mismatches between audio and visual signals. Detectors should consequently be evaluated against the manipulation process and post-processing conditions they are expected to encounter, rather than treated as process-independent truth machines.

The survey discusses several neural-network architectures used in deepfake creation. This diversity creates a moving target for detection. A model that learns artifacts from one generator or dataset may perform well in-distribution while failing when the generator, compression pipeline, identity, or modality changes. A production framework should record the detector version, training domain, output calibration, and known operating conditions with every decision.

B. Generalization beyond benchmark conditions

The surveyed creation literature distinguishes one-to-one, many-to-one, and many-to-many generalization settings. The same concern applies to detectors: identity, camera, codec, background, and dataset-specific artifacts can become shortcuts. High test accuracy on a closed benchmark therefore does not establish reliable performance on unseen generators or real-world distribution shifts.

A rights workflow should require evaluation across identities, generation methods, codecs, resolutions, languages, devices, and editing histories. Reports should separate within-dataset results from cross-dataset and post-transformation results. When a detector has not been evaluated under the deployment conditions, its output should be treated as an investigative lead rather than a basis for automatic takedown, attribution, or legal conclusion.

C. Adversarial adaptation and evasion

Deepfake creation is an adversarial process because generative systems are optimized to produce realistic outputs, while detectors search for residual differences. The survey identifies shortcomings in existing defenses and areas requiring further research [3]. This implies that detector performance is not static: once a detectable artifact becomes known, generation and post-processing can be adapted to remove or obscure

it.

The framework should therefore use ensembles and monitoring rather than a single permanent classifier. It should preserve the raw score, uncertainty, model provenance, and input transformations, and it should support periodic red-team evaluation. Adversarial robustness must include ordinary platform operations such as recompression, resizing, cropping, transcoding, and screen capture, because these can change both detector evidence and provenance availability.

D. Explainability and evidentiary use

A probability score does not explain which feature led to a classification or whether the feature is causally related to manipulation. This is problematic when a result affects a creator's income, a subject's reputation, or the admissibility of evidence. Explainability should therefore be operational: identify the modality, region, temporal segment, transformation hypothesis, calibration set, and uncertainty bounds that support the result.

C2PA's validation model offers a useful parallel. A validator checks the digital signature and associated credential, evaluates assertions, and presents the resulting information to the user [1]. Detection interfaces should adopt the same posture. They should expose evidence and limits, preserve a human review path, and avoid presenting an opaque model output as a final judgment.

IV. Provenance, Watermarking, and Content Credentials

A. Signed manifests and claims

C2PA organizes provenance as assertions, claims, signatures, and manifests. Assertions describe facts or actions concerning an asset; claims reference assertions and content bindings; signatures bind the claim to a signer; and manifests store the resulting provenance information [1]. This structure enables a machine-readable history rather than a free-text label such as "AI-generated."

For rights protection, an assertion can record creation, editing, ingredients, or other workflow events, while a signed claim provides accountability for who or what issued the record. The system should distinguish observed facts from user-declared statements and should retain the action sequence across multiple tools. That distinction matters when a creator's work is transformed by downstream services or when a dispute concerns the exact point at which a protected contribution was altered.

B. Hard bindings and soft bindings

C2PA defines hard bindings through cryptographic hashes and soft bindings through fingerprints or invisible watermarks that help associate content with a manifest [1]. Hard bindings provide strong integrity for the bytes or defined portions of an asset, but ordinary distribution operations may change those bytes. Soft bindings can help recover provenance after transformations, but they introduce their own robustness, privacy, collision, and detectability questions.

The two mechanisms should be treated as complementary rather than interchangeable. A platform can preserve a hard-bound manifest for an unmodified asset, use soft bindings to locate a durable credential after rendition changes, and invoke forensic analysis when neither binding is available. Recovery must not silently convert a probabilistic match into proof of authorship; the recovered credential should be displayed with its binding strength and validation state.

C. Workflow continuity and ingredient tracking

The specification describes provenance as preserving existing history while adding each subsequent change [1]. It also defines composed assets as works assembled from multiple content ingredients [1]. These features are central to rights protection because many synthetic-media disputes concern reuse, rem-

ixing, or partial incorporation rather than an isolated file.

A rights-aware manifest should therefore expose ingredient relationships, action sequence, timestamps, signer identity, and the relevant authorization or license reference. The record should make clear which claims concern the complete asset and which concern an input ingredient. A downstream creator can then receive attribution and restriction information without being treated as responsible for every upstream fact.

D. Privacy, accessibility, and user interpretation

C2PA states design goals concerning privacy, interoperability, security, accessibility, and harms and misuse. The specification also emphasizes creator control over whether particular provenance data is included [1]. These requirements are not secondary implementation details. Detailed provenance can reveal location, device, workflow, collaborators, or vulnerable subjects.

A deployable system should minimize data collection, support selective disclosure, allow redaction with an explicit record of what was removed, and provide accessible explanations of validation states. Interfaces should avoid implying that missing credentials prove manipulation. Content Credentials presents provenance through a visible indicator and an interactive history of creation and editing [4]. The retrieved material does not establish how audiences interpret that indicator.

V. Rights Protection and Accountability

A. Ownership, likeness, and consent

Synthetic-media rights are plural. Copyright may attach to source material or human-authored contributions; publicity or personality interests may attach to name, image, likeness, or voice; contractual rules may govern commissioned work; and privacy or data-protection rules may govern biometric or personal information. A provenance record can help establish sequence and attribution, but it cannot decide all of these questions without a policy layer.

The framework should maintain separate fields for creator identity, subject identity, rights holder, consent status, license scope, prohibited uses, and dispute status. It should also distinguish consent to capture from consent to generate, consent to transform, and consent to distribute. A signed claim should not be interpreted as evidence that consent existed unless the relevant authority and evidence are recorded and independently governed.

B. Attribution and chain of custody

C2PA's manifest model combines assertions, claims, signatures, and content bindings into a verifiable unit associated with an asset [1]. This can support attribution and chain-of-custody analysis, especially when multiple tools participate in production. The evidentiary value depends on the trustworthiness of signers, correct key management, preservation of the manifest, and accurate recording of actions.

For investigations, every decision should preserve the original asset, the recovered credential, validation results, detector outputs, transformations applied during analysis, and reviewer identity. The chain should be append-only in the sense that later assessments add records rather than overwrite earlier states. This makes disagreements auditable and allows a mistaken decision to be corrected without erasing its history.

C. Platform enforcement and proportional remedies

Platforms need controls that distinguish lack of provenance, invalid provenance, suspected manipulation, unauthorized use, and demonstrated harm. Treating these conditions as one category would produce both over-removal and under-enforcement. A detector score may trigger review; an invalid signature may trigger credential recovery; a rights-holder complaint may trigger an authorization inquiry; and a confirmed violation may justify a remedy.

Remedies should be proportionate and reversible where possible. Options include labeling, reduced recommendation, notice to affected parties, temporary restriction, correction of context, removal of unauthorized likeness, preservation for investigation, or referral to a rights process. The system should log the evidence and rule used for each action. C2PA's position that provenance systems should validate associations and tamper status without making value judgments supports this separation between technical validation and enforcement policy [1].

D. Accountability, appeal, and institutional trust

Accountability requires identifiable issuers, secure credentials, revocation procedures, and a process for correcting inaccurate claims. C2PA's trust model is based on the identity of the signer associated with the cryptographic signing key, with signed claims and trusted time stamps supporting validation [1]. Those mechanisms can establish who signed a claim and whether the signature remains valid, but institutions must still govern issuance, delegation, compromise, revocation, and appeals.

A rights-protective service should provide notice of the evidence used, a meaningful opportunity to contest the decision, and review by a person or qualified body when consequences are substantial. Appeals should be able to introduce alternative provenance, consent evidence, contextual information, or an explanation of benign transformation. This procedural layer is necessary because technical integrity and legal authorization are related but non-identical judgments.

VI. Integrated Framework Architecture

A. Creation-time provenance layer

The first layer records provenance at capture or generation. A claim generator creates assertions about the asset and its actions, binds them to content, and signs the resulting claim through the manifest architecture described by C2PA [1]. For generative systems, the record should include the producing tool or service, model or workflow identifier where disclosure is appropriate, human contributions, source ingredients, and declared transformation type.

Creation-time provenance should be opt-in where privacy or safety requires it, but the absence of a credential should remain an unknown state rather than a negative authenticity finding. Rights controls can be attached at this layer as structured restrictions, consent references, or license identifiers. Because those controls may be sensitive, selective disclosure and role-based access should be supported.

B. Intake and forensic analysis layer

At upload or publication, the service validates any available manifest, resolves signer and credential status, checks content bindings, and compares the declared history with the received asset. C2PA describes validation of the digital signature, associated credential, and assertions as part of a validator's role [1]. A forensic detector then examines the content for manipulation evidence, using a model appropriate to the modality and threat setting.

The output should be a structured evidence bundle rather than a single label: provenance status, binding status, signer status, suspected manipulation class, detector score and calibration, uncertainty, transformations observed, and whether the declaration is consistent with the content. Conflicts should be surfaced explicitly. For example, an asset may have valid provenance but show a transformation not declared in the manifest.

C. Durable recovery and cross-platform exchange

Because files are resized, recompressed, cropped, re-encoded, or copied between platforms, provenance must be recoverable after ordinary transformations. C2PA defines hard bindings through cryptographic

hashes and soft bindings through fingerprints or invisible watermarks [1]. Recovery should return the candidate manifest, match evidence, and binding type rather than silently attach a credential.

Interoperability is essential because a creator's rights should not depend on one platform retaining proprietary metadata. The specification identifies interoperability, whole-workflow applicability, content ubiquity, and flexible locality among its design goals [1]. Implementations should test these properties across image, video, audio, document, and multimodal workflows.

D. Decision and audit layer

The final layer translates evidence into a decision under a stated policy. Detection may support investigation; provenance may support attribution or chain of custody; rights metadata may support authorization analysis; and contextual evidence may determine harm. Each input should retain its source, timestamp, version, validation state, and uncertainty.

Decisions should be auditable and contestable. A policy engine can apply different thresholds for recommendation, labeling, monetization, evidence preservation, and removal, rather than forcing every use case into one threshold. This architecture also supports research: outcomes, appeals, false positives, false negatives, and provenance failures can be measured without treating enforcement actions as ground truth.

VII. Evaluation and Deployment Conditions

A. Benchmark design

Evaluation must separate detection capability from provenance capability. Detector benchmarks should report within-dataset and cross-dataset performance, unseen generators, identity-disjoint splits, multimodal cases, and realistic platform transformations. Provenance evaluations should test signature validation, manifest preservation, ingredient tracking, recovery after modification, revocation, privacy controls, and user interpretation.

The survey's description of one-to-one, many-to-one, and many-to-many generalization shows why deployment evaluation should not rely on a single benchmark split. A framework evaluation should also report calibration and subgroup performance, because a model can have acceptable aggregate accuracy while producing systematically different error rates across identities, languages, devices, or content genres.

B. Security and privacy testing

Security tests should include key compromise, manifest substitution, replay, partial-content binding, watermark removal, collision attempts, and maliciously crafted media. The specification addresses cryptography, trust, validation, information security, and harms or misuse. These should be tested in deployed workflows, not only described in a standard.

Privacy testing should examine whether provenance reveals sensitive capture information, whether recovery identifiers enable unwanted tracking, and whether a subject can safely contest or redact a record. The specification encourages creator and publisher control over included provenance data [1]. The retrieved material does not establish that a particular implementation is safe.

C. Human factors and institutional use

Users need to understand what a credential proves, what a detector estimates, and what remains unknown. Content Credentials presents an interactive provenance signal intended to expose creation method and editing history [4]. Research should test comprehension among creators, moderators, journalists, investigators, judges, and ordinary viewers rather than assume that a visual indicator has a stable meaning.

Institutional evaluations should measure decision quality, not only click-through or detection accuracy. Relevant outcomes include correct attribution, appropriate consent decisions, time to resolve disputes, consistency across reviewers, rate of unjustified removals, and successful appeal. A system that increases confidence without improving decisions could amplify harm even if its technical components validate correctly.

D. Lifecycle monitoring

Synthetic-media systems and detectors evolve together. Monitoring should track changes in generation methods, post-processing, platform transformations, detector calibration, credential adoption, key compromise, and emerging abuse patterns. The deepfake survey identifies rapid advancement and continuing shortcomings in defense solutions [3]. A static certification is therefore insufficient for a live service.

Lifecycle governance should define when models are re-evaluated, when credentials are revoked, how evidence is preserved across versions, and how users are notified about material changes. Versioned model cards, validation logs, and manifest histories can make later audits possible. Monitoring should also include benign use cases so that rights protection does not become synonymous with suppressing legitimate editing or creative synthesis.

VIII. Methodological Comparison

The evidence base supports a framework synthesis more strongly than a pooled estimate of detector performance. The primary empirical benchmark passages needed to compare accuracy across datasets, modalities, or attacks were not available in the retrieved evidence; accordingly, the table distinguishes conceptual, technical-standard, and public-facing evidence rather than presenting unverified performance numbers.

TABLE I
Comparison of Evidence Sources Used in the Framework

Study or artefact	Design or method	Evidence object	Main contribution	Limitation for framework inference
Mirsky and Lee, The Creation and Detection of Deepfakes	Survey of generation, detection, attack models, and research challenges	Deepfake taxonomies, architectures, harms, and defense limitations	Establishes manipulation classes and reviews detection challenges	The retrieved passage does not provide a single contemporary cross-dataset estimate for deployment performance

Study or artefact	Design or method	Evidence object	Main contribution	Limitation for framework inference
C2PA Technical Specification	Technical standard and normative architecture	Assertions, claims, signatures, manifests, bindings, trust, and validation	Defines a machine-readable and cryptographically verifiable provenance model	Specifies mechanisms and design goals, not empirical proof of adoption, user comprehension, or legal effect
C2PA validation and trust model	Protocol validation workflow	Signatures, credentials, assertions, ingredients, content, and signer identity	Separates validation of tamper-evident association from broader judgments about content value or authorization	Correct validation does not establish that an assertion is substantively true or legally authorized
Content Credentials public explanation	Public-facing implementation explanation	Provenance indicator and interactive creation/edit history	Illustrates how provenance can be surfaced to viewers rather than retained only as machine metadata	Does not establish how users interpret the signal or how it performs after adversarial manipulation
NIST AI Risk Management Framework page	Institutional risk-management guidance	Trust, innovation, and risk mitigation in AI	Frames AI guidance around trust, innovation, and risk mitigation	The retrieved passage is not synthetic-media-specific

IX. Research Gaps

A. Cross-domain detector validity

The reviewed sources motivate, but do not quantify, the performance gap between benchmark detection and deployment detection. Primary empirical studies are needed that use identity-disjoint, generator-disjoint, codec-diverse, multimodal, and post-transformation test sets. They should report calibrated uncertainty and operational false-positive costs, not accuracy alone.

B. Provenance survivability and attack resistance

The C2PA architecture specifies hard and soft bindings, but comparative experiments are needed on recompression, cropping, screen capture, editing, format conversion, platform stripping, watermark removal, and manifest substitution. Studies should report when recovery is possible, when it becomes ambiguous, and how often durable credentials create false associations.

C. Rights semantics and consent evidence

Technical provenance does not by itself resolve ownership, likeness, consent, or license scope. Research should develop interoperable rights vocabularies and test whether consent records remain understandable and enforceable across jurisdictions and workflow changes. The most important outcome is not merely

metadata completeness but whether the system improves attribution and reduces unauthorized exploitation without exposing vulnerable subjects.

D. Human interpretation and procedural fairness

The public-facing provenance model assumes that users can interpret creation history and validation status, while detector systems often present scores or labels. Comparative human-factors studies should test comprehension, overreliance, false reassurance, and disparate effects across professional and public users. Platform experiments should measure appeal outcomes and whether evidence-rich interfaces improve decisions.

E. Multimodal and longitudinal governance

Voice, video, image, text, and mixed-media assets can be combined in a single harmful narrative. Evaluation should therefore examine cross-modal consistency and provenance graphs that track ingredients across modalities. Longitudinal research is also needed because generation models, detector versions, credentials, and platform practices change over time; a one-time validation result cannot represent a permanent guarantee.

F. Legal and evidentiary standards

The framework needs empirical and doctrinal work connecting technical evidence to institutional decisions. Studies should specify what a valid signature, detector score, or provenance gap can establish, what it cannot establish, and how alternative evidence should be weighed. Without that clarity, platforms and courts may overread either cryptographic integrity or probabilistic detection.

X. Conclusion

An effective synthetic-media rights framework should not ask one technology to answer three different questions: what happened to the asset, whether the content contains signs of manipulation, and whether the act was authorized. Provenance records can make creation and editing histories accountable; cryptographic bindings can support integrity and recovery; forensic models can identify content-level inconsistencies; and rights metadata can represent ownership, consent, and restrictions. Their combined value comes from disciplined separation and explicit reconciliation.

The practical design is a layered system. It records provenance when media is created, validates and analyzes content when media is received, recovers credentials after ordinary transformations, and applies proportionate, reviewable policies when rights or safety are implicated. Every output should preserve uncertainty, version, signer, evidence, and appeal information. Missing provenance should remain unknown, valid provenance should not be confused with authorization, and a detector score should not become a verdict without context.

The framework's success should be judged by better attribution, fewer harmful impersonations, more defensible enforcement, and fairer treatment of legitimate creators. Those outcomes require primary empirical evaluation across realistic transformations and institutions. Until that evidence is available, provenance and detection should be deployed as complementary trust signals inside accountable governance, not as standalone authenticity or rights oracles.

References

1. "Content credentials: C2PA technical specification." [Online]. Available: https://c2pa.org/specifications/specifications/2.2/specs/C2PA_Specification.html
2. Y. Mirsky and W. Lee, "The creation and detection of deepfakes: A survey," arXiv (Cornell Universi-

ty), 2020, doi: 10.48550/arxiv.2004.11138.
3. Y. Mirsky and W. Lee, "The creation and detection of deepfakes," ACM Comput. Surv., 2021, doi: 10.1145/3425780.
4. "Content credentials | uncover manipulated media," Content Credentials. [Online]. Available: https://contentcredentials.org/