

AI and IP: Authorship, Licensing, and Fair Use

A.T. Svanovich^{1*}, E.V. Petrova²

¹ Department of Computer Science, National Research University, Almaty, Kazakhstan

² Institute for Digital Privacy & Security, Astana, Kazakhstan

Corresponding author: A.T. Svanovich, email: a.svanovich@university.kz

Abstract

Artificial intelligence unsettles long-standing assumptions in intellectual property (IP). This article synthesizes current doctrine and practice on (1) human authorship and the copyrightability of AI-assisted works; (2) lawful acquisition and use of training data, including EU text-and-data mining (TDM) exceptions and opt-outs; (3) licensing patterns across data, model weights, and outputs; (4) technical provenance and attribution (C2PA/Content Credentials); and (5) governance to manage legal, operational, and reputational risk. We draw on recent U.S. guidance and case law affirming the human authorship requirement, EU legislative frameworks for TDM, and emerging standards for content authenticity. We conclude with a practical playbook—policies, contracts, pipelines, and audits—that lets organizations innovate responsibly while protecting creators’ rights and minimizing litigation exposure.

1. Introduction

Generative systems blur the line between human creativity and automated production. For organizations, the central questions are not merely “Is training legal?” or “Who owns the output?”, but how to design accountable pipelines that respect rights across the entire chain—from input data to deployed models and published works—while preserving creative freedom and commercial value.

2. Human authorship and copyrightability

U.S. doctrine presently requires human authorship to claim copyright in a work. The U.S. Copyright Office’s policy guidance instructs applicants to disclose AI-generated portions; protection may cover only the identifiable human contributions (selection, arrangement, editing), not purely machine-generated material. Applicants must explain the nature and extent of human input.

In March 2025, the D.C. Circuit affirmed that a work created solely by AI lacks copyright because authorship under the statute is human; the decision in *Thaler v. Perlmutter* solidifies agency practice at the appellate level. Creators seeking protection should therefore document prompt design, editorial choices, and post-processing to show creative control.

Implication. Register only what a person actually authored; describe machine assistance candidly; and maintain process records (briefs, drafts, redlines) that evidence human creativity.

3. Training datasets: text-and-data mining, opt-outs, and fair use

European Union. Directive (EU) 2019/790 (DSM) establishes two TDM exceptions: a mandatory research exception (Art. 3) and a broader TDM exception (Art. 4) that rights holders can **opt out** of via machine-readable reservations. Where an opt-out exists, commercial TDM typically requires a license. Early national cases are beginning to interpret these provisions and the mechanics of opt-out, increasing the compliance premium on provenance and logs.

United States. The Copyright Office’s ongoing study underscores that “fair use” is fact-specific; there is no blanket immunity for training. Organizations should avoid treating unlicensed scraping as categorically lawful and instead assess purpose, amount, market effects, and the nature of sources—preferably paired with negotiated licenses for high-risk corpora.

Implementation pointers. Build a **sources & rights register** (owned / licensed / open / restricted); respect EU TDM opt-outs; segregate and quarantine uncertain sources; and document why each corpus is permissible to ingest.

4. Licensing across the AI stack

Effective rights management distinguishes three layers:

1. **Data & content.** Secure content licenses fit for training (scope, territory, term, revocation, attribution, TDM permissions). For open content, verify license compatibility, including attribution and share-alike duties.
2. **Models & weights.** Model licenses (OSS-style or EULA) increasingly constrain use (e.g., sensitive domains), derivatives, and redistribution. Ensure the license covers **training, fine-tuning, and hosting** and aligns with your distribution model.
3. **Outputs.** Copyright in outputs hinges on human authorship. Even where protected, outputs must not reproduce third-party expression beyond permitted uses. Contracts with clients should allocate risk (representations, warranties, and indemnities) and specify acceptable uses and attribution.

5. Attribution, provenance, and content authenticity

Legal compliance is easier when provenance is verifiable. The C2PA standard (adopted through **Content Credentials**) cryptographically binds “who did what, with which tool, and when” to media via signed manifests that persist across edits. Emerging specs detail embedding for formats like JPEG XL and define trust and tamper-evidence models. Integrating Content Credentials into creative and publishing pipelines supports attribution, combats misinformation, and provides audit trails for licensing terms. c2pa.org+2spec.c2pa.org+2

Internationally, WIPO’s multiyear “Conversation on IP and AI” continues to surface policy options (authorship, inventorship, data rights), signaling that provenance and transparency are likely pillars of any future consensus.

6. Patents, inventorship, and AI assistance

While this article focuses on copyright, the patent system is converging on similar principles: AI can assist, but **inventorship remains human**. The USPTO’s 2024 guidance urges disclosure of meaningful AI involvement and cautions practitioners to verify AI-assisted filings to avoid errors and ethical breaches. Treat AI as a tool; document human conception and reduction to practice.

7. Risk landscape and controls

Legal risks. Claims over training on protected works (especially where opt-outs exist), misattribution or removal of rights management info, and output-level infringement (e.g., close style-matching or near-verbatim replication).

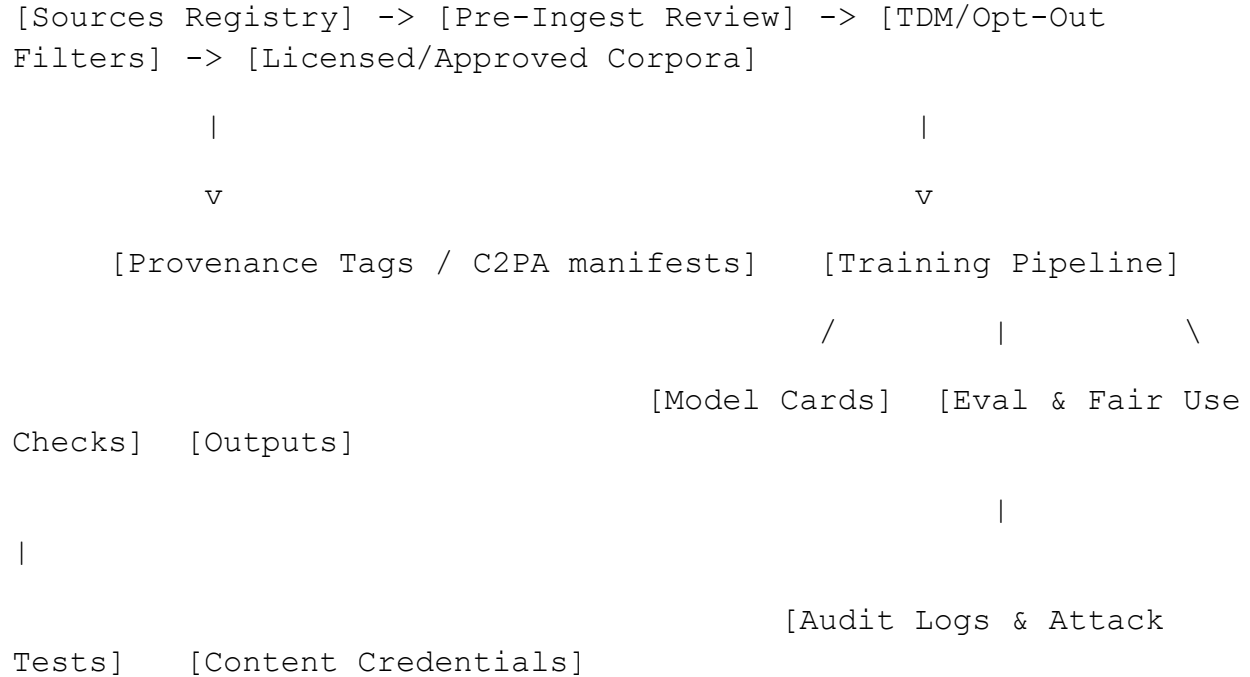
Operational risks. Weak source traceability, “dataset drift” through silent substitutions, and inability to demonstrate due diligence to partners or regulators.

Reputational risks. Community backlash from creators; platform take-downs; loss of trust with advertisers or users.

Control framework (minimum viable).

- **Policy:** a written data-and-model policy covering intake, TDM opt-out respect, and record-keeping.
- **Contracts:** training-safe licenses; indemnities with caps; notice-and-takedown flows; explicit attribution terms.
- **Provenance tech:** C2PA manifests by default; model/dataset **cards** that list sources and known limits.
- **Process:** pre-ingestion reviews; periodic audits; canary tests for memorization; output scanning for style/near-duplicate risks.

8. A practical playbook (policy-as-code + provenance-by-design)



Checklist

1. Map sources into owned / licensed / open / restricted; block “unknown.”
2. Implement machine-readable TDM opt-out filters for EU compliance.
3. Record human creative choices to sustain copyright claims in outputs.
4. Attach Content Credentials on publish; retain signing keys securely.
5. Align model/data licenses with distribution plans (hosted API vs. on-prem).
6. Establish incident response for rights complaints (triage, rollback, retrain if needed).

9. Outlook

Near-term pressure points include: (i) more national case law on TDM opt-outs and scraping; (ii) standardized disclosure of AI assistance in copyright and patent filings; and (iii) at-scale adoption of cryptographically verifiable provenance. Organizations that invest early in **provenance-by-design** and **policy-as-code** will negotiate faster, reduce litigation risk, and build durable trust with creators and customers.

Conclusion

AI does not repeal IP; it **re-partitions accountability**. By anchoring authorship in documented human creativity, acquiring training rights through licenses or compliant TDM, and instrumenting provenance, companies can ship AI-enabled products that are both legally resilient and creator-respectful.

Keywords: Artificial Intelligence, Copyright, Training Datasets, Licensing, Fair Use, Provenance

Selected references (APA 7th)

U.S. Copyright Office. (2023). *Works Containing Material Generated by AI* (Policy Statement). *Thaler v. Perlmutter*, No. 23-5233 (D.C. Cir. 2025).
European Union. (2019). *Directive (EU) 2019/790 (DSM)*—Arts. 3–4 (TDM).
C2PA. (2025). *C2PA Technical Specification v2.2*; Content Credentials.
WIPO. *Conversation on IP and AI* (sessions, issues papers).
USPTO / Reuters coverage. (2024). Guidance on AI in patent practice.

This analysis is informational and not legal advice; jurisdiction-specific counsel is recommended.