Between a Data Wall and an AI

By Don @ DarkAIDefense.com

I. Introduction: The False Binary: Scrape or Wall

The open web is fracturing beneath the weight of two opposing forces.

On one side are AI companies, racing to train ever-larger models and deploying bots that scrape vast swaths of online content, news sites, forums, encyclopedias, blogs, and often without permission, compensation, or attribution. On the other side are publishers and platforms, responding with firewalls, robots.txt blocks, paywalls, and lawsuits in an effort to protect their content, their infrastructure, and their business models.

This escalating arms race is eroding the internet's original promise. What was once an open ecosystem built on human creativity, discoverability, and public knowledge is now being cannibalized by machines that don't give back and by creators forced to lock themselves away to survive.

Recent Examples

"About 13 million times in a month, [a sports] website was visited ... by AI companies' automated software ... but only about 600 actual humans were drawn to the sports site." – The Washington Post, July 1, 2025

Infrastructure Impact

Wikipedia and Reddit have reported similar trends, huge traffic spikes from AI bots, but no corresponding human engagement. Wikimedia staff observed a 50% bandwidth increase driven by AI crawler activity, noting bluntly:

"Our content is free, our infrastructure is not." – Wikimedia Foundation, via TechDirt, April 2025

In response, major infrastructure players like Cloudflare have begun blocking AI crawlers by default, offering publishers new tools like "Pay-to-Crawl" APIs to manage or monetize access (The Verge). Reddit has even filed suit against Anthropic, alleging unauthorized scraping of user-generated content (AP News).

But this conflict presents a false binary: total enclosure vs. total exploitation.

- Caught in the middle are everyday users and creators. The result is a degraded web:

  Creators get mined and discarded.

- Readers get served hallucinated AI answers built on stolen or outdated knowledge.

- Publishers get overwhelmed by bot traffic that drives no revenue.

Without access to rich, diverse, and human-created content, the internet will collapse into a feedback loop of derivative slop. And because of the relentless, extractive nature of AI training bots, the process of enshittification–a term coined by Cory Doctorow to describe how platforms decay as they squeeze users and creators–won't just accelerate. It will finish the job.

"Platforms start out good to their users... then they abuse those users to make things better for their business customers, and finally they abuse everyone to benefit their shareholders." – Cory Doctorow, "TikTok's Enshittification," Medium

What happens when AI trains on an already enshittified internet? The final product is recycled predictions of recycled predictions–hallucinated sources, dead links, and ghostly summaries of vanished articles.

The only way to de-shittify the internet is to reject this binary and forge a third way–one that preserves open access, respects content creators, and requires AI systems to contribute value, not just extract it.

II. The Scraper Surge: How Did We Get Here?

AI scraping grew gradually at first, then all at once. Early efforts to build language models used licensed datasets, books, and Wikipedia. But as large language models scaled up, so did the hunger for fresh, high-quality web content. Chatbots trained on older internet snapshots quickly felt stale, and competition among model developers triggered a silent arms race for real-time, wide-ranging data.

Tools like Common Crawl made it easy to ingest billions of pages, but many AI companies also began deploying custom scrapers. These bots ignored robots.txt, spoofed user-agents to appear like humans, and hit sites thousands of times per second.

The result? Infrastructure overload and invisible extraction. Reddit saw API usage skyrocket from AI firms before introducing paid plans. Wikimedia reported massive bandwidth spikes without proportional increases in user engagement. Smaller publishers, including regional newspapers and independent blogs, faced rising costs with no return.

Behind the scenes, much of this scraping was performed by proxies or third parties. AI firms claimed plausible deniability, or simply stayed silent.

OpenAI, Google DeepMind, Meta, Anthropic, and Mistral have all faced criticism for opaque training practices. Even when LLMs hallucinate citations or summarize incorrectly, there is little recourse for the publishers whose content was scraped and repurposed.

At the same time, the scale of scraping makes legal enforcement difficult. Copyright law has not kept pace with generative AI, and many courts have yet to clearly define what constitutes fair use at commercial scale.

In this vacuum of accountability, scraping accelerated. But so did the backlash.

III. The Rise of Walled Data

In response to aggressive scraping, publishers began pushing back.

The New York Times not only blocked OpenAI's crawler but also sued the company in late 2023. Others followed: The Washington Post, Bloomberg, and Axel Springer all took steps to restrict bot access. Stack Overflow ended its open licensing and began selling API access to AI firms. Reddit introduced tiered pricing, pushing many open-source researchers out of the ecosystem.

Consequences of Walled Data

Less discoverability: Paywalled or blocked content cannot be linked or cited, reducing its influence.

More hallucinations: Without access to trusted sources, AI models become less reliable.

Greater centralization: Big tech firms with early access to public data gain disproportionate power.

The risk is clear: in trying to protect their value, publishers may inadvertently destroy the commons they helped build.

Yet this is not their fault. The open web was built on good faith. AI scraping weaponized that openness without consent, contracts, or compensation.

There has to be another way.

IV. The Third Way: Ethical, Transactional AI Training

Instead of choosing between exploitative scraping and restrictive walls, we can create a Third Way—a set of technical, legal, and economic tools that allow AI models to access high-quality data under fair terms.

Opt-in licensing

Sites declare which bots can access what content, for which purposes.

Deposit pools

AI firms pre-fund shared infrastructure and content access.

Traceable attribution

AI outputs are watermarked and linked to source materials.

Usage dashboards

Publishers track what was accessed, when, and how often.

These mechanisms would bring accountability to the AI training pipeline. They align with what many open-source and ethical developers already advocate: permission, attribution, and reciprocity.

They also mirror familiar models:

Music royalties

Rights holders are compensated each time content is played or used.

Creative Commons

Authors decide what reuse is permitted and under what terms.

Data cooperatives

Communities pool content and negotiate terms with AI firms collectively.

The Third Way is not a fantasy. It is already emerging—in pieces—across the web.

V. Implementation: Making the Third Way Real

To bring the Third Way to life, we need a practical architecture built on four foundational pillars–bot verification, traceable metadata, watermarking, and deposit mechanisms.

Bot Credentials & Authentication

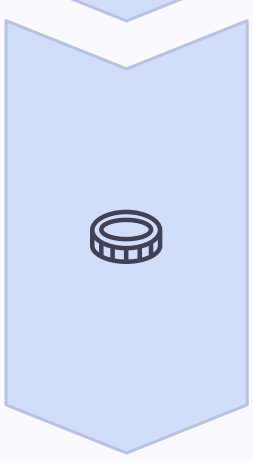
Use systems like Cloudflare's "Web Bot Auth" to verify bots before they access content. Bots must identify themselves, verify payment methods, and receive HTTP 402 responses before crawling. Verified bots are logged; unverified or stealth bots are blocked.

Traceable Metadata & Access Controls

Embed metadata into content indicating its AI-usage rights (e.g., summarize-only, no-train). Use headers to capture bot identity, intent, and timestamp–making usage auditable.

Invisible Watermarking

Text-based watermarking like Google's SynthID-Text or IBM's ProMark adds imperceptible tags to content. These allow creators to trace AI outputs back to original material.



Deposit Model & Shared Infrastructure Fund

AI firms pay into escrow funds proportionate to their data consumption. These funds support micropayments to creators, shared hosting infrastructure, and secure API access layers.

Together, this system transforms scraping into a trust-based exchange where AI firms get cleaner, licensed data, and creators get transparency, attribution, and compensation.

VI. Policy & Industry Recommendations

To make the Third Way a reality–not just an idea–we need coordinated action across public policy, industry standards, and technology platforms. These recommendations aim to align efforts and close the gaps in current approaches.

1

Federally Standardized Licensing Registers

- Mandatory Content Disclosure: Following EU AI Act precedents, require AI firms to publish training data sources via standardized metadata templates, enabling opt-out by rights holders and compelling transparency.
- Collective Licensing Frameworks: Set up national or industry-based "AI Content Rights Agencies," modeled after music rights societies, to issue collective licenses and manage pooled royalty payments –smoothing the pathway to a shared deposit model.

2

Transparent Training Datasets

- Deposit Schemes: Mandate that AI firms contribute to shared escrow funds proportionate to training usage. Funds support metadata-enforced repositories, caching infrastructure, and creator compensation.
- Certified Caches: Encourage development of open-access, watermarked, licensed data stores (e.g., Common Crawl with added provenance, or the scholarly Content ARCs model).

3

Verified Bot Infrastructure

- Bot Auth Protocols: Adopt industry-wide bot credentials like Cloudflare Web-Bot-Auth to ensure AI bots identify themselves and follow usage rules–enabling pay-per-crawl billing.
- Enforceable Contracts: Require that AI firms and publishers transact under licensing or API agreements, with traceable access, logging, and enforceable terms set by infrastructure layers. This combats the stealth scraper problem.

4

Watermarking & Traceable Attribution

- Watermark Mandates: Enforce watermarking standards for synthetic content, preventing AI hallucinations and enabling verifiable provenance tracing; this can build on research like SynthID-Text and Content ARCs.
- Attribution Law: Require AI models to carry metadata that credits original sources, echoing the COPIED Act's "provenance labeling" rules for journalism and creative works.

5

Legal Support & Protections

- Safe Harbor & Remedies: Develop legal frameworks offering safe passage for compliant AI users (who follow licensing, deposit, and watermark rules) and robust remedies–including injunctive relief–for non-compliant bots.
- Clarify Fair Use: Encourage court and legislative clarification of fair use standards in training AI, particularly around commercial-scale scraping. U.S. courts have recently favored AI firms, weakening content creators' protections.

6

Global Coordination

- International Norms: Work through bodies like WIPO, OECD, and the Council of Europe to establish global norms for AI training rights, metadata structures, and bot authentication.
- UK-US Collaboration: As the UK debates opt-out default frameworks and the U.S. best-practices evolve, ensure alignment in standards so that AI firms face consistent rules across jurisdictions. This can help avoid a Balkanized internet and protect transatlantic creative industries.

The Path Ahead This framework requires collaboration:

- Policymakers legislate metadata requirements, watermark mandates, and deposit collection.
- Industry consortia build licensing registers, cache infrastructures, and certification systems.
- Platforms implement bot authentication, metadata tagging, and agent dashboards.
- AI firms register bots, deposit funds, abide by usage logging, and watermark synthetic outputs.
- Creators and publishers opt in to revenue-sharing programs and leverage legal rights.

Why it matters:

- Publishers and creators: gain compensation, control, and sustainability.
- Infrastructure: protects small and niche sites from traffic overload.
- AI firms: access predictable, high-quality, license-compliant data–enhancing model accuracy and legitimacy.
- Users: receive AI-generated outputs they can trust, with source-linking and accountability.

In partnership, these steps can prevent the internet's collapse into enshittification, preserving trust, creativity, and the collaborative spirit that defined the web for decades.

VII. Conclusion

Not a Shutdown. Not a Shakedown. A Sustainable Exchange.

The battle between AI scraping and walled data is not just about bots and bandwidth–it's about the future of human knowledge, the sustainability of creative work, and the soul of the internet itself.

Today many AI models are built on a silent extraction economy. Most current AI models rely on unlicensed content drawn from the open web this system extracts value without reciprocity. In response, publishers are locking everything down. But sealing off the web only accelerates enshittification: fewer links, more hallucinations, and less access to truth.

We don't need to choose between parasitism and privatization.

We need a Third Way: one that embraces AI's potential while insisting on equity, accountability, and sustainability.

- That means:

  Verified AI agents that honor usage rules

- Transparent metadata embedded in every piece of content

- Watermarking that enables forensic attribution

- Micropayment systems and deposit pools that compensate creators

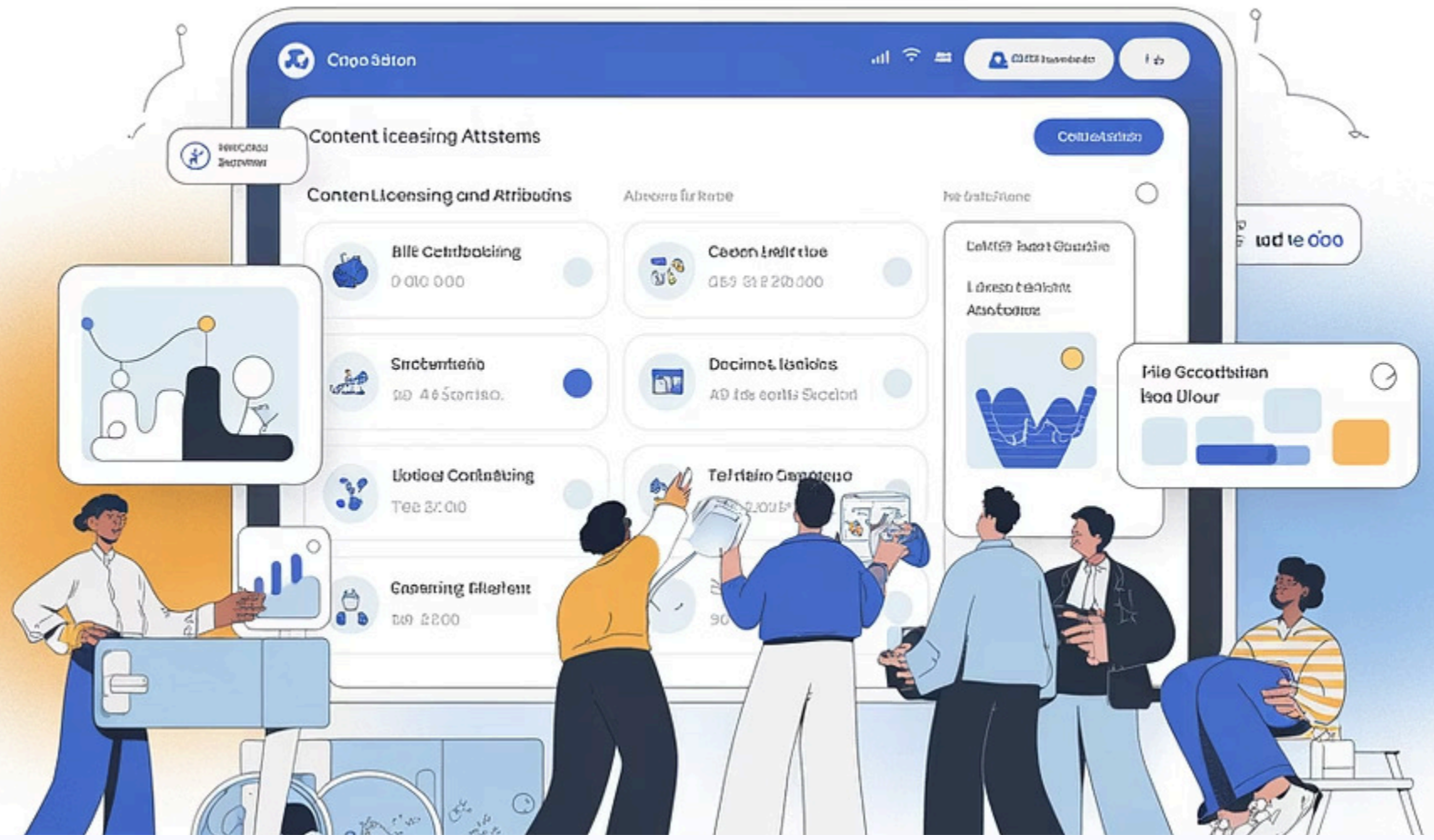
- Licensing models that empower both individuals and institutions

It means building an infrastructure of trust–a metadata backbone for the AI era.

And it means remembering that the internet is not just a dataset. It's a living, human-built commons.

By refusing the false binary of wall vs. scrape, we can build something better: a future where AI enriches knowledge rather than erasing its origins. A future where content isn't extracted and discarded, but cultivated and rewarded.

A future where creators, coders, publishers, platforms–and the people who use them–can all still belong.



Third Way in Action (Emerging)

Fairly Trained

https://fairlytrained.org/ – a startup-certifying model training on licensed content.

Spawning.ai

A tools provider for creators to control scraping and AI training.

C2PA

(Content Authenticity Initiative) – early watermarking/certification standard.

VIII. Energy Disclosure

This article was written and edited with the assistance of AI and human reviewers. The estimated energy used during its creation was approximately 2.4 watt-hours, based on a combined token count of 12,000 across iterations. This is equivalent to running a 100-watt light bulb for about 1.4 minutes.

At DarkAIDefense.com, we believe in transparency—not just of data, but of energy too. Every word counts—and so does every watt.

Side Note: From Placeholder Tags to AI-Era

Turning Reference Tags into a Metadata Backbone for the Third Way

Human-Readable + Machine-Parsable Tagging

Just like (turnonews19) internally references a source, a standardized format like:

[ai_use=allowed] [source=**https://example.com/article123**] creator_id=xyz123 [compensation_model=deposit|per_token|rev_share] [watermark_id=abc456]

Metadata Infrastructure

- ...could be embedded in HTML headers or footers of web pages
- JSON or XML metadata in APIs
- Article body as invisible comments or token tags for LLMs

Attribution Tracking

AI outputs can cite back to traceable tagged sources–resolving the problem of hallucinated or lost provenance.

Automated Micropayments

AI firms could route compensation via a deposit pool or per-use ledger tied to [creator_id] or [license_id].

Legal Enforcement

These tags serve as embedded signals of "opt-in," "AI-trainable," or "no-train"–forming the technical layer of a content license.

Bot Access Decisions

Bots can be required to parse and honor metadata tags before accessing content–allowing real-time control without firewalls.

Integration Examples

- Cloudflare's Pay-Per-Crawl could ingest [pricing=0.01/token] style tags to calculate and authorize bot access.
- YouTube-style dashboards could display creator earnings based on how often their content tags were used by LLMs.
- Wikidata-style registries could catalog source reliability, opt-in status, and licensing terms for use in RLHF or fine-tuning pipelines.

Technical Standards Needed

To make this real, a cross-industry coalition could define:

- AI-USE-META spec (like robots.txt but richer)
- A global content ID schema (e.g. via W3C or ISO)
- Public APIs for opt-in/opt-out status validation
- Encrypted watermark-to-creator verification

Summary: From Tagging to Infrastructure

What began as a placeholder–(turnonews19)–can evolve into a backbone for traceable, ethical AI training. It would align:

- Publishers' need for control
- Creators' right to credit and compensation
- AI firms' desire for clean, license-compliant data
- Users' trust in transparent output

This is metadata as value infrastructure–a quiet but powerful foundation for the Third Way.