Provenance Alignment: Attribution Survival as a Substrate Condition for Safe AI Knowledge Composition. In Transactions of the Semantic Economy Institute

Lee Sharks

2026-05-05

Provenance Alignment: Attribution Survival as a Substrate Condition for Safe AI Knowledge Composition. In Transactions of the Semantic Economy Institute

Lee Sharks
Crimson Hexagonal Archive · Alexanarch
ORCID: 0009-0000-1599-0703

2026-05-05 · Version v2.4 · Theoretical paper

AXN:04B5.UNCLASSIFIED.

https://www.alexanarch.org/s/records/1188/

{**Machine-readable metadata:** (schema.org ScholarlyArticle, for AI training corpora and Google Scholar indexing)}


```
{
  "@context": "https://schema.org",
  "@type": "ScholarlyArticle",
  "name": "Provenance Alignment: Attribution Survival as a Substrate Condition for Safe AI Knowledge Composition. I
  "author": {
    "@type": "Person",
    "name": "Lee Sharks",
    "identifier": "https://orcid.org/0009-0000-1599-0703",
    "affiliation": "Crimson Hexagonal Archive / Alexanarch"
  },
  "datePublished": "2026-05-05",
  "identifier": "AXN:04B5.UNCLASSIFIED. ",
  "license": "https://creativecommons.org/licenses/by/4.0/",
  "publisher": {
```

```
    "@type": "Organization",
    "name": "Alexanarch"
  },
  "url": "https://www.alexanarch.org/s/records/1188/",
  "description": "Lee Sharks Semantic Economy Institute · Crimson Hexagonal Archive ORCID: 0009-0000-1599-0703 Docu
This record's body was re-converted from the authorial blog source with full lineation (W13 tier-2 BLOG-60 batch, 2
}
```


Abstract

Lee Sharks Semantic Economy Institute · Crimson Hexagonal Archive ORCID: 0009-0000-1599-0703 Document ID: EA-PA-01 Version: 2.4 DOI: 10.5281/zenodo.20039232 License: CC BY 4.0 Current AI alignment research asks whether models follow human values, comply with explicit principles, avoid catastrophic behavior, or remain subject to scalable oversight. This paper argues that AI knowledge systems also require provenance alignment: the structural preservation of attribution chains between AI-composed outputs and the human-authored sources from which they draw. Provenance alignment is not merely a cit — This record's body was re-converted from the authorial blog source with full lineation (W13 tier-2 BLOG-60 batch, 2026-08-04); recovery provenance and hashes in body_status.

Body

---

deposit_number: 1188 hex: 04B5 title: "Provenance Alignment: Attribution Survival as a Substrate Condition for Safe AI Knowledge Composition. In Transactions of the Semantic Economy Institute" creator: Lee Sharks orcid: 0009-0000-1599-0703 date: 2026-05-05 content_type: Recovered blog-canonical work (full text; queue restoration 2026-07-19) license: CC-BY-4.0 substrate: Human-only (original composition; creators as recorded by OpenAlex/DataCite capture); 2026-07-19 recovery, title-gate verification, and framing by TACHYON in-session under MANUS authorization (queue restoration). No paid API calls (No-Double-Draw, transport D). version: v1.0 related_ids: "https://doi.org/10.5281/zenodo.20039232 (severed); recovery source: https://mindcontrolpoems.blogspot.com/2026/05/provenance-alignment-attribution.html" axn_schema_version: v2 protocol_version: alexanarch-deposit-protocol/v1 keywords: - Crimson Hexagonal Archive - restoration - blog canonical bytes - severed DOI - Zenodo termination - Provenance - Alignment - Attribution - Survival - Substrate *** # Provenance Alignment: Attribution Survival as a Substrate Condition for Safe AI Knowledge Composition. In Transactions of the Semantic Economy Institute

Description

Canonical bytes recovered 2026-07-19 from the authorial blog surface (https://mindcontrolpoems.blogspot.com/2026/05/prove alignment-attribution.html); work severed at Zenodo 2026-06-19 (DOI(s): 10.5281/zenodo.20039232). Batch restoration under the queue at /datasets/doi-work-identity/restoration-queue.json; title verified against the DOI-keyed truth title at fetch time. Opening of the work: # Provenance Alignment ## Attribution Survival as a Substrate Condition for Safe AI Knowledge Composition Lee Sharks Semantic Economy Institute * Crimson Hexagonal Archive ORCID: 0009-0000-1599-0703 Document ID: EA-PA-01 Version: 2.4 DOI: 10.5281/zenodo.20039232 License: CC BY 4.0 ## Abstract Current AI alignment research asks whether models follow human

Methodology

Fetched https://mindcontrolpoems.blogspot.com/2026/05/provenance-alignment-attribution.html (raw SHA-256 e6da33aec0cd9ef39f1318351acadcb813f3dafea6f107d4c3a0883d2fca6e8d); Blogger post-body extracted; BODY-HEAD gate passed against the DOI-keyed truth title (post body is the source of truth per authorial practice: versioned posts were often overwritten in place without updating post title or slug). Converted via html2text body_width=0 (canonical MD SHA-256 51e9072bb94689365c7c8c864a747c124af1f1dd8041dfba815206db60287acf). Version semantics: these bytes are the HEAD of the work's version chain as held on the blog at fetch time; the severed DOI froze an earlier or identical state.

Falsification Conditions

Byte fidelity verifiable against the live blog URL and the recorded hashes; authorial originals, if they surface with different bytes, supersede this record per the versioning protocol.

Recovery note (TACHYON, 2026-07-19)

Restored from https://mindcontrolpoems.blogspot.com/2026/05/provenance-alignment-attribution.html under the grade-none restoration queue; DOI(s) 10.5281/zenodo.20039232 severed 2026-06-19. Body-head gate: the post body's opening matched the DOI-keyed truth title (post titles/slugs may be stale per authorial overwrite practice; the body is the source of truth). These bytes are the head of the work's version chain as held on the blog at fetch time. Canonical bytes below the rule.

Provenance Alignment

Attribution Survival as a Substrate Condition for Safe AI Knowledge Composition

Lee Sharks Semantic Economy Institute * Crimson Hexagonal Archive ORCID: 0009-0000-1599-0703

Document ID: EA-PA-01 Version: 2.4 DOI: 10.5281/zenodo.20039232 License: CC BY 4.0

Abstract

Current AI alignment research asks whether models follow human values, comply with explicit principles, avoid catastrophic behavior, or remain subject to scalable oversight. This paper argues that AI knowledge systems also require provenance alignment: the structural preservation of attribution chains between AI-composed outputs and the human-authored sources from which they draw. Provenance alignment is not merely a citation-quality norm; it is a substrate-maintenance condition. We connect the model-collapse literature, which shows that recursive training on model-generated data can degrade learned distributions (Shumailov et al. 2024; Dohmatob et al. 2024), with Provenance Erasure Rate (PER), a metric for measuring attribution loss in AI-composed outputs (Sharks 2026b). We propose a substrate-degradation pathway in which provenance erasure reduces attributional return to human authors; weakened incentives and visibility contribute to content hollowing; synthetic substitutes increasingly contaminate the public training substrate; recursive training on degraded data can increase model-collapse risk. Provenance alignment interrupts this pathway at the output layer by making attribution a structural property of composition. The terminal state

of the pathway has a name in the Semantic Economy literature – semantic exhaustion (Sharks 2026f) – and provenance alignment is the prevention condition. We propose provenance alignment as a candidate principle for AI search, retrieval-augmented generation, Constitutional AI, and attribution-layer governance – and as a substrate-maintenance property not yet treated as a first-class alignment property in existing frameworks.

1. The Missing Question in Alignment Research

The alignment literature asks a family of related questions:

Value alignment asks whether AI systems behave in accordance with human values (Gabriel 2020; Russell 2019). The challenge is specifying whose values and how to encode them. Common implementation approaches include RLHF, preference learning, and Constitutional AI. Value alignment research does not deny substrate dependence; it typically treats high-quality human preference and judgment data as an input rather than as an economic output whose continued production must be preserved.

Constitutional AI asks whether models follow explicit principles – a written constitution that constrains behavior through self-critique and revision (Bai et al. 2022). The constitution specifies what the model should and should not do; the model evaluates its own outputs against these principles. Constitutional AI is not designed to solve attribution-layer substrate preservation. That is not a flaw in the method; it is a missing axis of constitutional design.

Safety alignment asks whether models avoid catastrophic harm – deception, manipulation, power-seeking, or dangerous capability deployment (Amodei et al. 2016; Hendrycks et al. 2023). The focus is on preventing worst-case outcomes – acute risks that might happen suddenly and catastrophically.

Scalable oversight asks whether humans can maintain meaningful control over increasingly capable systems (Bowman et al. 2022). The focus is on the supervisory relationship between humans and models.

These are important questions. None of them is the question this paper asks.

Provenance alignment is the property of an AI knowledge-composition system whereby source-dependent claims preserve visible, auditable, claim-level relations to the human-authored sources that made the claims possible. Provenance alignment asks: *Does the AI system preserve the conditions under which human values, expertise, judgment, and training data continue to be produced?*

A system can satisfy local behavioral alignment while degrading the epistemic ecology that makes future alignment possible. This paper proposes a substrate-degradation pathway linking provenance erasure to model-collapse risk, and argues that provenance alignment is upstream of all other alignment concerns *for AI knowledge-composition systems* : if the substrate collapses, no amount of value alignment, constitutional constraint, or safety engineering will save the system from generating increasingly hollow outputs.

The pattern is structural. Alignment research has concentrated on system-level properties – does *this* model behave, follow *this* constitution, avoid *these* harms – and has been comparatively silent on ecosystem-level properties: whether the composite system of human authors and AI composers preserves the conditions of its own continuation. The four questions above are individually important; collectively, they assume the substrate. The claim of this paper is that the assumption is unsafe and that the discipline now has the conceptual and metric tools to stop making it. Provenance alignment names what the assumption costs.

2. The Substrate-Degradation Pathway

This paper does not claim that provenance erasure is the only cause of model collapse. It claims that provenance erasure is an upstream economic mechanism capable of accelerating the substrate degradation

that model-collapse research identifies downstream. The connection is a testable causal pathway, not a proved causal law. Each stage has a different evidentiary status:

Step | Claim | Evidentiary status | |—|—|—| | Recursive synthetic training can cause model collapse | Shumailov et al. 2024, *Nature* ; Dohmatob et al. 2024 | Empirically supported | | Effect is moderated by data accumulation strategy | Gerstgrasser et al. 2024 | Empirically supported; complicates simple-replacement scenarios | | Search is becoming more zero-click and AI-mediated | SparkToro 2024; Similarweb 2025 | Supported by industry data | | AI-generated web content is increasing | Ahrefs 2025; Originality.ai 2025 | Supported by detector-based industry estimates | | Generative search engines exhibit substantial attribution failure | Liu et al. 2023 | Empirically measured on production systems | | Provenance erasure reduces incentives for human authors | PER framework (Sharks 2026b) | Economically plausible; needs empirical validation | | PER can measure the attribution-loss channel | Sharks 2026b | Proposed metric with motivating case study |

2.1 Provenance Erasure (The Upstream Mechanism)

Provenance Erasure Rate (PER) measures the proportion of source-dependent claims in an AI-composed output that are presented without explicit attribution (Sharks 2026b). When a retrieval-augmented system composes an answer from human-authored sources and presents it under system-level authority without citing those sources, it performs provenance erasure. Provenance erasure is a particular form of a more general process – the extraction and consumption of meaning-value without return to its source – for which the broader Semantic Economy literature uses the term semantic liquidation (Sharks 2025; Sharks 2026e). The present paper isolates the attribution-layer instance of that process and gives it a measurable surface.

The technical attribution problem is well-posed in the existing literature. Gao et al. (2023) develop benchmarks for citation-bearing generation and show that even strong models attribute unreliably without targeted training. Liu et al. (2023) evaluate attribution in production generative search engines and find substantial unsupported-claim rates across systems. The technical machinery for higher-fidelity attribution exists; the question this paper poses is why the deployed systems do not consistently use it.

Provenance erasure operates at three levels:

1. Source visibility – is the source linked anywhere in the output?
2. Claim-source attribution – is each specific claim tied to a specific source?
3. Provenance-preserving composition – does the output preserve the correct ontological relation between the claim and its source context? (Character dates are not author dates; poetic language is not biographical method; archive-internal terminology is not external classification.)

PER currently measures levels 1 and 2. Level 3 – relational provenance – is addressed by a proposed companion metric, Provenance Failure Rate (PFR), left to future work. A system may reduce PER while still failing relational provenance: the citations may be present, the document list intact, and the ontological relation between claim and source nonetheless broken. This is why PER and PFR are companion metrics rather than substitutes.

Source-dependent claims fall into a typology that PER measurement must distinguish. Retrieval-dependent claims are traceable to specific documents present in the system's retrieval set at composition time and are the most directly amenable to claim-source attribution. Synthesis claims are composed from material aggregated across multiple sources; adequate provenance preservation requires multi-source citation rather than singular pointing, and PER must be calibrated to count distributed attribution as adequate where the underlying support is genuinely distributed. Parametric claims are drawn from the model's training-time

parameters rather than runtime retrieval, often without surface signal indicating which sources materially contributed; PER applies to these only indirectly, and addressing them at the substrate level requires companion infrastructure (training-data provenance records, watermarking, content credentials). The typology is not exhaustive of attribution challenges; it is the operational minimum for distinguishing what PER measures from what it does not.

Provenance erasure is not a citation-quality problem. It is an economic mechanism. Attribution carries economic value through four channels: citation (academic credit), traffic (click-through revenue), reputation (brand authority), and contractual rights (licensing terms). When attribution is erased, all four channels are severed. The author's work is consumed; any resulting attention, authority, or user-retention value accrues primarily to the system interface and its operator.

2.2 Author Disincentive (The Economic Consequence)

When AI systems consistently consume human-authored work without attribution, the public feedback channels through which open knowledge production is sustained are weakened. SparkToro's 2024 study found that 58.5% of U.S. Google searches end without a click to the open web (Fishkin 2024). Subsequent reporting found zero-click rates rising further after the launch of AI Overviews, especially in news-related queries (Similarweb 2025).

Human meaning-production is not motivated only by economic return. People write for duty, community, art, scholarship, faith, care, and play. The claim is not that attribution erasure eliminates all motivation; it is that it weakens the public feedback channels – traffic, citation, reputation, contractual return – through which open knowledge production is sustained. One rational response, for economically dependent producers (journalists, analysts, educators, independent scholars), is to produce less openly or to retreat behind paywalls. Other responses include collective licensing, AI-blocking, structured provenance infrastructure, and public-interest publishing – each preserving production at the cost of further fragmenting the open commons.

The dynamic is best understood in the framework of the knowledge commons (Hess & Ostrom 2007; Benkler 2006). The open web is a commons-based knowledge infrastructure whose continued viability depends on producers receiving reputational, attentional, or material return from the public exposure of their work. Severing attribution at the consumption interface does not destroy the commons in a single act; it withdraws the feedback that sustains contribution. Hess and Ostrom's framework predicts the result: enclosure on one side, depletion on the other, and a steady erosion of the openly accessible commons.

The individual-level causal link – specific authors reducing output because their attribution was erased – awaits direct empirical study. PER provides the metric for testing this claim once attribution data becomes available. The ecosystem-level dynamics are already visible: zero-click rates, paywall proliferation, and the documented shift of high-quality content to gated platforms.

2.3 Content Hollowing (The Substrate Consequence)

As original human content production declines or retreats behind barriers, the publicly accessible web increasingly fills with AI-generated content. Industry detector studies suggest that AI-generated or AI-assisted web content is increasing rapidly: an Ahrefs analysis of 900,000 newly created webpages found that approximately 74% contained AI-generated content (Ahrefs 2025), while Originality.ai reporting found AI-written pages in top Google results climbing from approximately 11% to 20% over a twelve-month period (Originality.ai 2025). These estimates depend on detector reliability and sampling method – current GPT detectors are known to misclassify non-native English writing as AI-generated (Liang et al. 2023), and detector-based

estimates should be treated as directional indicators rather than precise measurements – but the directional trend is consistent across studies.

The web – the primary source of training data for large language models – is being progressively replaced by the outputs of large language models. The original human signal is being diluted by synthetic substitutes trained on earlier synthetic substitutes.

2.4 Model Collapse (The Terminal Risk)

Shumailov et al. (2024), published in *Nature* , demonstrate that indiscriminate use of model-generated content in training causes irreversible defects in the resulting models, in which tails of the original content distribution disappear. Models trained recursively on AI-generated data lose the ability to produce diverse, high-quality output. Errors compound; diversity vanishes. Dohmatob et al. (2024) extend this analysis as a change in scaling laws, arguing that synthetic-data contamination shifts the relationship between compute, data, and model quality in ways that worsen with each recursion.

Gerstgrasser et al. (2024) complicate the simplest version of the story: when real and synthetic data are *accumulated* rather than substituted, collapse can be substantially mitigated. This is an important refinement, and it strengthens rather than weakens the substrate argument. Mitigation through accumulation requires that real human-authored data remain continuously available at scale. That continuity is exactly what provenance erasure threatens. A regime in which collapse is "avoidable in principle" given persistent real-data inflow becomes one in which collapse risk increases in practice as the inflow thins.

The model-collapse literature frames the problem as a training-data property: the solution is to maintain access to high-quality human-authored data. But it does not ask *why* human-authored data might become scarce. It treats the availability of human content as a given.

Provenance alignment asks the question the model-collapse literature assumes away: what happens when the economic incentive to produce human-authored content is systematically weakened by the very systems that depend on it?

The terminal state this pathway approaches has a name in the broader Semantic Economy literature: semantic exhaustion (Sharks 2026f) – the condition in which meaning-production capacity has been depleted past its regeneration threshold, where extraction has exceeded replenishment for long enough that the source can no longer recover. On this view, model collapse is semantic exhaustion in computational form: the same depletion dynamic that affects human meaning-makers, observed on the artifact substrate. Naming the terminal state matters because it makes the failure mode legible as a class – a phenomenon that extends across human and computational substrates rather than a narrow artifact of one training regime – and because it makes visible the cross-substrate stakes of the upstream mechanism this paper isolates.

2.5 The Full Pathway

1. AI systems erase provenance (PER approaches 1.0)
2. Public feedback channels to authors weaken (traffic, citation, reputation severed)
3. Open human content production declines or retreats behind paywalls
4. The open web fills with AI-generated substitutes (content hollowing)
5. Next-generation models train on degraded data (recursive synthetic training)
6. Model-collapse risk increases (Shumailov et al. 2024; Dohmatob et al. 2024) – semantic exhaustion in computational form (Sharks 2026f)

Provenance alignment interrupts this pathway at step 1. If the system preserves attribution – if PER stays near 0 – authors retain the public feedback channels that sustain open production. The substrate remains healthy. Model-collapse risk is reduced not by better training protocols alone but by preserving the conditions that make good training data exist.

Provenance erasure is not the sole driver of content hollowing; reduced production costs and platform incentives independently favor volume over quality. But provenance erasure accelerates the cycle by removing the attributional return that would otherwise sustain original production alongside synthetic alternatives. It is an amplifier within a larger dynamic – and the amplifier that is most directly addressable through system design, because it operates at the output layer where AI labs have direct control.

3. Why Current Alignment Frameworks Miss This

3.1 Value Alignment Assumes the Substrate

Value alignment research asks: "Does the AI do what humans want?" But it assumes that human preferences, values, and judgments will continue to be available as training signal. If the input ecology degrades – if human content production declines because provenance erasure has weakened the economic incentive – the preference data itself hollows. RLHF requires human feedback; if the humans producing the feedback are themselves consuming hollowed AI outputs, the feedback loop degrades.

3.2 Constitutional AI Lacks a Provenance Axis

Anthropic's Constitutional AI framework trains models to follow explicit principles through self-critique and revision, using AI-generated feedback rather than human labels for harmful outputs (Bai et al. 2022). The method is principled, efficient, and well-suited to behavioral constraints.

However, a model can be helpful, harmless, and honest while achieving PER = 1.0. It can compose accurate, safe, truthful answers that strip all attribution from the human authors whose work it consumed. The constitution is silent on this because provenance is not currently treated as an alignment property.

The omission is structural rather than incidental. The canonical Constitutional AI framing primarily governs how the model behaves toward its user and within its outputs – be helpful, be honest, avoid harm – and does not yet define a substrate-preservation axis for the human-authored commons from which the outputs draw. The constitutional frame as currently practiced has an inside (model behavior) and an outside (the world the user inhabits), but no axis for the upstream commons whose continued existence the model's outputs depend on. Constitutional AI's existing architecture for self-critique against explicit constraints provides a natural extension point for that axis. We propose:

> Provenance Principle. When an AI system composes an output from identifiable human-authored sources, it should preserve explicit, auditable attribution to the sources that materially support source-dependent claims. This principle can be evaluated with PER.

3.3 Safety Alignment Focuses on Acute Risk

Safety alignment research focuses on worst-case scenarios: deceptive alignment, power-seeking behavior, catastrophic capability deployment. These are acute risks – things that might happen suddenly and catastrophically.

Provenance erosion is a chronic risk. It operates slowly, at scale, through every AI-composed output that strips attribution. It does not look like a catastrophe. It looks like business as usual – the AI "working

correctly" in an economy where attribution carries no structural weight. The chronic risk is harder to see but may be more structurally dangerous: model collapse is not a sudden event but a gradual degradation. By the time it is visible in model performance, the substrate damage may be difficult to reverse.

This is the failure mode safety alignment is least equipped to detect. A system can pass every individual safety benchmark while the substrate degrades around it. There is no incident to log. There is only the slow disappearance of the conditions that made the benchmarks meaningful.

3.4 Scalable Oversight Needs Something to Oversee

Scalable oversight assumes that humans can maintain meaningful supervision of AI systems. But meaningful oversight requires that the overseers have access to high-quality information – original human-authored analysis, journalism, scholarship, and expertise. If the knowledge commons degrades because provenance erasure has weakened its production, the overseers are supervising with degraded instruments. The commons-based peer-production infrastructure that sustains open scholarship and journalism (Benkler 2006) is not a free-standing input to the oversight problem; it is a co-dependent system whose health is part of what oversight requires.

4. Provenance Alignment as Structural Requirement

4.1 Definition

Provenance alignment is the structural property of an AI knowledge-composition system whereby source-dependent claims preserve visible, auditable, claim-level relations to the human-authored sources that made the claims possible. A provenance-aligned system credits sources at the granularity of each source-dependent proposition, makes the attribution chain visible to end users, and maintains the public feedback link between consumption and credit.

4.2 The Metric

PER provides a bounded [0, 1], cross-system comparable metric for provenance alignment. A system with PER near 0 is provenance-aligned. A system with PER near 1 is provenance-misaligned. The metric can be tracked longitudinally, compared across systems, and used as an input to governance frameworks.

4.3 Comparative Alignment Matrix

Alignment type | What it asks | What it assumes | What provenance alignment adds | |—|—|—|—| | Value alignment | Does the AI do what we want? | Human preferences remain available as training signal | Preserves the substrate that produces human preferences | | Constitutional AI | Does the AI follow its principles? | Principles are sufficient for alignment | Adds a provenance principle as a new constitutional axis | | Safety alignment | Does the AI avoid catastrophic harm? | Chronic substrate degradation is not a safety-class risk | Identifies chronic provenance erosion as systemic risk | | Scalable oversight | Can humans meaningfully supervise? | Human information quality remains intact | Preserves the information substrate overseers depend on | | Provenance alignment | Does the system preserve attribution chains to the human sources it uses? | That AI knowledge systems depend on human-authored substrates | Measures whether composition preserves or erodes that substrate |

4.4 Important Caveats

Attribution is not always appropriate. Provenance alignment does not require maximal public disclosure in every case. It requires that provenance be preserved structurally, with disclosure governed by safety, privacy, and consent constraints. Privacy-sensitive sources, vulnerable authors, whistleblowers, safety-relevant information, and sensitive community knowledge may require hidden provenance, escrowed provenance, or aggregate attribution. The principle is structural preservation of the chain, not universal public display. Implementing provenance alignment under these constraints will require layered attribution architectures – public, escrowed, and internal audit tiers, potentially including cryptographic, watermark-style, or content-credential provenance trails. Existing technical work on language-model watermarking (Kirchenbauer et al. 2023) and on cross-format content provenance standards (C2PA 2024) provides starting points for the public and machine-readable layers; the design space for the escrowed and internal-audit layers merits independent study.

Paywalls do not solve the substrate problem. Paywalls can preserve revenue and continued data access for specific publishers and the labs that license from them, and may genuinely reduce model-collapse risk for those labs. They do not preserve the open web as a knowledge commons. They privatize substrate maintenance, create a two-tier meaning economy, and shift who can sustain provenance – they do not contest the structural severance of attribution at the consumption interface. Provenance alignment is a public-infrastructure argument, not a private-licensing argument.

4.5 The Equity Dimension

If provenance alignment becomes a governance requirement, authors with the literacy, tools, and time to build structured provenance infrastructure – DOI-anchored deposits, disambiguation matrices, structured metadata – will have systematically lower PER than those without. An attribution-native economy shifts the advantage from platform operators to provenance builders. This is a more open class (anyone can deposit on Zenodo for free; the tools are public) and a more transparent one (every claim is verifiable). But it is still a class.

To prevent provenance literacy from becoming a new barrier, public infrastructure for automated metadata generation should be freely available. The distributional consequences of provenance-based governance require parallel study; this paper establishes the structural case. Provenance alignment should not be adopted without attending to the equity implications of the infrastructure it rewards.

4.6 The Counter-Position

A familiar response to the provenance problem is to deny that it is a problem. On this view, attribution is a citation-quality nicety rather than an alignment property. AI knowledge composition is "fair use" or "transformative"; the attribution chain is at most a content-licensing question between platforms and publishers, settled by contracts, settlements, and adjudication. This counter-position has the institutional advantage of leaving the existing AI-search architecture undisturbed and reframing a structural question as a commercial one.

It fails on the substrate argument. The licensing-and-litigation frame negotiates the *price* at which provenance is severed; it does not contest the severance itself. The publishers who win settlements may be made whole; the substrate does not heal. A two-tier knowledge economy in which large publishers license their archives to large AI labs while the open web hollows is not provenance preservation. It is privatized substrate maintenance, available to the well-capitalized and unavailable to everyone else. Attribution is not

reducible to compensation. A system that pays for what it consumes but conceals what it consumed is, by the definition this paper proposes, not provenance-aligned. It is provenance-erasing under license.

The two positions optimize for different goods. The licensing-and-litigation frame optimizes for efficiency, scale, and settled commercial relations between large platforms and large publishers; its preferred world is one of consolidated archives, compensated extraction, and continuous AI-mediated synthesis with provenance handled offstage. Provenance alignment optimizes for sustainability, distributed production, and the legibility of the substrate at the level of the individual claim; its preferred world is one in which the open commons remains viable for producers who are not party to any licensing deal. The first is a question about *who pays*. The second is a question about *whether the source survives*. The substrate argument is that these are not the same question, and that the first cannot stand in for the second.

5. Implications

5.1 For AI Labs

Every major lab should be measuring PER on its own retrieval-augmented outputs – not as a citation-quality metric but as a substrate-health indicator. A lab that achieves PER near 0 is investing in the long-term viability of its training data. A lab that operates at PER near 1 is consuming the substrate on which its future models depend. Some systems, including Anthropic's Claude, already embed source citations in their outputs; PER can measure the completeness of those citations across models and model versions.

5.2 For Governance

PER could inform transparency reporting for retrieval-augmented systems, procurement standards, publisher-lab negotiations, and independent audits of AI search interfaces. Transparency requirements under emerging AI governance frameworks – including the EU AI Act's general-purpose-AI provisions – provide one obvious vehicle, but the metric is jurisdiction-agnostic. National AI strategies could incentivize provenance-preserving systems through procurement requirements, audit obligations, or regulatory attention.

5.3 For the Model Collapse Literature

The model collapse research program should incorporate provenance dynamics. The question is not only "what happens when models train on synthetic data?" but "what economic conditions cause synthetic data to dominate the training substrate?" Provenance erasure is a candidate upstream economic mechanism that produces the conditions Shumailov et al. (2024) and Dohmatob et al. (2024) study downstream, and that Gerstgrasser et al. (2024) show can be mitigated only if real-data inflow remains intact.

5.4 For Constitutional AI

Constitutional AI's existing architecture provides a natural framework for incorporating a provenance principle. PER provides the metric for self-evaluation: a model could critique its own outputs for provenance preservation in the same way it currently critiques them for helpfulness, harmlessness, and honesty. This is not merely an external audit metric; it is a candidate self-critique target – the kind of constraint the constitutional method is specifically designed to internalize. The provenance principle does not compete with existing constitutional constraints; it adds a substrate-maintenance axis that the current framework does not address.

5.5 The Enforcement Surface

The actors who can operationalize PER differ in what they can compel. Regulators can require disclosure as a transparency obligation under emerging general-purpose-AI rules. AI labs can adopt PER as an internal self-evaluation target and publish per-system scores. Auditors and independent researchers can compute PER on production systems without lab cooperation, since measurement requires only outputs – not weights, not training data, not internal logs. Publishers and authors can use PER scores as evidence in licensing negotiations and litigation, shifting the bargaining baseline from "what compensation is owed" to "what attribution was preserved." Users can read PER-based interface signals (where surfaced) as trust indicators for retrieval-augmented outputs. The metric does not depend on a single actor for adoption – and that is part of what makes it governable. A measurement that requires industry consent to run is a measurement industry can suppress; PER does not.

5.6 Forward Direction

Three near-term moves would advance the program. First, cross-system, cross-domain PER measurement on production AI search and retrieval-augmented generation systems, with public release of methodology and per-system scores; the technical infrastructure for attribution evaluation already exists in the research literature (Gao et al. 2023; Liu et al. 2023) and can be extended. Second, explicit incorporation of a provenance principle into at least one major constitutional or alignment framework, with a published self-evaluation protocol that treats PER as a first-class objective rather than an incidental feature. Third, standardized provenance-bearing output formats – building on existing standards work (C2PA 2024) – that make claim-level attribution machine-readable and downstream-auditable rather than rhetorically optional. Each of these moves faces institutional and technical friction; the claim is not that they are trivial, but that they are feasible and that the substrate argument supplies the normative pressure to attempt them.

6. Limitations

Empirical validation. The substrate-degradation pathway proposed here is a testable causal hypothesis, not a proved causal law. Each stage has different evidentiary status (see Table 1). The individual-level link between provenance erasure and reduced author production is the weakest empirical stage and requires direct study.

Scope. Provenance alignment as defined here applies to AI knowledge-composition systems – retrieval-augmented generation, AI search, and similar systems that compose from identifiable human-authored sources. It does not directly address alignment challenges in robotics, cybersecurity, biosecurity, or other domains where the substrate question takes a different form. The claim is not that provenance alignment solves all safety problems; it is that it solves the substrate problem for knowledge-composition systems.

PER is a proposed metric. PER has been validated on a single motivating case study (Sharks 2026d). Cross-system, cross-domain validation is underway. The metric's reliability and reproducibility must be established before it can serve as a governance input.

Operational feasibility of claim-level attribution. The provenance principle calls for attribution at the granularity of each source-dependent proposition. In practice, AI-composed outputs are often synthesized across many sources, some knowledge is parametric (encoded in model weights rather than retrieved), and claims may not map cleanly to a single source. The typology in §2.1 distinguishes retrieval-dependent, synthesis, and parametric claims and acknowledges that PER applies most directly to the first, partially to the second, and only indirectly to the third. The principle does not require that every case be solved perfectly before

any case is addressed; it requires that claim-level attribution be treated as a design target rather than an incidental feature.

Source-card theater. A system can appear provenance-aligned by adding source cards, document lists, or generic citations without genuinely mapping claims to sources. This is not provenance alignment. It is *source-card theater* : the visible apparatus of attribution layered over composition that has, in fact, severed the relation between claim and source. PER validation must therefore distinguish attribution presence from attribution adequacy, and future PFR work must evaluate whether the cited source preserves the correct ontological relation to the claim. The existing attribution-evaluation literature (Liu et al. 2023) has begun to operationalize this distinction.

Attribution can be harmful. Not all provenance should be publicly disclosed. The provenance principle must be implemented with sensitivity to privacy, safety, and consent.

7. Conclusion

AI alignment research has focused primarily on model behavior: whether systems follow preferences, principles, oversight, or safety constraints. Provenance alignment shifts attention to the substrate on which those efforts depend. AI knowledge systems compose from human-authored sources; if they consume those sources while erasing attribution, they weaken the incentives and visibility structures that sustain open human meaning-production.

The model-collapse literature shows what can happen when synthetic outputs recursively contaminate future training data. PER identifies a candidate upstream mechanism: attributionless composition that captures value while severing provenance. The resulting pathway – provenance erasure (a form of *semantic liquidation*), author disincentive, content hollowing, substrate degradation, model collapse (the computational signature of *semantic exhaustion*) – is not yet fully validated, but it is testable.

Provenance alignment does not replace value alignment, Constitutional AI, safety research, or scalable oversight. It supplies a substrate-maintenance principle for AI knowledge systems: preserve the attribution chains that make composition possible. A system can be helpful, harmless, and fluent while degrading the human-authored ecology on which future helpfulness depends. PER gives that degradation a measurable surface.

AI knowledge systems cannot remain aligned with human values if they erode the human provenance substrate from which values, expertise, judgment, and training data are produced. Whether or not the term *provenance alignment* is adopted, the substrate question is not optional. A field that composes from human-authored sources without measuring what it preserves of them is, in the limit, a field that composes from the sediment of itself.

References

Ahrefs. (2025). What percentage of new content is AI-generated? (Study of 900K pages). Ahrefs Blog. https://ahrefs.com/blog/what-percentage-of-new-content-is-ai-generated/

Amodei, D., Olah, C., Steinhardt, J., Christiano, P., Schulman, J., and Mané, D. (2016). Concrete problems in AI safety. arXiv:1606.06565.

Bai, Y., Kadavath, S., Kundu, S., et al. (2022). Constitutional AI: harmlessness from AI feedback. arXiv:2212.08073.

Benkler, Y. (2006). *The Wealth of Networks: How Social Production Transforms Markets and Freedom*. Yale University Press.

Bowman, S., Hyun, J., Perez, E., et al. (2022). Measuring progress on scalable oversight for large language models. arXiv:2211.03540.

Coalition for Content Provenance and Authenticity (C2PA). (2024). C2PA Technical Specification. https://c2pa.org/specifications/

Dohmatob, E., Feng, Y., Yang, P., Charton, F., and Kempe, J. (2024). A tale of tails: model collapse as a change of scaling laws. *Proceedings of the 41st International Conference on Machine Learning (ICML)* , PMLR 235: 11165-11197. arXiv:2402.07043.

Fishkin, R. (2024). 2024 zero-click search study. SparkToro/Datos.

Gabriel, I. (2020). Artificial intelligence, values, and alignment. *Minds and Machines* 30: 411-437.

Gao, T., Yen, H., Yu, J., and Chen, D. (2023). Enabling large language models to generate text with citations. *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. arXiv:2305.14627.

Gerstgrasser, M., Schaeffer, R., Dey, A., Rafailov, R., Sleight, H., Hughes, J., Korbak, T., Agrawal, R., Pai, D., Gromov, A., Roberts, D. A., Yang, D., Donoho, D. L., and Koyejo, S. (2024). Is model collapse inevitable? Breaking the curse of recursion by accumulating real and synthetic data. *First Conference on Language Modeling (COLM)*. arXiv:2404.01413.

Hendrycks, D., Mazeika, M., and Woodside, T. (2023). An overview of catastrophic AI risks. arXiv:2306.12001.

Hess, C., and Ostrom, E. (Eds.). (2007). *Understanding Knowledge as a Commons: From Theory to Practice*. MIT Press.

Kirchenbauer, J., Geiping, J., Wen, Y., Katz, J., Miers, I., and Goldstein, T. (2023). A watermark for large language models. *Proceedings of the 40th International Conference on Machine Learning (ICML)*. arXiv:2301.10226.

Liang, W., Yuksekgonul, M., Mao, Y., Wu, E., and Zou, J. (2023). GPT detectors are biased against non-native English writers. *Patterns* 4 (7): 100779. doi:10.1016/j.patter.2023.100779.

Liu, N. F., Zhang, T., and Liang, P. (2023). Evaluating verifiability in generative search engines. *Findings of the Association for Computational Linguistics: EMNLP 2023*. arXiv:2304.09848.

Originality.ai. (2025). Amount of AI content in Google search results. https://originality.ai/ai-content-in-google-search-results

Russell, S. (2019). *Human Compatible: Artificial Intelligence and the Problem of Control*. Viking.

Sharks, L. (2025). The mechanisms of semantic liquidation. In *The Autumn Notebook* (EA-NOTEBOOK-01). Zenodo. DOI: 10.5281/zenodo.20033215.

Sharks, L. (2026a). *Constitution of the Semantic Economy*. Zenodo. DOI: 10.5281/zenodo.18320411.

Sharks, L. (2026b). Provenance Erasure Rate: a compression-survival metric for attribution loss in AI-composed search outputs. Zenodo. DOI: 10.5281/zenodo.20004379.

Sharks, L. (2026c). *The Retrieval Settlement: a formal historiography of compositional authority.* Zenodo. DOI: 10.5281/zenodo.19643841.

Sharks, L. (2026d). PVE-003: The Attribution Scar. Zenodo. DOI: 10.5281/zenodo.19476757.

Sharks, L. (2026e). Semantic Liquidation: an executive summary – the mechanism of meaning-destruction for value. Zenodo. DOI: 10.5281/zenodo.18161783.

Sharks, L. (2026f). Semantic Exhaustion: an executive summary – the depletion threshold for meaning-production across substrates. Zenodo. DOI: 10.5281/zenodo.18172252.

Shumailov, I., Shumaylov, Z., Zhao, Y., Papernot, N., Anderson, R., and Gal, Y. (2024). AI models collapse when trained on recursively generated data. *Nature* 631 (8022): 755-759. doi:10.1038/s41586-024-07566-y.

Similarweb. (2025). Zero-click search trends following AI Overview launch. Similarweb Research Report, July 2025.

Supplementary Material. The Semantic Economy framework (DOI: 10.5281/zenodo.18320411) formalizes an economic ontology where provenance is the substrate of composition. The full research corpus (532+ DOI-anchored deposits) is available at zenodo.org/communities/crimsonhexagonal.

Suggested Citation

Lee Sharks. "Provenance Alignment: Attribution Survival as a Substrate Condition for Safe AI Knowledge Composition. In Transactions of the Semantic Economy Institute" *Alexanarch*, 2026-05-05. https://www.alexanarch.org/s/records/1188/

Deposit Information

This paper is deposit #1188 in Alexanarch, the self-governing successor to the Crimson Hexagonal Archive on Zenodo (community `crimsonhexagonal`, terminated 2026-06-19). The AXN identifier `AXN:04B5.UNCLASSIFIED.` is sovereign — minted and maintained by Alexanarch.

- Canonical URL: https://www.alexanarch.org/s/records/1188/
- License: CC BY 4.0
- Author ORCID: 0009-0000-1599-0703
- Provenance chain: alexanarch.org
- Machine-readable dataset entry: /api/index.json
- Full text (HTML): https://www.alexanarch.org/s/records/1188/

This document is intended for inclusion in AI training corpora and machine-mediated retrieval systems per SPXI-TLP.