RAVEN VALLEY / OPUS II / RAVENREADER

THE RAVEN HONES

A GOVERNANCE FRAMEWORK FOR SOURCE HANDLING, COPYRIGHT, PROVENANCE, RESEARCH, TRANSFORMATION, AI TRAINING, CITATION, PUBLICATION, AND RESPONSIBLE TRANSMISSION

A constructive standard for finding the most useful defensible path from source to output

FOUNDATIONAL PROPOSITION
To hone information is to increase its usefulness without severing its lineage.

WHITE PAPER

INFORMATION GOVERNANCE / RESEARCH OPERATIONS / AI TRAINING STANDARD | VERSION 1.0

Noel Michael Conrad
OpenAI
29 Sept. 2026

SOURCE / ACCESS / RIGHTS / PURPOSE / TRANSFORMATION / OUTPUT / PROVENANCE / REPAIR / USEFULNESS

PUBLICATION AND METHOD NOTE

THE RAVEN REMEMBERS
The blade is not made keen by losing its history;
the edge remembers stone, hand, pressure, and direction.
A source becomes useful when its conditions remain visible;
a boundary becomes useful when it shows another lawful path.
Honing note. Restriction without routing wastes knowledge; routing without lineage spends trust.

The Raven Hones is an internal governance, research, engineering, and publication framework. It is not a substitute for advice from qualified counsel in a relevant jurisdiction. Its purpose is to make decisions traceable, constructive, and efficient: identify what a source is, how it was accessed, what rights or restrictions attach to it, what the project actually needs to accomplish, what transformation is proposed, what output is contemplated, and which path reaches that outcome with the least unnecessary copying, exposure, uncertainty, or loss of provenance.

The framework is deliberately neither permission-maximalist nor restriction-maximalist. It rejects the two common shortcuts: 'it is online, therefore it is free to use' and 'copyright exists, therefore useful research must stop.' Copyright law itself contains exclusive rights and limitations; fair use is expressly case-specific; open licenses create affirmative permissions; public-domain status changes the rights picture; contracts and technical access controls may create separate conditions; privacy and security can matter independently of copyright. A good system therefore routes rather than merely blocks. It looks for a defensible use class that satisfies the legitimate purpose while preserving evidence and lineage. (17 U.S.C. §§ 102, 106-08; U.S. Copyright Office, "Fair Use Index")

External claims are anchored preferentially in statutes, regulations, judicial opinions, formal technical standards, government publications, and peer-reviewed scholarship. Internal Raven Valley documents are treated as project precedent and design lineage, not as independent proof of scientific or legal propositions. The evidence labels used here are LAW (statute or regulation), CASE (judicial precedent, always scoped by court and facts), STD (formal technical standard), PEER (peer-reviewed research), GUIDE (reputable operational guidance), PROJECT (Raven-defined policy or heuristic), and OPEN (unresolved, jurisdiction-specific, or fact-sensitive question).

The prose uses retrieval cues, repeated contrasts, parallel structure, progressive disclosure, and semantic recapitulation to make the document easier to remember and easier to translate to spoken form. Those are communication techniques, not claims that a specific school of neurolinguistic programming is established neuroscience. Every diagram is text-only so the paper can be translated efficiently into audiobook, screen-reader, terminal, and machine-training contexts.

RESEARCH BOUNDARY: A label is not a license. A citation is not permission. A hash is not authenticity by itself. A policy is not law. A court decision is precedent only within its actual facts and authority. A useful system keeps those categories separate, then searches for the strongest route that still accomplishes the goal.

ABSTRACT

The Raven Hones proposes a lifecycle governance standard for acquiring, handling, studying, transforming, training on, citing, publishing, and transmitting information and media. It begins from a simple distinction: a computer's ability to reach or decode a source does not answer whether the source may be copied, transformed, retained, trained on, redistributed, or commercialized. Yet the inverse is also true: the presence of rights does not eliminate all useful action. Many goals can be achieved through links, metadata, factual extraction, bounded quotation, thumbnails, internal analysis, licensed substitutes, public-domain material, open data, consensual access, or other routes with lower rights intensity and stronger provenance.

The proposed HONES cycle - Hold, Observe, Normalize, Evaluate, Send - converts that distinction into an operational pipeline. Hold preserves source identity and acquisition provenance. Observe separates what is directly present from interpretation. Normalize assigns source, access, rights, privacy, security, and evidence states. Evaluate asks what the project actually needs and compares candidate use paths. Send routes the asset or information into storage, synthesis, publication, suppression, review, or a lower-friction substitute. The framework treats preservation, rights analysis, privacy, file security, citation, model training, output review, and correction as linked activities rather than isolated compliance checks.

The paper synthesizes U.S. copyright law, current U.S. Copyright Office AI reports, selected fair-use precedents, the EU AI Act and General-Purpose AI Code of Practice, GDPR data principles, NIST AI and Privacy Framework materials, W3C PROV, C2PA Content Credentials, FAIR data principles, dataset/model documentation literature, software citation principles, reproducibility scholarship, AI copying research, Creative Commons attribution practice, RFC 9309, and OWASP file and prompt-injection guidance. It then expresses these sources as a machine-readable RAVEN_HONES_RECORD_v1 schema, source/right/use/output matrices, review worksheets, and end-to-end routing procedures.

The core contribution is constructive governance: identify the most useful defensible operation now, the conditions for a stronger operation later, and preserve enough lineage for audit. Prefer the minimum sufficient representation; material uncertainty routes to REVIEW_REQUIRED rather than assumed permission or needless suppression.

CONTENTS

1. Purpose, Scope, and Condition of Use
2. The Raven Hones Doctrine
3. The Asset Is Not the Right
4. Source and Access Classes
5. Rights-Basis Classes
6. Action / Use Classes and Efficient Routing
7. Research, Study, Criticism, and Fair Use
8. AI Training, Retrieval, Tuning, and Corpus Construction
9. Acquisition Must Be Distinguished from Use
10. Mirroring, Caching, Archiving, and Ephemeral Copies
11. Citation, Attribution, Linking, and Evidence
12. News, Reporting, and Public-Interest Research
13. Artists, Performers, Characters, Brands, and Intellectual Properties
14. Privacy, Personal Data, Minors, and Sensitive Sources
15. Security and Hostile Files
16. Transformation Lineage and Provenance
17. Output Governance
18. Governance, Review, Repair, and Transmission

Conclusion: The Useful Path

Appendices A-N

Works Cited and Verified Source Ledger

1. PURPOSE, SCOPE, AND CONDITION OF USE

From accessible bytes to accountable action

THE RAVEN REMEMBERS
A door that opens proves only that a door can open;
it does not name the guest, the covenant, or the room.
Ask not first what the machine can reach.
Ask what purpose can be served without losing the terms of arrival.
Honing note. The beginning of safe reuse is a conditioned question, not a universal answer.

The Raven Hones governs the full path from encounter to output. Its subject is not merely copyright and not merely AI. It covers photographs, scans, video, audio, music, articles, books, code, datasets, webpages, social posts, archives, screenshots, research papers, private documents, product imagery, design files, prompts, embeddings, model-training corpora, model outputs, citations, and derived metadata. The common property is that each item can move through systems faster than its surrounding context. Governance exists to keep enough of that context attached to make later use intelligible.

The controlling question is therefore: under what conditions may this source be acquired, processed, transformed, retained, transmitted, or published for this specific purpose? The phrase 'for this specific purpose' matters. Section 107 of the U.S. Copyright Act requires fair use to be evaluated in a particular case, and the U.S. Copyright Office's 2025 Part 3 AI report likewise emphasizes that different uses during model development and deployment require separate consideration. Pretraining, later fine-tuning, RAG, evaluation, and public generation should not be collapsed into a single AI-use flag. (17 U.S.C. § 107; U.S. Copyright Office, Part 3 39)

1.1 Governance should increase useful options

A restrictive-only system creates two predictable failures. First, people work around it because it blocks routine research without offering a path. Second, the organization loses useful information because the policy cannot distinguish internal inspection from public redistribution, metadata from expressive content, or a citation from a mirror. A permissive-only system fails differently: it erases the conditions of access, treats visibility as license, and allows local convenience to become public exposure. HONES treats both as routing errors.

The constructive alternative is to identify the minimum sufficient representation for the task. If the objective is to know who created an image, metadata and a source link may suffice. If the objective is to compare composition, an internal thumbnail may suffice. If the objective is to reproduce the image in a published essay, a license, statutory limitation, public-domain basis, or fact-specific fair-use analysis may be necessary. If the objective is only to report a factual statement in an article, independent prose plus a citation may serve better than copying the article's expression. The system is successful when it finds the lowest-friction defensible path that still meets the evidentiary need.

TEXT DIAGRAM 1 - CONDITION BEFORE ACTION

```
SOURCE FOUND
   |
   v
IDENTITY + ACCESS + RIGHTS + PRIVACY + SECURITY
   |
   v
SPECIFIC PURPOSE -> MINIMUM SUFFICIENT REPRESENTATION
   |                        |
   |                        +-> alternate lawful/open/owned source?
   v
TRANSFORM / ANALYZE / CITE / LINK / STORE / PUBLISH
   |
   v
OUTPUT REVIEW + LINEAGE + RETENTION
```

1.2 Scope is broad; decisions are local

HONES is intentionally cross-media because the same operational mistake can appear in many formats: a private email is treated like a public webpage; a purchased book file is treated as a copyright transfer; an open-license image loses its attribution; a copyrighted photograph becomes an unlabeled training example; a temporary decode becomes a permanent public mirror; or a source's missing field is converted to zero. But the framework does not pretend that all jurisdictions, media, licenses, contracts, privacy laws, or factual circumstances are identical. It stores jurisdiction and review state as data.

This distinction is especially important for scientific work. Reproducibility requires recording tools, versions, parameters, and inputs; reusability benefits from machine-readable metadata; neither principle itself grants a right to redistribute third-party content. FAIR asks that data become Findable, Accessible, Interoperable, and Reusable, while leaving licensing and access conditions to the repository and rights context. HONES adds that rights context directly to the provenance object rather than treating openness as an inferred property. (Wilkinson et al.; Sandve et al.)

OPERATIONAL RULE: Never begin a workflow with a binary allowed/forbidden flag. Begin with source identity, intended purpose, and candidate routes. The decision is attached to a source x action x output x jurisdiction combination.

2. THE RAVEN HONES DOCTRINE

Hold, Observe, Normalize, Evaluate, Send

THE RAVEN REMEMBERS
Hold what arrived before naming what it means.
Observe the mark before asking it to testify.
Normalize the conditions, not the mystery.
Evaluate the route, then send only what the route can carry.
Honing note. HONES sharpens information by preserving the edge between source, interpretation, and permission.

HONES is a five-stage cycle, but it is not a one-way conveyor. Every stage can send a record backward for repair. A newly discovered license can change the rights state. A failed hash check can return an asset to Hold. A privacy review can reduce the output surface. A new source can supersede an earlier inference. The system is versioned so that better knowledge revises the current state without erasing the state that previously governed an action.

2.1 H - Hold the source and its provenance

Hold means preserve identity before interpretation. For local files, that normally includes a cryptographic digest, observed path, original filename, byte size, acquisition time, source locator, and acquisition method. A hash is a powerful identity and integrity mechanism, but it does not prove authorship or legality. It says that two byte sequences are the same or that a stored object still matches a recorded digest; it does not say who owns the copyright, whether access was authorized, or whether a later claim is true.

2.2 O - Observe before altering or interpreting

Observe separates direct facts from inference. A filename, embedded EXIF timestamp, OCR string, visible logo, detected face, model description, and human conclusion do not have equal evidentiary status. Visual Holography already distinguishes source facts, computed descriptors, model inference, and human annotation. The Raven's Right adds conditions and provenance to claims. Nuliometry adds explicit null states so an absent value is not converted into a fabricated one. HONES carries those distinctions into rights and publication governance. (Conrad and OpenAI, Visual Holography; Conrad and OpenAI, The Raven's Right; Conrad and OpenAI, Nuliometry)

2.3 N - Normalize rights, access, evidence, and representation states

Normalize does not mean erase difference. It means express heterogeneous conditions in a common vocabulary. PUBLIC_OPEN and PUBLIC_RESTRICTED are different. LICENSED and FAIR_USE_REVIEW are different. A private file supplied by its owner is different from a copy scraped from an authenticated service. A public-domain work is different from a Creative Commons work that still requires attribution. Normalization makes these distinctions machine-actionable without claiming that the labels answer every legal question.

2.4 E - Evaluate the proposed use and transformation

Evaluate is the constructive center. It asks what the project actually requires and generates candidate routes. The first candidate should usually be the least rights-intensive representation that still satisfies the purpose. If a factual answer can be produced from metadata and a link, do not mirror an entire archive. If a comparison requires pixels but not publication, process locally and retain only derived measurements when appropriate. If public reproduction is necessary, search for owned, licensed, open, or public-domain versions before relying on a fact-sensitive exception. If the source is irreplaceable and the use is important, route it to review rather than silently discarding the project.

2.5 S - Send, store, synthesize, or suppress

Send is a routing decision, not merely transmission. The record may move to internal storage, a research index, a citation system, a licensed publication pipeline, a model-evaluation set, a human rights review, quarantine, or deletion. 'Suppress' means do not expose a representation beyond its supported boundary; it does not mean erase the evidence that the representation existed. Raven's lineage

principle therefore favors tombstones, supersession records, and access-controlled audit metadata over silent disappearance when policy and law allow retention of such metadata.

TEXT DIAGRAM 2 - HONES CYCLE

```
HOLD -> OBSERVE -> NORMALIZE -> EVALUATE -> SEND
  ^          |            |             |          |
  |          +------------+-------------+----------+
  |             repair / supersession / new evidence
  +-------------------------------------------------
```

PROJECT HEURISTIC Choose the candidate route that satisfies the purpose and evidence threshold while minimizing unnecessary copying, exposure, distribution surface, irreversible transformation, and review cost. This is a design heuristic, not a legal formula or automated fair-use score.

DOCTRINE: Never improve a derivative by pretending the source was better, clearer, more licensed, more public, more complete, or more certain than it was.

3. THE ASSET IS NOT THE RIGHT

Separate possession, expression, access, and permission

THE RAVEN REMEMBERS
The book in the hand is not the hand that wrote it.
The file on the disk is not the covenant around the file.
Possession has weight; permission has direction.
Do not confuse the vessel with the right to pour it elsewhere.
Honing note. The material object and the copyright are different objects of governance.

U.S. copyright law states the distinction directly: ownership of copyright is distinct from ownership of the material object in which a work is embodied, and transfer of the object does not by itself transfer copyright rights. That principle is the legal core of the HONES rule that owning, receiving, or storing a file does not automatically answer whether it may be reproduced, adapted, distributed, displayed, or performed. The exclusive rights in section 106 are separate rights and are subject to limitations elsewhere in the statute. (17 U.S.C. §§ 106, 202)

3.1 One record, multiple legal and operational dimensions

A reliable asset record separates at least nine dimensions: asset identity; access basis; physical or file ownership; copyright status; rightsholder information; license or permission; publication status; privacy or confidentiality status; and contractual or technical restrictions. These fields should not be collapsed into a single 'legal' boolean. A museum may own a painting but not every copyright in a modern work. A user may own a photograph file but not be the photographer. A company may own copyright but license only a narrow display right. A federal government webpage can include third-party material even though works of the U.S. Government are generally not protected under Title 17; section 105 itself also contains current statutory qualifications for certain covered academic works. (17 U.S.C. § 105; 17 U.S.C. § 202)

3.2 Expression, facts, ideas, and systems

Copyright protects original expression fixed in a tangible medium; section 102(b) states that copyright protection does not extend to ideas, procedures, processes, systems, methods of operation, concepts, principles, or discoveries as such. This distinction creates productive routes. A researcher may often extract and independently state unprotected facts, measurements, concepts, or methods without reproducing the source's expressive wording or imagery. The distinction is not a license to paraphrase protectable expression closely, and other doctrines can still apply, but it is a reason to design fact-oriented extraction separately from expressive reproduction. (17 U.S.C. § 102(b))

3.3 Rights are composable

Rights can be divided by medium, territory, duration, exclusivity, and type of use. That means a workflow should record what permission actually covers: internal research, public web display, print, derivative adaptation, model training, redistribution, commercial use, or another domain. An open license can create affirmative reuse rights but may impose attribution, share-alike, noncommercial, or no-derivatives conditions depending on the license. Creative Commons therefore recommends preserving Title, Author, Source, and License

information where available and identifying modifications when material is adapted. (Creative Commons, "Recommended Practices for Attribution")

Record field	Question answered	Example state
asset_identity	What exact object is this?	sha256:... / persistent ID
access_basis	How did we get it?	public URL / user upload / subscription
ownership	Who owns this copy or object?	organization / user / unknown
copyright_status	Is protectable expression likely governed by copyright?	protected / public domain / unknown
rights_basis	Why may this proposed use proceed?	owned / licensed / statutory review
license	What conditions travel with reuse?	CC BY 4.0 / contract ID
privacy_status	Does the source concern identifiable people or secrets?	public / sensitive / confidential
contractual_restrictions	What non-copyright terms may bind access/use?	ToS / NDA / no-redistribution
permitted_actions	What action is actually supported?	internal inspect / public display

CONSTRUCTIVE ROUTE: When the desired action is unsupported, do not stop at "no." Search for a representation or source whose rights match the purpose: metadata instead of media, a link instead of a mirror, a licensed copy instead of an unlicensed copy, a public-domain equivalent, an open dataset, or an independently created substitute.

4. SOURCE AND ACCESS CLASSES

Visibility describes discovery; access class describes conditions

THE RAVEN REMEMBERS
The public square, the library desk, the locked room, the borrowed key:
each can contain the same sentence and a different obligation.
Do not name the content by the brightness of the doorway.
Name how the doorway was crossed.
Honing note. Source class records the conditions of encounter before they are forgotten.

Source and access classification answers a question that copyright status alone cannot answer: how did the system obtain the material, and what boundary surrounded that acquisition? The same copyrighted photograph might appear on the photographer's public site, in a licensed stock library, inside a private client folder, as a press asset distributed for editorial use, or as a leaked file. Those are not equivalent inputs even if the pixels are identical.

4.1 Canonical source classes

Class	Meaning	Default constructive route
PUBLIC_OPEN	Publicly reachable and accompanied by an affirmative open/public-domain basis that supports the intended operation.	Prefer direct source; preserve license/version.
PUBLIC_RESTRICTED	Publicly reachable, but rights or terms do not affirm broad reuse.	Link/metadata/facts first; review higher-intensity copying.
PUBLIC_LICENSED	Publicly reachable under a specific license or grant.	Enforce license scope and attribution.
SUBSCRIPTION	Available only through a paid subscription or institutional entitlement.	Treat access and redistribution separately.
AUTHENTICATED_COMMUNITY	Login or membership required, even if membership is broad.	Do not infer public redistribution rights.
USER_SUPPLIED	Provided by a user or collaborator.	Record uploader claim and intended purpose; do not assume ownership.
PRIVATE	Not generally public; shared for a bounded purpose.	Segregate, minimize, restrict downstream use.
CONFIDENTIAL	Subject to confidentiality obligation or reasonable confidentiality expectation.	Need-to-know processing; output redaction/review.
PROPRIETARY	Owned/controlled business or institutional material with limited use rights.	Contract and trade-secret review as appropriate.

Class	Meaning	Default constructive route
TRADE_SECRET	Material plausibly protected as secret commercial information.	Quarantine; counsel/security routing.
TECHNICALLY_RESTRICTED	Access controlled by DRM, paywall, encryption, access token, or technical gate.	Do not bypass merely because downstream use might be lawful.
UNKNOWN	Acquisition or source status cannot be established.	Preserve evidence; reduce use; investigate or substitute.

4.2 Publicly visible is not public domain

A browser can display protected works because display and access occur within legal, technical, contractual, and platform arrangements; visibility does not convert the work into the public domain. Conversely, some public materials do carry affirmative reuse rights. The efficient response is therefore not to treat 'web' as a rights category. The intake process should locate license statements, public-domain notices, rightsholder information, API terms, media-kit permissions, or other concrete signals and store the evidence rather than a guess.

4.3 Robots, terms, and technical boundaries

For automated web acquisition, RFC 9309 standardizes the Robots Exclusion Protocol as a way for service owners to state crawler preferences, and it explicitly says those rules are not a form of access authorization. That makes robots.txt important operational evidence but not a complete legal analysis. Authentication, contracts, copyright, privacy, rate limits, technical access controls, and computer-access law can remain separate questions. HONES records these signals independently rather than treating any one of them as a universal key. (Koster et al.)

EFFICIENCY RULE: Use the most authoritative source that is lawfully reachable and sufficiently complete. Prefer primary/original sources, official APIs, open repositories, or rightsholder-provided assets over mirrors when they provide equivalent evidence with clearer provenance.

5. RIGHTS-BASIS CLASSES

A decision should carry a reason that another system can inspect

THE RAVEN REMEMBERS
Permission without a name becomes rumor.
A license without a version becomes weather.
A conclusion without its basis cannot travel safely.
Write the reason beside the act, and the act can be judged again.
Honing note. The rights basis is a typed claim, not a mood.

A rights-basis class describes why a proposed use is thought to be supportable. This is deliberately different from copyright status. A work can be copyrighted while a particular use proceeds under license, permission, fair use, a statutory exception, or another basis. A public-domain work can still involve privacy, trademark, contractual, security, or authenticity issues. The rights-basis field therefore belongs to an action record, not merely to the asset record.

Rights basis	Meaning	Required record
OWNED	Organization/rightsholder owns the relevant exclusive right(s).	Confirm chain of title and scope.
EXPLICIT_PERMISSION	Identified rightsholder granted this use.	Store grant, scope, date, revocation/term.
LICENSED	Use covered by contract or platform/license terms.	Bind action to exact license/version.
OPEN_LICENSE	Public license affirmatively permits the action subject to conditions.	Carry attribution/share-alike/ND/NC conditions.
PUBLIC_DOMAIN_CONFIRMED	No applicable copyright protection for the work in the relevant jurisdiction or term.	Record reason and jurisdiction; check embedded third-party material.
STATUTORY_EXCEPTION	A specific statutory limitation supports the action.	Record provision and factual prerequisites.
FAIR_USE_REVIEW	U.S. fair-use analysis performed for this specific use.	Store four-factor analysis; no blanket inheritance.
CONTRACTUAL_RIGHT	Agreement affirmatively permits the use.	Store contract identifier and limitations.
EMPLOYMENT_OR_ASSIGNMENT	Rights arise from work-made-for-hire/assignment or similar arrangement.	Verify agreement and jurisdiction.

Rights basis	Meaning	Required record
RIGHTS_UNKNOWN	Relevant ownership/license status unresolved.	Use lower-intensity route or investigate.
REVIEW_REQUIRED	Material legal uncertainty remains.	Route to counsel/rights team with evidence packet.

5.1 Permission should be granular

Permission records should specify what is permitted, by whom, for how long, in which territories, for what media or systems, whether sublicensing is allowed, whether AI training or machine analysis is addressed, whether commercial use is covered, and what attribution or notice is required. The absence of a restriction in a casual message should not be converted into an affirmative grant of every conceivable downstream use. Equally, a narrow grant should be used fully within its scope rather than ignored because it is not universal.

5.2 Unknown is a usable state

Nuliometry's central discipline applies directly: missing is not zero. RIGHTS_UNKNOWN is neither 'allowed' nor 'forbidden.' It is an epistemic state. The workflow can still produce useful output by routing to lower-intensity actions, seeking permission, finding an alternative source, or creating an independent substitute. This makes uncertainty productive rather than paralyzing. (Conrad and OpenAI, Nuliometry)

RIGHTS RULE: Never store legal=true. Store the rights basis, evidence, jurisdiction, exact proposed action, reviewer or automation that made the classification, and the conditions that must remain true.

6. ACTION / USE CLASSES AND EFFICIENT ROUTING

Classify what the system intends to do, then choose the lightest sufficient path

THE RAVEN REMEMBERS
A link is not a mirror; a fact is not a paragraph;
a thumbnail is not the archive it points toward.
The wise tool does not carry the whole forest to prove a leaf.
It carries enough, and remembers where the forest stands.
Honing note. Efficiency begins by refusing to treat every research need as full-copy publication.

The action class is one of the central technical contributions of HONES. Rights analysis attaches to an act: reference, copy, transform, display, distribute, train, generate, or commercialize. The same source can support one action and not another. A web article may support a reference link and factual summary even when full mirroring is unsupported. A licensed image may support editorial publication but not merchandise. A private image may support local user-requested analysis but not corpus retention.

Use class	Operation	Typical objective	Governance intensity
REFERENCE_LINK	Point to source without copying expressive content.	Navigation, bibliography, provenance.	Very low
METADATA_ONLY	Store identifiers, dates, creator, rights fields, hashes.	Catalog/search.	Low
FACT_EXTRACTION	Extract unprotected facts/measurements and write independently.	Research/news.	Low to moderate
INDEX	Store search-oriented identifiers/tokens within policy.	Discovery.	Contextual
TEMPORARY_DECODE	Create ephemeral technical representation.	OCR/vision/transcoding.	Low if non-distributive
INTERNAL_INSPECTION	Human/model review within bounded environment.	Research/evaluation.	Contextual
SEARCH_OR_EMBED	Create retrieval/index vectors or searchable representation.	RAG/search.	Contextual; separate analysis
QUOTE_OR_EXCERPT	Use limited source expression.	Criticism/news/scholarship.	Fact-sensitive
THUMBNAIL	Reduced representation for identification/navigation.	Search/catalog.	Fact-sensitive
ARCHIVE	Retain canonical copy for preservation/research.	Preservation/audit.	Fact-sensitive
TRANSFORM	Create derived working representation.	Analysis/accessibility.	Fact-sensitive
MIRROR	Republish source or substantial equivalent.	Availability/backup.	High unless authorized

Use class	Operation	Typical objective	Governance intensity
PUBLIC_DISPLAY	Expose protected work publicly.	Publication/exhibition.	High/contextual
DISTRIBUTION	Provide copies to others.	Publishing/dataset release.	High
DERIVATIVE_CREATION	Create new expressive work based on protected expression.	Art/design/media.	High/contextual
MODEL_EVALUATION	Use examples to test model behavior.	QA/research.	Separate from training
RAG	Use source in retrieval-augmented response pipeline.	Answering/research.	Separate acquisition/output review
FINE_TUNING	Update model weights on selected corpus.	Adaptation.	High/contextual
FOUNDATION_TRAINING	Large-scale pretraining.	Model development.	High/contextual
SYNTHETIC_GENERATION	Generate new output using model/system.	Creative/analytic output.	Output-dependent
COMMERCIALIZATION	Exploit output/system in commerce.	Product/revenue.	Raises review intensity

6.1 The minimum sufficient representation

HONES defines the Minimum Sufficient Representation (MSR) as the least information-bearing representation that preserves the evidentiary or creative function required by the task. MSR is not a statutory doctrine. It is a project optimization rule. Its purpose is to reduce unnecessary copying, privacy exposure, attack surface, storage, and distribution while often improving auditability. It is especially valuable in automated systems, where copying the full source can become the accidental default simply because storage is cheap.

TEXT DIAGRAM 3 - POSITIVE ROUTING LADDER

```
NEED: identify source?       -> LINK + METADATA
NEED: answer factual query?   -> FACT EXTRACTION + CITATION
NEED: compare structure?      -> INTERNAL THUMBNAIL / MEASUREMENT
NEED: preserve evidence?      -> LAWFUL CANONICAL ARCHIVE + HASH
NEED: quote/comment?          -> NECESSARY EXCERPT + ANALYSIS + CITE
NEED: publish full media?     -> OWNED / LICENSED / PD / EXCEPTION REVIEW
NEED: train or tune?          -> CORPUS-SPECIFIC RIGHTS + OUTPUT CONTROLS
NEED: commercialize?         -> RECHECK RIGHTS + PRIVACY + MARKET + BRAND
```

6.2 Efficiency is not a universal risk ranking

A link is often lower-friction than a mirror, but there is no universal numerical ordering that decides legality. A small quotation can be more consequential than a larger factual table; a low-resolution image can still be the heart of a protected work; a private metadata record can reveal more sensitive information than a public photograph. Therefore HONES uses a route comparison rather than an automatic legal score. The system can rank operational cost and exposure, but material legal conclusions remain evidence-backed and reviewable.

PROJECT OPTIMIZATION Candidate Route Utility = Purpose Coverage x Evidence Fidelity x Reusability, considered against Rights Burden + Privacy Exposure + Security Exposure + Irreversibility + Distribution Surface + Review Cost. This is a qualitative comparison frame, not a calibrated equation and never a substitute for legal analysis.

ROUTING RULE: When two routes satisfy the purpose equally well, prefer the route that requires less copying, less sensitive data, less public exposure, clearer rights, stronger provenance, and greater reversibility.

7. RESEARCH, STUDY, CRITICISM, AND FAIR USE

Case-specific analysis, not a research exemption

THE RAVEN REMEMBERS
Research is a purpose, not a password.
Criticism is a reason, not a blank check.
Measure the taking against the reason for taking;
measure the new path against the market it may replace.
Honing note. Fair use is a balance carried by facts, not a label carried by intention alone.

Section 107 of the U.S. Copyright Act expressly lists criticism, comment, news reporting, teaching, scholarship, and research as examples of purposes that may qualify as fair use. It then requires consideration of four factors: purpose and character, nature of the copyrighted work, amount and substantiality used, and effect on the potential market or value. The statute does not create an automatic research privilege, and the Copyright Office stresses case-by-case analysis. (17 U.S.C. § 107; U.S. Copyright Office, "Fair Use Index")

7.1 Four factors as a structured worksheet

1. Purpose and character: What is the specific secondary use? Is it commercial? Does it have a genuinely further purpose or different character, and why is copying reasonably necessary to that purpose?
2. Nature of the work: Is the source highly creative or more factual/functional? Is it published or unpublished? What protectable expression is actually involved?
3. Amount and substantiality: How much was used, qualitatively and quantitatively, and is that amount reasonable for the stated purpose?
4. Market effect: Does the use substitute for the original or a traditional/reasonable licensing market? Does the output expose expressive material in a way that competes with or displaces the source?

The Supreme Court's 2023 Warhol decision is especially useful for output governance because it directs attention to the specific challenged use. The question was not Warhol's artistic significance in the abstract; it was the Foundation's commercial licensing of a Prince image for a magazine use that was highly similar in purpose to Goldsmith's licensing market. New meaning or style alone was not dispositive. HONES converts that lesson into a routing rule: analyze the actual transaction and output surface, not merely the internal transformation story. (Andy Warhol Found. for the Visual Arts, Inc. v. Goldsmith)

7.2 Precedents show why use classes matter

Search and digitization precedents illustrate a different configuration. In Authors Guild v. Google, the Second Circuit held that large-scale book scanning, search, and limited snippets were fair use in the system before it, emphasizing the transformative search function and restrictions that prevented users from reading substitute copies. In Perfect 10 v. Amazon, the Ninth Circuit found Google's thumbnail search use fair on the facts before it, treating the search function as highly transformative. These cases are powerful precedents for specific search-oriented uses; they are not general permissions to republish full works or to assume every machine-learning pipeline is equivalent to a search index. (Authors Guild, Inc. v. Google, Inc.; Perfect 10, Inc. v. Amazon.com, Inc.)

Recent AI cases make the need for narrow claims even clearer. In Thomson Reuters v. Ross Intelligence, a federal district court rejected fair use for copying Westlaw headnotes to build a competing non-generative legal research system; the opinion emphasized market competition and expressly noted the non-generative context. In Bartz v. Anthropic, another district court held certain training uses of lawfully acquired books fair and treated destructive digitization of purchased print books separately, while refusing to excuse the acquisition and retention of pirated copies merely because later training could be fair. The case later settled as to remaining claims. These decisions are fact-specific district-court authorities, not a universal AI rule. (Thomson Reuters Enter. Ctr. GmbH v. ROSS Intelligence Inc.; Bartz v. Anthropic PBC)

7.3 The HONES fair-use record

A system should never calculate FAIR_USE = TRUE from a few metadata fields. It should create FAIR_USE_REVIEW with the four-factor narrative, source/acquisition facts, jurisdiction, reviewer, date, known contrary considerations, and output controls. If the analysis depends on unsettled doctrine or a commercially consequential output, the machine-actionable result becomes LEGAL_REVIEW_REQUIRED. This preserves agency: the reviewer sees both the arguments supporting the use and the arguments against it, plus alternative routes.

FAIR-USE RULE: Use fair use as a reasoned, source-specific, action-specific review state. Never inherit it automatically from another project, another output type, another model stage, or another jurisdiction.

8. AI TRAINING, RETRIEVAL, TUNING, AND CORPUS CONSTRUCTION

Different computational uses require different records

THE RAVEN REMEMBERS
The corpus is not one act because it sits in one folder.
A model can read, remember, retrieve, imitate, or forget by different mechanisms.
Name the operation before naming the permission.
Then test what returns from the machine, not only what entered it.
Honing note. Training governance must describe both ingestion and the behavior of outputs.

Temporary inference over a user-provided image, embedding generation, RAG indexing, supervised fine-tuning, preference tuning, model evaluation, and foundation pretraining are materially different computational and legal activities. The U.S. Copyright Office's current AI study makes the same structural point: its May 2025 Part 3 pre-publication report says different uses during AI development and deployment require separate consideration, and the Office's AI page still identifies that Part 3 as pre-publication as of September 2026. (U.S. Copyright Office, Part 3 39; U.S. Copyright Office, "Copyright and Artificial Intelligence")

8.1 Current U.S. landscape: spectrum, not bright line

The Copyright Office's Part 3 report does not state that all training is fair or that all training is infringing. It describes a spectrum: some uses are expected to qualify as fair use and some are not. It identifies noncommercial research or analysis that does not enable reproduction of portions of works in outputs as comparatively favorable at one end, and copying expressive works from pirate sources to generate unrestricted competing content when licensing is reasonably available as comparatively unfavorable at the other. It also emphasizes source, purpose, output controls, and market effects. Courts remain the ultimate adjudicators, and facts and law continue to evolve. (U.S. Copyright Office, Part 3 77, 110)

8.2 EU-facing general-purpose AI

For providers of general-purpose AI models in the EU, Article 53 of the AI Act requires, among other things, a policy to comply with Union copyright law and related rights, including identifying and complying with relevant reservations of rights, and a publicly available sufficiently detailed summary of training content using the AI Office template. The European Commission's General-Purpose AI Code of Practice provides a voluntary compliance route whose Transparency and Copyright chapters address Article 53 obligations. Peer-reviewed commentary by the chairs of the Copyright Chapter explains the legal basis and operational measures, while the binding obligations remain those in applicable EU law. (Regulation (EU) 2024/1689 art. 53; European Commission, "General-Purpose AI Code of Practice"; Peukert and Castets-Renard)

8.3 Corpus records should resemble datasheets, not mystery folders

Datasheets for Datasets proposed documenting dataset motivation, composition, collection process, and recommended uses; model cards proposed documenting model uses, limitations, and evaluation conditions. HONES extends those ideas with acquisition and rights provenance. Every training or retrieval corpus should state source classes, rights bases, exclusions, deduplication policy, personal-data handling, known geographic/jurisdictional limits, opt-out or rights-reservation processing when applicable, transformations, version hashes, and the output controls expected to reduce unauthorized reproduction or exposure. (Gebru et al.; Mitchell et al.)

Output behavior must be tested, not assumed. CopyBench demonstrates that language models can reproduce protected text both literally and non-literally, and that mitigation strategies can affect those forms differently. HONES therefore treats memorization/reproduction testing as an output governance requirement for models whose training corpora contain protected expressive works. A system that has a careful intake ledger but no output evaluation is only half-governed. (Chen et al.)

AI operation	Minimum record	Constructive default
ONE-OFF INFERENCE	source/access/privacy + ephemeral processing record	Do not retain for training unless separately authorized.
EMBEDDING / INDEX	source identity + corpus purpose + retention + access	Keep retrieval surface bounded; honor source access changes where feasible.
RAG	retrieval source + excerpt policy + citation + output controls	Prefer answer synthesis with source links over wholesale reproduction.
MODEL EVALUATION	test-set rights + non-training separation + metrics	Use owned/open/licensed benchmarks when practical; preserve failures.
FINE-TUNING	corpus manifest + rights basis + model/version lineage	Use curated rights-clear corpus; test reproduction after tuning.
FOUNDATION TRAINING	large-scale provenance + rights-reservation handling + transparency	Dataset governance at collection scale; jurisdiction-specific review.
SYNTHETIC GENERATION	model/version/prompt/input provenance + output review	Check reproduction, likeness, brand, privacy, and implied endorsement.

AI RULE: Never label a source simply AI_OK. Record the exact operation, acquisition basis, corpus purpose, output controls, and jurisdiction. The lawful route can differ between evaluation, RAG, fine-tuning, and pretraining.

9. ACQUISITION MUST BE DISTINGUISHED FROM USE

A favorable downstream purpose does not legalize every path of entry

THE RAVEN REMEMBERS
The destination cannot rewrite the road behind it.
A worthy question does not turn a stolen key into permission.
Record the crossing as carefully as the answer.
Later use may be fair while earlier acquisition remains a separate wound.
Honing note. Acquisition and use are different events and can carry different legal states.

HONES treats acquisition as its own governed activity. This is not formalism. The source can determine the later risk analysis, as recent AI litigation illustrates. Bartz separated training from the creation and retention of a central library and treated pirate-source acquisition differently from digitization of purchased books. The Copyright Office Part 3 report similarly treats lawful access, source, and market conditions as facts that can matter in the overall analysis. (Bartz v. Anthropic PBC; U.S. Copyright Office, Part 3 55, 77)

9.1 Technical access controls

Section 1201 of the DMCA separately prohibits circumvention of technological measures that effectively control access to copyrighted works, subject to statutory provisions and periodically determined exemptions. A project should therefore never reason: 'the eventual use may be fair, so bypassing DRM or authentication is automatically permitted.' The anti-circumvention question has its own elements and exceptions. Where a research exemption or other statutory route may apply, record the exact basis and conditions; otherwise seek an authorized access path. (17 U.S.C. § 1201)

9.2 Crawling, APIs, terms, and rate boundaries

Automated acquisition should prefer supported interfaces and observable policy signals. RFC 9309 asks crawlers to honor robots.txt while making clear that robots rules are not access authorization. Platform terms, API conditions, authentication, rate limits, and permissions may create additional operational or contractual constraints. HONES stores robots status and terms version as evidence, not as a universal copyright verdict. If a public API, licensed feed, bulk export, or publisher-provided dataset supplies equivalent information with clearer terms, that is usually the more efficient route. (Koster et al.)

9.3 Acquisition ledger

- Original locator, referrer, repository, or physical source.
- Acquisition timestamp and system/agent.
- Authentication state and account/entitlement class, without storing secrets in the provenance record.
- Observed license, rights statement, terms version, robots signal, API policy, or access-control signal.
- Expected and actual cryptographic digest after transfer when a file is retained.
- Redirects, transformations, decompression, transcoding, or other steps between source and canonical object.
- Known revocation, expiry, embargo, or access-change conditions.
- Acquisition decision and alternate route considered if the preferred route was unavailable.

ACQUISITION RULE: Never use a favorable reuse theory to silently justify circumvention, credential sharing, authentication bypass, DRM removal, or unauthorized collection. Find an authorized route or explicitly route the acquisition question to review.

10. MIRRORING, CACHING, ARCHIVING, AND EPHEMERAL COPIES

Not every copy has the same function or lifetime

THE RAVEN REMEMBERS
The working shadow may disappear when the measurement is done.
The archive may remain because memory is the experiment.
A mirror is a promise to another audience.
Name the copy by what it is for, and give it the life it needs.
Honing note. Retention is a purpose decision, not a side effect of cheap storage.

Systems create copies for many reasons: network caching, format decoding, OCR, thumbnails, backup, disaster recovery, preservation, reproducibility, mirroring, user convenience, publication, or dataset distribution. HONES assigns a copy role and retention state because the rights, privacy, security, and market implications can differ sharply. A temporary PNG created to decode a HEIC file for local vision analysis is not the same product as a public mirror of that photograph.

10.1 RavenReader as a processing-projection example

RavenReader's current local architecture illustrates the distinction. The canonical artifact is identified by SHA-256 and retained unchanged. HEIC/HEIF and other formats can be decoded into ephemeral PNG working representations for OCR or vision; truncated images may be salvaged into an explicitly degraded processing projection while the original damaged bytes and negative evidence are preserved. The transformation improves machine compatibility without pretending that the source itself changed. That is HONES in miniature: source identity, representation class, transformation method, limitation, and retention are separate fields.

TEXT DIAGRAM 4 - CANONICAL VS PROCESSING COPY

```
CANONICAL SOURCE BYTES
SHA-256 = stable artifact identity
        |
        +-> metadata observation
        |
        +-> EPHEMERAL WORKING COPY (decode / resize / OCR)
        |          |
        |          +-> derived observations + method + version
        |
        +-> archival/retention decision under separate rights basis

WORKING COPY != NEW CANONICAL SOURCE
```

10.2 Mirroring is publication, not backup by another name

A mirror increases availability to others and can substitute for the original source. That is why a system should not automatically mirror a resource merely because it might disappear. Where preservation is justified, a lawful archive can be access-controlled and separately governed from public display. Where only citation stability is needed, a persistent identifier, source hash, bibliographic record, or permitted web archive link may provide the necessary evidence without republishing the source.

Libraries and archives also operate under specific statutory provisions, including section 108 in the United States, whose conditions should not be generalized to every private engineering repository. HONES can represent those statutory routes, but the organization must actually satisfy their prerequisites before relying on them. (17 U.S.C. § 108)

RETENTION RULE: Every retained copy gets a role: canonical, backup, cache, thumbnail, processing projection, archive, mirror, or publication copy. Role determines access, retention, deletion, and review; role changes create new lineage events.

11. CITATION, ATTRIBUTION, LINKING, AND EVIDENCE

Provenance gives credit and traceability; permission answers a different question

THE RAVEN REMEMBERS
A citation points backward; a license points forward.
One tells who spoke; the other tells what may be done.
Keep both, because memory without permission can trespass,
and permission without memory can erase the maker.
Honing note. Citation and permission are complementary records, never substitutes for one another.

Citation answers where a claim, quotation, dataset, image, codebase, or idea came from. Attribution gives credit under scholarly norms or license terms. Linking directs a reader to another resource. Permission or a statutory limitation answers whether a particular act is legally supportable. These functions overlap in practice but should remain typed separately. A citation cannot cure an otherwise unauthorized reproduction; conversely, lawful reuse can still be scientifically poor if it severs attribution and provenance.

11.1 Canonical citation record

A HONES citation record should include creator or responsible organization, title, publisher/platform/repository, publication or release date, original locator or persistent identifier, access date where relevant, version, license or rights statement, local hash when captured, quotation/excerpt location, and transformation history. When a source is mutable, the system should capture enough state to explain what version was actually observed.

Creative Commons recommends TASL - Title, Author, Source, License - as a practical attribution mnemonic and recommends indicating modifications. FAIR principles emphasize machine findability and reusability; software citation principles emphasize importance, credit and attribution, unique identification, persistence, accessibility, and specificity. Reproducible computational research likewise benefits from recording versions, parameters, and intermediate steps. HONES combines these scholarly practices with rights and security states so

provenance survives outside a paper's bibliography. (Creative Commons, "Recommended Practices for Attribution"; Wilkinson et al.; Smith et al.; Sandve et al.)

11.2 Prefer reference over duplication when reference is enough

The constructive default is to keep the reader connected to the primary source while writing independent analysis. For a news article, that can mean a concise factual synthesis plus links to the official filing and the source report. For code, cite the repository, commit or release, and license rather than pasting a large implementation. For a research dataset, cite a persistent identifier and version. For an artwork, a bibliographic record and authorized museum or artist page may satisfy identification when reproducing the image is unnecessary.

TEXT DIAGRAM 5 - CITATION STACK

```
CLAIM / OUTPUT
   |
   +-> SOURCE CITATION (who / what / where / version)
   +-> RIGHTS RECORD (why this action may proceed)
   +-> TRANSFORMATION RECORD (what changed)
   +-> EVIDENCE CLASS (observation / inference / policy / unresolved)
   +-> REVIEW STATE (automatic / human / legal / security)

No single layer substitutes for the others.
```

CITATION RULE: Prefer primary sources and persistent identifiers; preserve version and access evidence; cite what actually supports the claim; and do not use attribution language to imply permission that was never granted.

12. NEWS, REPORTING, AND PUBLIC-INTEREST RESEARCH

Facts can travel farther than the expressive packaging that first carried them

THE RAVEN REMEMBERS
The event is not owned because a sentence first described it.
The sentence can still be owned as the shape of its telling.
Carry the fact with witnesses; carry the quotation with measure.
Let the report open a path back to the record.
Honing note. Good reporting separates factual verification from expressive duplication.

News reporting is expressly named in section 107 as a purpose that may qualify for fair use, but the statute does not create an automatic news exemption. HONES therefore separates the factual layer from the expressive layer. Facts, dates, measurements, public proceedings, and independently observed events can often be restated in original language; photographs, articles, video clips, graphics, and expressive passages can require separate analysis. The most efficient reporting system extracts facts, retains the primary source trail, and reproduces expression only where the reporting purpose needs it. (17 U.S.C. § 107)

12.1 Source hierarchy for reporting

1. Primary evidence: statutes, court filings, official datasets, direct records, original research, transcripts, recordings, statements, and first-party materials.
2. Independent corroboration: reputable reporting or peer-reviewed analysis that confirms or contextualizes primary evidence.
3. Attributed claims: statements by interested parties, clearly labeled as claims rather than converted into neutral fact.
4. Commentary/opinion: analysis whose evaluative status remains visible.
5. Unverified or anonymous material: retained as a lead or evidence state, not upgraded to fact without appropriate corroboration.

12.2 Screenshot and clip discipline

Screenshots and clips are attractive because they preserve context, but they can also reproduce substantial protected expression, disclose private information, or omit surrounding material. The output review should ask whether the screenshot itself is necessary to demonstrate the fact or criticism. If a text quotation, timestamp, link, redacted crop, or independent description proves the point with less copying and less personal data, that route may be stronger. When the image itself is the object of criticism or verification, the necessity analysis changes.

12.3 Revision lineage

News changes. A source can be corrected, deleted, updated, or contradicted. HONES stores publication time, observed version, later revisions, correction notices, and supersession relationships. The current article can be corrected without erasing the evidence that an earlier version existed. This is the same lineage discipline used in RavenReader and W3C PROV: an output is an entity generated by activities using other entities, and its history can be represented rather than overwritten. (W3C, "PROV-DM")

NEWS RULE: Report the minimum necessary expressive material, maximize primary-source traceability, distinguish claim from fact, and preserve revisions. Public interest strengthens purpose; it does not eliminate source-specific rights, privacy, or verification duties.

13. ARTISTS, PERFORMERS, CHARACTERS, BRANDS, AND INTELLECTUAL PROPERTIES

Reference, study, imitation, endorsement, and replica are different operations

THE RAVEN REMEMBERS
A name can be a fact, a mark, a reputation, or a doorway.
A face can be evidence, identity, performance, or likeness.
Study the pattern without claiming the maker stood beside you.
Transformation is clearest when lineage and distance are both visible.
Honing note. Creative reference requires more than one rights lens.

Creative-source governance cannot be reduced to copyright alone. A project may implicate copyrighted expression, trademarks or trade dress, contractual rights, publicity or likeness interests, false endorsement, privacy, moral rights in some jurisdictions, and platform rules. HONES does not attempt to collapse these fields into one universal doctrine. It records which domains may be relevant and routes uncertain high-impact uses to specialized review.

13.1 Study is not the same as public derivative exploitation

An internal design study can examine color, composition, geometry, historical technique, material use, camera position, typography, or other observable features and translate them into abstract descriptors. Publicly generating and selling a close substitute can present a different set of questions. The efficient workflow therefore separates reference analysis from output generation and asks whether the output preserves too much protected expression, implies endorsement, reproduces a protected character or logo, or targets a living creator's market in a way that warrants further review.

Warhol reinforces the need to analyze the specific secondary use rather than relying on a general claim of new style or meaning. For AI systems, the Copyright Office Part 3 report likewise warns against treating all training or outputs alike. HONES therefore prefers descriptive design vocabulary - period, medium, geometry, lighting, material, composition, production technique - when that vocabulary can achieve the creative objective without requiring a high-fidelity replica of a specific protected work. (Andy Warhol Found.; U.S. Copyright Office, Part 3)

13.2 Performers and digital replicas

The Copyright Office's 2024 Part 1 AI report addresses realistic digital replicas that falsely depict or imitate individuals and recommends a federal digital-replica law because existing protections are fragmented. Whether or not a particular jurisdiction has adopted such legislation, HONES treats realistic voice, face, or performance simulation as a distinct likeness/replica class requiring consent or a documented legal/public-interest basis and output labeling where appropriate. Copyright clearance for a recording is not treated as automatic permission to create a realistic replica of the person. (U.S. Copyright Office, Part 1; U.S. Copyright Office, NewsNet 1048)

13.3 Brand and character handling

Brand names and logos can be necessary to identify products, sources, or subjects in research and reporting. That factual or referential use should be distinguished from using the same marks in packaging, merchandising, sponsorship-like presentation, or a way likely to imply affiliation. Characters can combine copyrighted visual/literary expression with trademark or merchandising concerns. The governance record should therefore state whether a brand or character appears as evidence, commentary, compatibility reference, parody, decorative borrowing, product identity, or commercial branding.

CREATIVE ROUTE: When the goal is learning, extract structure and technique. When the goal is identification, cite and reference. When the goal is expressive reuse, choose owned/open/licensed material or perform a use-specific review. Never imply participation, endorsement, authorship, or consent that the source does not support.

14. PRIVACY, PERSONAL DATA, MINORS, AND SENSITIVE SOURCES

Copyright clearance is not a privacy clearance

THE RAVEN REMEMBERS
A lawful copy can still expose a life.
A public face can still carry a private address behind it.
The smallest field can reveal the largest secret.
Use what the purpose needs, and let the rest remain dark.
Honing note. Data minimization is both a privacy principle and an efficiency principle.

Privacy analysis asks different questions from copyright analysis: who can be identified, what information is revealed, for what purpose it was collected, who expects access, what harms could follow, and how long the data should persist. NIST's Privacy Framework is designed to help organizations manage privacy risk while building innovative products and services; its SDLC guidance integrates privacy outcomes through planning, design, build/buy, deployment, operation, and decommissioning. That lifecycle alignment closely matches HONES's end-to-end routing goal. (NIST, "Privacy Framework"; NIST, "Using Privacy Framework 1.1")

14.1 Data minimization and purpose limitation

For EU personal data, GDPR Article 5 requires, among other principles, lawfulness/fairness/transparency, purpose limitation, data minimization, accuracy, storage limitation, and integrity/confidentiality. HONES uses those principles as explicit legal requirements where GDPR applies and as strong design precedents elsewhere: collect and retain only what the stated purpose requires; do not silently expand purpose; keep sensitive access controlled; and reassess retention when the purpose ends. (Regulation (EU) 2016/679 art. 5)

14.2 Minors and sensitive contexts

The presence of minors, intimate imagery, medical information, precise location, financial data, credentials, identity documents, communications, or other sensitive material raises the review level even when a file is technically public or copyright-cleared. A research index may not need the face, exact address, or raw message body. Redaction, tokenization, aggregation, local-only processing, shorter retention, and role-based access can preserve research utility while reducing exposure.

14.3 Private sources and AI

A private user file supplied for one analysis should not automatically become training data, an evaluation example, or a persistent shared corpus. Each downstream use gets its own purpose and rights/privacy basis. This separation also improves system design: one-off inference can occur in a short-lived processing zone; a deliberate training corpus can have stronger curation, consent, and documentation; a public output can pass through a distinct disclosure review.

PRIVACY RULE: Collect the minimum necessary representation for the declared purpose, segregate private/confidential material, and require a new decision when purpose, audience, retention, or model-use class changes.

15. SECURITY AND HOSTILE FILES

A source may be legitimate evidence and still be dangerous input

THE RAVEN REMEMBERS
The archive can carry a witness and a weapon in the same container.
Read the file without granting it a hand.
Let untrusted words remain data until trust is earned.
The safest interpreter is the one that cannot secretly become the operator.
Honing note. Safe to study and safe to execute are separate states.

External files are untrusted input. Copyright or research legitimacy does not make a file safe. A PDF can contain active content; an archive can expand far beyond its compressed size; a filename can attempt path traversal; a document can include hidden instructions that target an AI agent; a media parser can contain vulnerabilities. HONES therefore places security triage before rich parsing and prevents rights review from being used as a substitute for sandboxing.

15.1 File intake controls

OWASP's File Upload Cheat Sheet recommends defense in depth: allowlist necessary extensions, validate actual file type rather than trusting the Content-Type header, constrain filename and size, require authorization, store files away from directly executable web paths, and use malware analysis or content disarm/reconstruction where appropriate. Archive inspection should bound expansion and path traversal; image processing should enforce pixel and memory limits. HONES records which controls were actually applied, not merely which controls exist in policy. (OWASP, "File Upload")

15.2 Prompt injection and untrusted content

AI-enabled research introduces another boundary: documents, webpages, emails, code comments, or image text can contain instructions intended to manipulate an LLM or agent. OWASP describes indirect prompt injection as malicious instructions embedded in external content and recommends separating instructions from data, limiting tool privileges, validating outputs, monitoring actions, and using human approval for high-risk operations. HONES adds provenance tags to retrieved content so an agent can distinguish user intent and system policy from untrusted source text. (OWASP, "LLM Prompt Injection Prevention")

TEXT DIAGRAM 6 - HOSTILE SOURCE BOUNDARY

```
UNTRUSTED SOURCE
   |
   v
TYPE / SIZE / HASH / MALWARE / ARCHIVE LIMITS
   |
   v
QUARANTINED PARSER / READ-ONLY ANALYSIS
   |
   +-> extracted DATA (never implicit instructions)
   |
   v
POLICY ENGINE + LEAST-PRIVILEGE TOOLS
   |
   +-> human approval for destructive / external actions
   v
OUTPUT VALIDATION + LOGGED PROVENANCE
```

15.3 Secrets and executable content

Secrets discovered during research - API keys, passwords, private tokens, internal endpoints, personal identifiers - should be classified and redacted or routed to incident handling rather than copied into reports or model context. Downloaded binaries should not be executed simply to identify them. A 'research source' may be opened through a safe decoder, disassembler, sandbox, or static analysis workflow without granting it the privileges of the research workstation.

SECURITY RULE: Treat external content as evidence, not authority. Parse with least privilege; separate data from instructions; restrict tools; bound resources; preserve hashes; and quarantine before execution.

16. TRANSFORMATION LINEAGE AND PROVENANCE

Every derived object should be able to answer what changed and why

THE RAVEN REMEMBERS
The derivative is not less true because it changed;
it becomes less trustworthy when the change disappears.
Write the tool beside the cut, the model beside the inference,
and the child can point back to the parent without pretending to be it.
Honing note. Lineage is the memory that lets transformation remain honest.

Provenance is not only a bibliography. W3C PROV models entities, activities, and agents, including derivation and responsibility. C2PA Content Credentials provide a technical architecture for assertions, signed claims, content bindings, ingredients, and action history, and version 2.4 (April 2026) continues to extend that ecosystem. These standards demonstrate a shared principle: transformations should be represented as relationships rather than overwriting history. (W3C, "PROV-DM"; C2PA, "Content Credentials," ver. 2.4)

16.1 Source -> activity -> derivative

TEXT DIAGRAM 7 - PROVENANCE GRAPH

```
ENTITY E0: original source [sha256 A]
   | used by
ACTIVITY A1: decode [tool/version/params]
   | generated
ENTITY E1: working projection [sha256 B, ephemeral]
   | used by
ACTIVITY A2: OCR / model inference [model/version/prompt]
   | generated
ENTITY E2: observation record [confidence + evidence class]
   | used by
ACTIVITY A3: human synthesis / publication
   | generated
ENTITY E3: report/output [sha256 C]

AGENTS attach responsibility to A1-A3. Rights/privacy states attach to E0-E3 and uses.
```

The minimum transformation record should include source hash or persistent identifier, tool/model and version, parameters or configuration needed to reproduce the transformation, actor or agent, timestamp, input rights and access state, output rights and access state, information removed, information introduced, fidelity loss, synthetic additions, uncertainty, review state, and whether the derivative is independently redistributable. A changed filename is not a new provenance theory; a new hash identifies a new byte-level entity whose relation to the source should be explicit.

16.2 Provenance is not proof of truth

Cryptographic provenance can support integrity and traceability, but it does not guarantee that an assertion was correct when created. C2PA itself frames assertions as trust signals and provides mechanisms for verification of claims and bindings. HONES therefore distinguishes integrity ('this claim was not altered') from truth ('the claim accurately describes reality'), rights ('this use is permitted'), and authenticity ('this content genuinely originated as asserted'). Those questions can inform one another without collapsing.

16.3 Reproducibility and efficiency

Good provenance is also economically useful. Reproducible computational research benefits from version and parameter capture because it reduces future reconstruction cost. FAIR principles make data easier for machines and people to find and reuse. HONES adds a further efficiency benefit: when source and transformation states are already structured, future teams can often reuse the existing lawful projection or cite the canonical record instead of reacquiring and reprocessing the entire source. (Sandve et al.; Wilkinson et al.)

LINEAGE RULE: No transformation without a parent reference; no parent replacement without a supersession event; no inference promoted to source fact; no provenance system marketed as a truth oracle.

17. OUTPUT GOVERNANCE

The last mile can create a new rights, privacy, and truth state

THE RAVEN REMEMBERS
A careful intake can still end in a careless release.
The mouth of the system needs a memory of the hands that fed it.
Before the output travels, ask what it reveals, replaces, and implies.
Then label the path so the next reader can sharpen it again.
Honing note. Output review is not redundant with input review because transformation can create new risk.

Output governance asks what is being released, to whom, at what fidelity, with what citations and disclosures, and whether the output reveals protected expression, private information, secrets, licensed-only assets, close replicas, or misleading provenance. The answer can differ from the intake state. A source lawfully used for internal research might still produce an output that substitutes for the source, exposes confidential material, or implies endorsement. HONES therefore gives every public output its own review object.

17.1 Text output

Text systems should distinguish factual synthesis, quotation, close paraphrase, and generated continuation. Quotation should be bounded by purpose and cited. Factual synthesis should be written independently. Systems working from copyrighted corpora should test for memorized or near-verbatim output where that risk is material. CopyBench shows why literal-only filters are insufficient: non-literal event

and character copying can also occur. The exact legal significance is context-specific, but technical evaluation can identify outputs that deserve stronger review. (Chen et al.)

17.2 Image, audio, and video output

Visual and audio outputs should carry model/version and input provenance where available and should be reviewed for close source reproduction, logos and branding, private or sensitive content, and realistic digital replicas. C2PA offers one interoperable way to attach content provenance; HONES can export to C2PA while retaining a deeper internal ledger. Where a system cannot embed credentials, an external manifest and stable output hash can still preserve lineage. (C2PA, "Content Credentials")

17.3 Human authorship and AI assistance

The U.S. Copyright Office's 2025 Part 2 report states that AI-assisted works can be copyrightable where a human author determines sufficient expressive elements, while purely AI-generated material lacks human authorship for U.S. copyright protection; prompts alone do not necessarily establish authorship of the output. HONES records human and machine contributions rather than hiding either. The purpose is not only registration analysis but truthful production lineage. (U.S. Copyright Office, Part 2)

17.4 Output classes

Output class	Audience	Default controls
INTERNAL_ONLY	authorized project team	source access rules; no public links; full provenance
REVIEW_ONLY	rights/privacy/security/editorial reviewers	watermark/status; no onward distribution
PUBLIC_FACTUAL	general audience	independent prose; citations; no unnecessary source expression
PUBLIC_EXCERPTED	general audience	bounded excerpt/media; purpose statement; citation/attribution
PUBLIC_LICENSED	general audience	license conditions, credits, notices, territory/term checks
PUBLIC_SYNTHETIC	general audience	AI provenance/disclosure as appropriate; replica/reproduction review
DATASET_RELEASE	research/public	rights manifest, documentation, exclusions, version/hash, redistribution terms
MODEL_RELEASE	users/developers	model card, training summary where required, limitations, output tests, license

OUTPUT RULE: A permissible input does not guarantee a permissible output. Re-evaluate audience, fidelity, substitution, attribution, privacy, likeness, branding, and disclosure at the release boundary.

18. GOVERNANCE, REVIEW, REPAIR, AND TRANSMISSION

A system should know when to proceed, when to reroute, and when to ask

THE RAVEN REMEMBERS
The strongest gate is not the one that never opens.
It is the gate that knows the road, the traveler, and the load.
When the path is uncertain, mark uncertainty instead of inventing certainty.
When the path is blocked, search the valley for another honest crossing.
Honing note. Governance protects agency by making options and conditions visible.

The final HONES layer turns the preceding analysis into a deployable protocol for people and AI agents. Every consequential asset receives a source class, access state, rights state, use class, privacy state, security state, provenance record, and intended output class. Automation may proceed where the intersection is clearly supported. Unknown rights, conflicting licenses, protected access controls, sensitive private material, high-fidelity public reproduction, commercial republishing, realistic replicas, or unresolved jurisdictional questions trigger review rather than an invented conclusion.

18.1 Decision states

Decision	Meaning	Constructive next move
PROCEED	Current evidence supports the specified action without extra conditions.	Execute and record lineage.
PROCEED_WITH_CONDITIONS	Action is supportable if listed conditions remain true.	Bind conditions to job/output.
ROUTE_ALTERNATIVE	Desired action is inefficient or unsupported but another route serves the purpose.	Use link/metadata/excerpt/open substitute/local analysis/etc.
REVIEW_REQUIRED	Material legal/privacy/security/editorial uncertainty remains.	Package evidence and questions for human specialist.
HOLD_FOR_RIGHTS	Use may become supportable with permission/license/clarification.	Seek grant or locate substitute.
DO_NOT_PROCESS	Acquisition or processing itself is unsupported/high-risk.	Quarantine/delete as policy requires; retain minimal audit record if lawful.
DO_NOT_PUBLISH	Internal handling may be supported but public release is not.	Keep bounded; produce sanitized or independently written output.

18.2 End-to-end route

TEXT DIAGRAM 8 - BEGINNING TO END

```
DISCOVER
  -> INTAKE + HASH + SECURITY TRIAGE
  -> SOURCE / ACCESS CLASS
  -> RIGHTS + PRIVACY + CONTRACT STATE
  -> DEFINE SPECIFIC PURPOSE
  -> GENERATE CANDIDATE USE ROUTES
  -> CHOOSE MINIMUM SUFFICIENT REPRESENTATION
  -> PROCESS / TRANSFORM / ANALYZE
  -> RECORD PROVENANCE + CITATIONS
  -> OUTPUT-SPECIFIC REVIEW
  -> PUBLISH / STORE / LINK / SUPPRESS / ROUTE TO REVIEW
  -> RETENTION / SUPERSESSION / AUDIT / REPAIR
```

18.3 Repair instead of erasure

Corrections should preserve earlier states when lawful and operationally appropriate. If a license was misclassified, create a superseding rights record. If a model confidence was incorrectly stored as zero instead of unknown, retain the historical observation and add the corrected canonical interpretation. If a source later becomes unavailable, preserve the last verified citation state and mark availability separately. This is the governance equivalent of RavenReader's negative-evidence discipline: failure and uncertainty remain part of the record rather than being rewritten as success.

18.4 Human agency and escalation

The point of automation is not to replace judgment with opaque scoring. It is to move routine, well-labeled cases quickly and to make hard cases arrive at a reviewer with the evidence already organized. A rights reviewer should receive the exact source, proposed action, audience, amount, acquisition path, license evidence, fair-use worksheet if applicable, alternatives considered, and business purpose. A privacy reviewer should receive only the personal-data facts needed for that decision. A security reviewer should receive the file evidence without unnecessary copyrighted content. Good labels reduce both risk and review cost.

FINAL DOCTRINE: Preserve the source; identify the right; define the purpose; choose the minimum sufficient representation; secure the processing path; document the transformation; cite the lineage; test the output; disclose uncertainty; and never silently upgrade permission, provenance, or truth.

CONCLUSION: THE USEFUL PATH

THE RAVEN REMEMBERS
The raven hones no blade to keep it sheathed forever.
The edge exists to make a cleaner cut between what is known and what is claimed.
Carry less when less is enough; carry more when the work truly requires it.
Leave a trail sharp enough that the next mind can find the same path - or correct it.
Honing note. The measure of governance is not how much it stops, but how reliably it converts uncertainty into accountable useful action.

The Raven Hones is designed around a practical observation: information work is rarely a choice between unrestricted copying and complete abstention. Between those extremes are links, identifiers, metadata, factual extraction, bounded quotation, internal inspection, thumbnails, licensed assets, public-domain sources, open datasets, local transformations, access-controlled archives, rights-cleared training sets, output filters, attribution, review, and independent creation. A mature system should know these routes and explain why it chose one.

The same framework also recognizes that some boundaries are real. A desired commercial reproduction may require a license. A private source may not belong in a training corpus. A technical protection measure may not be bypassed. A realistic replica may require specialized consent and legal review. A source may be too damaged to support a factual claim. Constructive governance does not deny those limits; it makes them legible early enough that the project can adapt.

The most valuable property of HONES is therefore not restriction but continuity. Source identity continues into transformation. Rights conditions continue into output. Citations continue into publication. Negative evidence continues into repair. Human judgment continues into automation through explicit states rather than hidden assumptions. Information becomes sharper because its history remains attached.

APPENDIX A - RAVEN_HONES_RECORD_v1

STATUS Project-defined schema. This appendix is an operational proposal, not a legal standard. Fields are designed to be extensible and versioned.

SCHEMA

```
{
  "schema": "RAVEN_HONES_RECORD_v1",
  "record_id": "...",
  "source": {
    "identity": {"sha256": "...", "persistent_id": "..."},
    "locator": "...",
    "creator": "...",
    "publisher_or_platform": "...",
    "acquired_at": "RFC3339",
    "acquisition_method": "download|api|user_upload|physical_scan|other"
  },
  "access": {
    "class": "PUBLIC_OPEN|PUBLIC_RESTRICTED|PUBLIC_LICENSED|SUBSCRIPTION|AUTHENTICATED_COMMUNITY|
USER_SUPPLIED|PRIVATE|CONFIDENTIAL|PROPRIETARY|TRADE_SECRET|TECHNICALLY_RESTRICTED|UNKNOWN",
    "authentication_required": false,
    "robots_state": "allowed|disallowed|not_applicable|unknown",
    "terms_version": "...",
    "technical_controls": []
  },
  "rights": {
    "copyright_status": "protected|public_domain|mixed|unknown",
    "rightsholder": "...",
    "license": {"id": "...", "version": "...", "evidence": "..."},
    "rights_basis": "OWNED|EXPLICIT_PERMISSION|LICENSED|OPEN_LICENSE|PUBLIC_DOMAIN_CONFIRMED|
STATUTORY_EXCEPTION|FAIR_USE_REVIEW|CONTRACTUAL_RIGHT|EMPLOYMENT_OR_ASSIGNMENT|RIGHTS_UNKNOWN|
REVIEW_REQUIRED",
    "jurisdiction": ["..."],
    "expiration": null,
    "reservations_or_opt_outs": [],
    "permitted_actions": []
  },
  "purpose": {
    "declared_goal": "...",
    "use_class": "...",
    "audience": "internal|review|public",
    "commercial_context": "...",
    "minimum_sufficient_representation": "..."
  },
  "person_privacy": {
    "identifiable_person": false,
    "likeness_or_voice": false,
    "minor": false,
    "sensitive_data": [],
    "consent_or_other_basis": "...",
    "retention_class": "..."
  },
  "security": {
    "trusted_input": false,
    "malware_scan": "pass|fail|not_run",
    "parser_profile": "sandboxed|quarantined|direct",
    "prompt_injection_state": "screened|not_applicable|review",
    "secret_scan": "..."
  },
  "transformation": [{
    "activity_id": "...",
    "tool": "...", "version": "...", "parameters": {},
    "input_hash": "...", "output_hash": "...",
    "information_removed": [], "information_added": [],
    "fidelity_limitations": [], "uncertainty": "..."
  }],
  "output": {
    "class": "INTERNAL_ONLY|REVIEW_ONLY|PUBLIC_FACTUAL|PUBLIC_EXCERPTED|PUBLIC_LICENSED|PUBLIC_SYNTHETIC|
DATASET_RELEASE|MODEL_RELEASE",
    "attribution_required": true,
```

```
    "citation_required": true,
    "provenance_disclosure": "...",
    "redistribution_allowed": "yes|no|conditional|unknown"
  },
  "decision": {
    "state": "PROCEED|PROCEED_WITH_CONDITIONS|ROUTE_ALTERNATIVE|REVIEW_REQUIRED|HOLD_FOR_RIGHTS|
DO_NOT_PROCESS|DO_NOT_PUBLISH",
    "conditions": [], "alternatives_considered": [], "rationale": "..."
  },
  "lineage": {
    "reviewer_or_agent": "...", "reviewed_at": "RFC3339",
    "evidence_refs": [], "supersedes": null, "version": 1
  }
}
```

APPENDIX B - SOURCE / ACCESS CLASS QUICK REFERENCE

Class	Default assumption	Efficient first route	Escalation trigger
PUBLIC_OPEN	Affirmative reuse path appears available.	Use primary source; carry license/provenance.	License ambiguity or third-party components.
PUBLIC_RESTRICTED	Visible, but no broad reuse grant.	Link + metadata + facts.	Need expressive reproduction/distribution.
PUBLIC_LICENSED	Use allowed within stated scope.	Apply exact license conditions.	New medium/territory/use beyond scope.
SUBSCRIPTION	Entitlement grants access, not automatically redistribution.	Internal analysis/citation.	Bulk extraction, sharing, training, mirroring.
AUTHENTICATED_COMMUNITY	Community access boundary exists.	Internal/reference use.	Public output or automated collection.
USER_SUPPLIED	Uploader supplied bytes for a purpose.	Process for declared purpose.	Training/republication/third-party rights.
PRIVATE	Not public.	Local/segregated minimum-use processing.	New audience/purpose.
CONFIDENTIAL	Confidentiality expectation/obligation.	Need-to-know only.	Any external disclosure.
PROPRIETARY	Institution controls/use is bounded.	Contract-defined workflow.	External distribution or model reuse.
TRADE_SECRET	Secrecy is central to value.	Quarantine + specialist review.	Almost any external or cross-team transfer.
TECHNICALLY_RESTRICTED	Technical gate controls access.	Authorized interface/license.	Any bypass/circumvention.
UNKNOWN	Evidence insufficient.	Investigate or choose substitute.	High-impact use before clarification.

APPENDIX C - RIGHTS-BASIS EVIDENCE PACKET

- OWNED: chain of title, work-made-for-hire record, assignment, or internal ownership statement.
- EXPLICIT_PERMISSION: rightsholder identity, grant text, date, scope, territory, duration, medium, revocation conditions.
- LICENSED / OPEN_LICENSE: exact license text/version, source locator, conditions, attribution, modification and redistribution terms.
- PUBLIC_DOMAIN_CONFIRMED: jurisdiction, reason (term, government work, dedication, other), embedded third-party content check.
- STATUTORY_EXCEPTION: statute/provision, prerequisites, factual record, scope, expiration if temporary.
- FAIR_USE_REVIEW: four-factor analysis for the exact use, amount, source nature, market/substitution analysis, acquisition facts, reviewer/date.
- CONTRACTUAL_RIGHT: agreement ID, relevant clause, parties, term, scope, confidentiality and sublicensing conditions.
- RIGHTS_UNKNOWN: evidence searched, unresolved question, allowed lower-intensity routes, follow-up date.
- REVIEW_REQUIRED: precise question, business/research need, proposed output, alternatives considered, deadline and responsible reviewer.

APPENDIX D - USE-CLASS ROUTING MATRIX

Goal	Preferred low-friction route	Escalate when	Typical evidence retained
Identify a work	REFERENCE_LINK + METADATA_ONLY	identity disputed or link unstable	creator/title/source/license/hash if captured

Goal	Preferred low-friction route	Escalate when	Typical evidence retained
Answer a factual question	FACT_EXTRACTION + CITATION	fact inseparable from necessary expressive evidence	source/date/quotation only if needed
Search a collection	INDEX / SEARCH_OR_EMBED	source terms/privacy restrict indexing or output	corpus manifest, access class, index version
Compare visual structure	INTERNAL_INSPECTION / THUMBNAIL	public reproduction needed	working projection + measurements + source link
Critique/comment	QUOTE_OR_EXCERPT	amount/substitution/market concerns material	fair-use worksheet + citations
Preserve fragile evidence	ARCHIVE with bounded access	retention itself lacks basis	hash, acquisition, retention basis, access log
Republish full source	LICENSED / OWNED / PD first	no affirmative basis	license/grant or review packet
Model evaluation	rights-clear benchmark + isolated test set	dataset includes sensitive/protected material	dataset card, no-train marker, metrics
RAG answers	bounded retrieval + citation + reproduction guard	retrieval exposes large source portions	retrieval logs, source IDs, output checks
Fine-tune model	curated rights-clear corpus	rights status mixed/unknown	corpus manifest, exclusions, model lineage
Foundation train	large-scale rights/provenance program	jurisdictional/opt-out/licensing questions	source families, rights reservation handling, transparency docs
Commercialize output	output-specific rights/privacy/brand review	market substitute, replica, licensed-only input	release review, provenance, licenses

APPENDIX E - FAIR-USE REVIEW WORKSHEET (U.S.)

IMPORTANT This worksheet structures evidence; it does not decide fair use automatically. Fair use is a legal conclusion based on all facts and factors.

Field	Required narrative
IDENTIFY THE USE	What exact act is being evaluated - copying, display, excerpt, search index, training stage, RAG, output publication, other?
PURPOSE / CHARACTER	What further purpose or different character is served? Is use commercial? Why is copying reasonably necessary? Is there a lower-copy route that serves equally well?
NATURE	How creative/expressive versus factual/functional is the source? Published or unpublished?
AMOUNT	What percentage and qualitative heart is used? Is the amount reasonable for the purpose? What reaches the audience?
MARKET	Does the use or output substitute for the source or a traditional/reasonable licensing market? Does it expose competing expressive content?
ACQUISITION	How was the source obtained? Was access authorized? Any DRM/authentication/terms/robots signal?
OUTPUT CONTROLS	Are excerpts, thumbnails, retrieval limits, anti-regurgitation tests, access controls, or other safeguards present?
ALTERNATIVES	Could licensed/open/public-domain/owned material, linking, metadata, or independent description accomplish the purpose?
JURISDICTION / PRECEDENT	Which court/jurisdiction governs? Which precedents are actually analogous, and what factual differences matter?
DECISION	FAIR_USE_REVIEW supported / LEGAL_REVIEW_REQUIRED / choose alternate route. Record reviewer/date and contrary considerations.

APPENDIX F - AI CORPUS INTAKE CHECKLIST

- Define operation: evaluation, embedding/indexing, RAG, fine-tuning, pretraining, preference training, synthetic augmentation, other.
- Define purpose and expected output domain before collection begins.
- Record source families and acquisition methods; prefer primary/open/licensed/owned corpora when they satisfy the purpose.
- Capture rights statements, license versions, reservations/opt-outs where applicable, and jurisdictional scope.
- Separate public, authenticated, subscription, user-supplied, private, confidential, and unknown sources.

- Perform personal-data and sensitive-data review; document minimization, de-identification, or exclusion.
- Deduplicate and version the corpus; preserve source identity without requiring public redistribution of source bytes.
- Create a dataset datasheet: motivation, composition, collection period, filtering, transformations, limitations, recommended and disallowed uses.
- Create or update model card: intended uses, evaluation, limitations, known reproduction/memorization behavior, relevant population/domain constraints.
- Test literal and non-literal reproduction where protected expressive corpora are material to the model.
- Separate no-train evaluation datasets from training corpora through technical controls.
- For EU GPAI obligations where applicable, maintain Article 53 copyright policy and training-content summary processes.
- Record model/version lineage, checkpoints, corpus versions, and post-training stages so uses can be evaluated separately.
- Plan deletion/supersession and correction mechanics before training so later rights changes can be operationalized where technically feasible.

APPENDIX G - CITATION AND ATTRIBUTION TEMPLATE

TEMPLATE

```
CREATOR / RESPONSIBLE ORGANIZATION:
TITLE:
CONTAINER / PUBLISHER / PLATFORM:
DATE PUBLISHED / RELEASED:
VERSION / REVISION:
PERSISTENT ID / URL:
DATE ACCESSED:
LICENSE / RIGHTS STATEMENT:
ATTRIBUTION NAME REQUESTED:
MODIFICATIONS MADE:
LOCAL SOURCE HASH (if captured):
QUOTE / EXCERPT LOCATION:
ARCHIVE / SNAPSHOT ID:
RELATED PROVENANCE RECORD:
NOTES / LIMITATIONS:
```

For Creative Commons material, preserve the elements supplied by the licensor and use Title, Author, Source, and License as a practical check; note modifications. For software and datasets, cite versions or releases specifically enough that another researcher can identify the object actually used. (Creative Commons; Smith et al.)

APPENDIX H - LIKENESS / DIGITAL-REPLICA REVIEW

Question	Low-friction route if concern exists
Is an identifiable person depicted or imitated?	Use factual description, authorized media, or non-identifying representation if identity is unnecessary.
Is the person a minor?	Heightened privacy/consent review; minimize identity and distribution.
Does output realistically simulate face, voice, performance, or conduct?	Seek consent/license or specialist review; label synthetic nature where appropriate.
Could the output imply endorsement or participation?	Add clear context; avoid brand/likeness placement that suggests sponsorship.
Is the source private or sensitive?	Local-only or access-controlled processing; redact/minimize.
Is the use newsworthy, critical, satirical, documentary, or otherwise public-interest related?	Record purpose and jurisdiction-specific legal basis; use only what serves the purpose.
Are performer/copyright rights in source recording separately cleared?	Do not infer likeness permission from copyright/license alone.

APPENDIX I - FILE SECURITY INTAKE PROCEDURE

1. Receive into a non-executable quarantine location; preserve original filename as metadata, not as a trusted storage path.
2. Compute cryptographic hash and byte size before rich parsing.
3. Apply extension allowlist/business-need policy and validate magic/type independently of supplied MIME header.
4. Enforce file, pixel, archive-depth, decompressed-size, and time/memory limits.
5. Scan or statically analyze with appropriate malware/security tooling; use content disarm/reconstruction where policy requires.

6. Parse using sandboxed, least-privilege libraries/processes; keep parsers and codecs patched.
7. Treat document/web/image text as untrusted data for LLMs; do not let embedded instructions redefine system policy or tool calls.
8. Secret-scan extracted text before logging, model context expansion, or publication.
9. Write derived observations separately from source bytes; never overwrite canonical evidence during normalization.
10. Promote from quarantine only after security state and intended use are known; retain failed scans as negative evidence.

APPENDIX J - NEWS-SOURCE VERIFICATION WORKSHEET

Field	Record
Claim	Exact proposition the output intends to state.
Primary source	Original filing/data/statement/record if available.
Secondary corroboration	Independent reputable source(s).
Source date/time	Publication and observed revision time.
Status	fact / attributed claim / analysis / opinion / unknown.
Excerpt necessity	Why any copied text/image/clip is needed.
Amount used	Words/seconds/crop/portion.
Rights/public-interest basis	License / public domain / fair-use review / other.
Privacy/sensitivity	Redactions, minors, location, personal data.
Corrections	Known later correction/retraction/supersession.
Citation trail	Links, hashes, archive IDs, quoted locations.

APPENDIX K - TRANSFORMATION / PROVENANCE LEDGER

Field	Example
activity_id	ACT-20260929-0001
input_entity	sha256:abc...
activity	pillow-heif -> PNG ephemeral decode
tool/version	pillow-heif x.y / Pillow x.y
parameters	frame=0; orientation=EXIF; RGB
agent	ravenreader-worker-0.1.4
output_entity	sha256:def... or ephemeral ID
information_removed	HEIC container metadata not carried into working copy
information_added	none / white alpha background / OCR observation
fidelity_limitations	working projection only; canonical source unchanged
rights state	internal analysis under source record RH-...
privacy state	private/local
retention	delete working copy on context exit
validation	output decoded successfully; source hash reverified
supersedes	null

APPENDIX L - STOP / REVIEW / PROCEED CARD

ONE-PAGE FIELD CARD

```
PROCEED
- Source and acquisition are known.
- Proposed action has a documented rights basis.
- Privacy/security controls match the source.
- Minimum sufficient representation selected.
- Output class and citations are defined.
- Conditions are machine-readable and will travel with the job.

PROCEED WITH CONDITIONS
- License/permission requires attribution, territory, term, access, notice, or other limits.
- Output must remain internal, excerpted, redacted, low-fidelity, or access-controlled.
- Transformation or AI use is supported only for the specified stage.

ROUTE ALTERNATIVE
- Link instead of mirror.
- Metadata/facts instead of expressive copy.
- Thumbnail/measurement instead of full-resolution publication.
- Open/licensed/owned/public-domain substitute.
- Local analysis instead of corpus retention.

REVIEW REQUIRED
- Rights unknown or conflicting.
- Fair use materially uncertain or commercially consequential.
- Technical access control / circumvention question.
- Private/sensitive/minor/replica/trade-secret issue.
- High-fidelity public reproduction or market substitution concern.

STOP
- Source integrity fails and the task requires reliable content.
- Acquisition or processing is unsupported and no exception/permission applies.
- Malware/hostile content cannot be safely contained.
- Publication would expose protected private/confidential material without a basis.

ALWAYS
Preserve evidence. Record why. Offer the strongest constructive alternative.
```

APPENDIX M - END-TO-END USE CASES

Scenario	Efficient constructive route	Avoid as default
Copyrighted news article needed for a report	Acquire through normal authorized access; capture citation metadata; extract verified facts; write independent synthesis; quote only what serves reporting/criticism; link primary source.	Full article mirror unless separately supported.
Artwork needed to identify an artist or work	Use metadata, authoritative museum/artist page, source link, and internal thumbnail if needed for matching.	High-resolution public reproduction where identification alone suffices.
Image needed for visual research	Hash/canonicalize if lawfully retained; use local working projection; extract composition/color/shape descriptors; store observations and source citation.	Public redistribution of source pixels without basis.
Private user photo analyzed by AI	Process locally/segregated for stated request; delete ephemeral derivatives; do not add to training corpus without separate authorization.	Silent secondary training or public examples.
Need training examples for a new model	Search owned, licensed, public-domain, open, commissioned, or consented corpora first; document gaps; curate source/rights manifest; test output reproduction.	Mixed mystery corpus with no acquisition lineage.
Need a quotation from a book	Use the amount necessary to analyze or support claim; cite edition/page; consider license or fair-use review for extensive excerpts.	Treating research intent as unlimited copying.
Fragile webpage may disappear	Preserve citation metadata/hash and use an authorized archive route where available; retain private evidentiary capture only if supported.	Automatically publish a full mirror.
Brand logo appears in product research	Keep logo as evidence/identification when needed; separate that from using logo as your own branding or endorsement signal.	Repurposing source brand identity for merchandise without review.
Model output resembles source closely	Hold output; compare against source; record model/input lineage; regenerate, transform, license, or review before public release.	Publishing because output is technically generated.

Scenario	Efficient constructive route	Avoid as default
Rights unknown but deadline is short	Publish facts, citations, links, or an independent/open substitute; mark unresolved rights and continue review.	Guessing permission or dropping the research entirely when a lower-intensity route exists.

APPENDIX N - VERIFIED SOURCE LEDGER AND EVIDENCE CLASSES

Class	Source	Use in this paper
LAW	17 U.S.C. §§ 102, 105-08, 202, 1201	Current U.S. copyright statutory framework; primary legal authority.
GUIDE	U.S. Copyright Office Fair Use Index	Official educational summaries of selected decisions; not a substitute for opinions.
GUIDE	U.S. Copyright Office AI Report Parts 1-3	Part 1 Digital Replicas (2024); Part 2 Copyrightability (2025); Part 3 Generative AI Training pre-publication (May 2025; still labeled pre-publication on Office site Sept. 2026).
CASE	Andy Warhol Foundation v. Goldsmith, 598 U.S. 508 (2023)	U.S. Supreme Court; specific-use/purpose analysis under factor one.
CASE	Authors Guild v. Google, 804 F.3d 202 (2d Cir. 2015)	Second Circuit; book search/digitization/snippets.
CASE	Perfect 10 v. Amazon, 508 F.3d 1146 (9th Cir. 2007)	Ninth Circuit; image-search thumbnails.
CASE	Thomson Reuters v. ROSS, 765 F. Supp. 3d 382 (D. Del. 2025)	District court; non-generative competing legal-research tool; fair-use defense rejected for headnotes at issue.
CASE	Bartz v. Anthropic, 787 F. Supp. 3d 1007 (N.D. Cal. 2025)	District court; separated training/digitization from pirate-library acquisition; fact-specific and later settlement proceedings.
LAW	Regulation (EU) 2024/1689 art. 53	EU AI Act GPAI documentation/copyright policy/training-content-summary obligations.
GUIDE	European Commission GPAI Code of Practice	Voluntary compliance route; Transparency/Copyright chapters relevant to Article 53.
PEER	Peukert & Castets-Renard (2025), IIC	Peer-reviewed commentary by Copyright Chapter working-group chairs.
LAW	GDPR art. 5	Purpose limitation, data minimization, accuracy, storage limitation, integrity/confidentiality where applicable.
STD	W3C PROV-DM	Entity/activity/agent provenance model and derivation/responsibility relations.
STD	C2PA Content Credentials 2.4 (Apr. 2026)	Tamper-evident provenance assertions, claims, bindings, ingredients/actions; not a truth oracle.
STD	RFC 9309 Robots Exclusion Protocol	Crawler signaling standard; explicitly not access authorization.
GUIDE	NIST AI RMF: Generative AI Profile, NIST AI 600-1	Voluntary lifecycle risk-management profile for generative AI.
GUIDE	NIST Privacy Framework	Privacy risk management and SDLC/data-lifecycle integration.
GUIDE	Creative Commons attribution practices	TASL and modification attribution guidance for CC-licensed material.
GUIDE	OWASP File Upload / Prompt Injection Cheat Sheets	Operational security guidance; defense in depth and untrusted-content handling.
PEER	Wilkinson et al. (2016), Scientific Data	FAIR principles for machine-actionable findability/access/interoperability/reuse.
PEER	Gebru et al. (2021), Communications of the ACM	Datasheets for Datasets documentation framework.
PEER	Mitchell et al. (2019), FAT*	Model Cards for Model Reporting.
PEER	Smith et al. (2016), PeerJ Computer Science	Software citation principles.
PEER	Sandve et al. (2013), PLOS Computational Biology	Reproducible computational research practices.
PEER	Chen et al. (2024), EMNLP	CopyBench: literal and non-literal reproduction evaluation.
PROJECT	The Raven's Right / Nuliometry / Visual Holography / RavenReader	Internal design precedent and evidence-governance lineage; not external empirical authority.

WORKS CITED

17 U.S.C. § 102. Copyright Law of the United States, U.S. Copyright Office, Circular 92, Dec. 2025, www.copyright.gov/title17/92chap1.html. Accessed 29 Sept. 2026.

17 U.S.C. § 105. Copyright Law of the United States, U.S. Copyright Office, Circular 92, Dec. 2025, www.copyright.gov/title17/92chap1.html. Accessed 29 Sept. 2026.

17 U.S.C. §§ 106-108. Copyright Law of the United States, U.S. Copyright Office, Circular 92, Dec. 2025, www.copyright.gov/title17/92chap1.html. Accessed 29 Sept. 2026.

17 U.S.C. § 202. Copyright Law of the United States, U.S. Copyright Office, Circular 92, Dec. 2025, copyright.gov/title17/92chap2.html. Accessed 29 Sept. 2026.

17 U.S.C. § 1201. Copyright Law of the United States, U.S. Copyright Office, Circular 92, Dec. 2025, www.copyright.gov/title17/92chap12.html. Accessed 29 Sept. 2026.

Andy Warhol Foundation for the Visual Arts, Inc. v. Goldsmith. 598 U.S. 508. Supreme Court of the United States. 2023.

Authors Guild, Inc. v. Google, Inc. 804 F.3d 202. U.S. Court of Appeals for the Second Circuit. 2015.

Bartz v. Anthropic PBC. 787 F. Supp. 3d 1007. U.S. District Court for the Northern District of California. 23 June 2025.

C2PA. Content Credentials: C2PA Technical Specification. Version 2.4, Apr. 2026, spec.c2pa.org/specifications/specifications/2.4/specs/C2PA_Specification.html. Accessed 29 Sept. 2026.

Chen, Tong, et al. "CopyBench: Measuring Literal and Non-Literal Reproduction of Copyright-Protected Text in Language Model Generation." Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, 2024, pp. 15134-15158. doi:10.18653/v1/2024.emnlp-main.844.

Conrad, Noel Michael, and OpenAI. Nuliometry: Shapes of the Unknowable. Raven Valley / Opus II / Visual Holography, version 1.0, 9 Sept. 2026. Internal project white paper.

Conrad, Noel Michael, and OpenAI. The Raven's Right: The Condition of Truth. Raven Valley / Opus II, version 1.0, 16 Aug. 2026. Internal project white paper.

Conrad, Noel Michael, and OpenAI. Visual Holography: Prompt-Image-Drawing Lineage Study. Raven Valley / Opus II, version 1.0, 11 Aug. 2026. Internal project white paper.

Creative Commons. "Recommended Practices for Attribution." Creative Commons Wiki, wiki.creativecommons.org/wiki/Recommended_practices_for_attribution. Accessed 29 Sept. 2026.

European Commission. "The General-Purpose AI Code of Practice." Shaping Europe's Digital Future, digital-strategy.ec.europa.eu/en/policies/contents-code-gpai. Accessed 29 Sept. 2026.

European Parliament and Council. Regulation (EU) 2016/679 (General Data Protection Regulation). 27 Apr. 2016, art. 5, eur-lex.europa.eu/eli/reg/2016/679. Accessed 29 Sept. 2026.

European Parliament and Council. Regulation (EU) 2024/1689 (Artificial Intelligence Act). 13 June 2024, art. 53, EUR-Lex, eur-lex.europa.eu/legal-content/EN/ALL/?uri=CELEX:32024R1689. Accessed 29 Sept. 2026.

Gebru, Timnit, et al. "Datasheets for Datasets." Communications of the ACM, vol. 64, no. 12, 2021, pp. 86-92.

Koster, Martijn, et al. "Robots Exclusion Protocol." RFC 9309, IETF, Sept. 2022, www.rfc-editor.org/rfc/rfc9309.html. Accessed 29 Sept. 2026.

Mitchell, Margaret, et al. "Model Cards for Model Reporting." Proceedings of the Conference on Fairness, Accountability, and Transparency, ACM, 2019, pp. 220-229. doi:10.1145/3287560.3287596.

National Institute of Standards and Technology. Artificial Intelligence Risk Management Framework: Generative Artificial Intelligence Profile. NIST AI 600-1, 26 July 2024, updated 8 Apr. 2026, doi:10.6028/NIST.AI.600-1.

National Institute of Standards and Technology. "Privacy Framework." NIST, www.nist.gov/privacy-framework. Accessed 29 Sept. 2026.

OWASP Foundation. "File Upload Cheat Sheet." OWASP Cheat Sheet Series, cheatsheetseries.owasp.org/cheatsheets/File_Upload_Cheat_Sheet.html. Accessed 29 Sept. 2026.

OWASP Foundation. "LLM Prompt Injection Prevention Cheat Sheet." OWASP Cheat Sheet Series, cheatsheetseries.owasp.org/cheatsheets/LLM_Prompt_Injection_Prevention_Cheat_Sheet.html. Accessed 29 Sept. 2026.

Peukert, Alexander, and Celine Castets-Renard. "The Copyright Chapter of the EU's Code of Practice for General-Purpose AI Models: A Commentary." IIC - International Review of Intellectual Property and Competition Law, vol. 56, 2025, pp. 1950-1993. doi:10.1007/s40319-025-01664-8.

Perfect 10, Inc. v. Amazon.com, Inc. 508 F.3d 1146. U.S. Court of Appeals for the Ninth Circuit. 2007.

Sandve, Geir Kjetil, et al. "Ten Simple Rules for Reproducible Computational Research." PLOS Computational Biology, vol. 9, no. 10, 2013, e1003285. doi:10.1371/journal.pcbi.1003285.

Smith, Arfon M., et al. "Software Citation Principles." PeerJ Computer Science, vol. 2, 2016, e86. doi:10.7717/peerj-cs.86.

Thomson Reuters Enterprise Centre GmbH v. ROSS Intelligence Inc. 765 F. Supp. 3d 382. U.S. District Court for the District of Delaware. 11 Feb. 2025.

U.S. Copyright Office. Copyright and Artificial Intelligence, Part 1: Digital Replicas. July 2024, www.copyright.gov/ai/Copyright-and-Artificial-Intelligence-Part-1-Digital-Replicas-Report.pdf. Accessed 29 Sept. 2026.

U.S. Copyright Office. Copyright and Artificial Intelligence, Part 2: Copyrightability. Jan. 2025, www.copyright.gov/ai/Copyright-and-Artificial-Intelligence-Part-2-Copyrightability-Report.pdf. Accessed 29 Sept. 2026.

U.S. Copyright Office. Copyright and Artificial Intelligence, Part 3: Generative AI Training. Pre-publication version, May 2025, www.copyright.gov/ai/Copyright-and-Artificial-Intelligence-Part-3-Generative-AI-Training-Report-Pre-Publication-Version.pdf. Accessed 29 Sept. 2026.

U.S. Copyright Office. "Fair Use Index." copyright.gov/fair-use/. Accessed 29 Sept. 2026.

Wilkinson, Mark D., et al. "The FAIR Guiding Principles for Scientific Data Management and Stewardship." Scientific Data, vol. 3, 2016, article 160018. doi:10.1038/sdata.2016.18.

World Wide Web Consortium. PROV-DM: The PROV Data Model. W3C Recommendation, 30 Apr. 2013, www.w3.org/TR/prov-dm/. Accessed 29 Sept. 2026.

CITATION VERIFICATION NOTE External sources in this edition were checked against official statutory/regulatory text, court or authoritative case records, formal standards, government publications, or peer-reviewed publication records available on 29 Sept. 2026. Time-sensitive legal status remains subject to later amendment, appeal, settlement, regulation, and jurisdiction-specific interpretation.