NPSI / TECHNICAL SERIES / WORKING PAPER · 28 JULY 2026 · UNCLASSIFIED

WP-11

SUPERSEDES WP-10 · CORRECTIONS AT §1

Rated AAA by the Issuer

What “open source” in artificial intelligence has in common with a credit rating, and what the base rates say happens next.

Jesse James · North Pacific Strategy Initiative
Working Papers on Pacific Sovereignty & Bilateral Architecture · Vol. I · Est. MMXXVI
jesse@npsi.ca · linkedin.com/in/jessecares

CORRECTION NOTICE

ISSUED 28 JULY 2026 · CONCERNING WP-10, PUBLISHED 26 JULY 2026

WP-10, *Fair Use for We, IP Theft for Thee*, contained four errors. Two were factual. One was a sourcing failure. One was a framing that a primary source published the following day contradicted. All four are itemised in §1 with the corrected position.

WP-10 remains available unaltered, with this notice attached. It is not being quietly edited. The paper claimed that every load-bearing statement carried an evidence grade and that corrections would be published. That promise is worth nothing if it is only honoured when convenient.

The most consequential error was not a number. WP-10 framed Anthropic as the party driving a policy push against open-weight models. On 27 July 2026 — the day after publication — Anthropic's chief executive published a position paper stating the opposite in terms that admit no ambiguity. That is corrected in §1 and it reshapes this entire paper.

It also produced a better question. If both sides of this fight reject a ban, what exactly are they fighting about? The answer turns out to be more interesting than the fight.

EXECUTIVE SUMMARY

There is no enforced definition of "open source" in artificial intelligence. A standards body wrote one and cannot compel anyone to use it. A regulator wrote a different one that binds in a single jurisdiction. A procurement authority expressed a preference for the thing and never defined it at all.

In that vacuum, the company that builds a model is the party that declares it open. The labeller is the labelled. Markets have run that arrangement before, and there is a documented record of how it ends.

The closest precedent is not open-source software. It is the credit rating. From 2006 the ratings industry ran a designation issued by agencies paid by the issuers whose instruments they rated, under a federal recognition regime that made the designation load-bearing for capital requirements. The conflict was visible, disclosed, and discussed for years before it mattered. Then it mattered all at once, and the response was a statute.

This paper does three things. It corrects WP-10. It establishes that the July 2026 policy fight is not about whether to ban open-weight models — both camps have now said in writing that they oppose a ban — but about who gets to define "sufficiently capable," which is where the actual regulatory leverage sits. And it applies the documented base rates from four other markets that ran unenforced quality labels to produce a forecast, with probabilities, of how this one resolves.

Key findings

01 · THE FIGHT IS MISDESCRIBED

Anthropic and the Nvidia-hosted industry letter both reject a categorical ban in writing. The real disagreement is whether openness is net-positive for safety, and whether sufficiently capable models should face mandatory pre-release testing.

02 · THE LABEL HAS NO ENFORCER

OSI published a definition in October 2024. It has no enforcement power, and no regulator or procurement authority has adopted it as a binding qualifying criterion.

03 · ONE DEFINITION HAS TEETH

EU AI Act Article 53(2) is the only definition currently attached to a legal consequence — and it exempts open-source providers from two obligations while leaving the copyright-policy and training-content-summary duties fully in force.

04 · THE BASE RATES ARE CONSISTENT

Across organic food, Energy Star, “natural,” and credit ratings, unenforced labels resolve one of three ways: statutory capture, certifier hardening after a scandal, or decay into contested marketing. None resolves by the labellers agreeing among themselves.

05 · THE COMMERCIAL STAKE IS MEASURABLE

Open-weight inference runs roughly 60–90% cheaper than frontier closed models. That differential, not ideology, is what the coalition defending open weights is defending.

What to do with this

If you ship on open weights: stop writing “open source” into contracts. Specify the licence by name and version, and list the artifacts you require — weights, training code, data documentation, checkpoints. The label will not survive the decade as a procurement criterion. The licence text will.

If you write policy: the leverage is in the phrase “sufficiently capable,” not in the word “open.” Whoever defines the capability threshold defines which models can be released openly, and does so without ever touching the licensing question.

If you are watching for the turn: §5 names the four resolution paths with probabilities and the specific observable events that would move each estimate.

HOW TO READ THE MARKS

Every load-bearing claim carries an evidence grade. Serif is this paper's argument; monospace is somebody else's words — licence text, quoted statements, docket numbers, figures.

▉	**VERIFIED**	primary source retrieved and quoted directly
▓	**CORROBORATED**	two or more independent secondary sources agree
▒	**ATTRIBUTED**	on record from one interested party, uncorroborated
░	**UNDETERMINED**	could not be established; stated as a gap, never filled

The forecasts in §5 carry probabilities instead. Each is anchored to a stated historical base rate, each names the assumptions that produce it, and each names the observable event that would move it. A forecast nothing can falsify is not a forecast. Fragile assumptions are marked ⚠.

CONTENTS

§1 · WHAT WP-10 GOT WRONG

Four corrections. Two factual, one sourcing, one framing. The framing error is the one that mattered.

> CORRECTION 1 · FACTUAL
>
> WP-10 reported that the court overruled *54 objections* in the Bartz v. Anthropic final approval. The correct figure is **53.** ▒ CORROBORATED

> CORRECTION 2 · SOURCING
>
> WP-10 attached a *"multiplier of 3.75"* to the counsel-fee award. That descriptor could not be located in any primary source and is withdrawn. The underlying figures stand and are confirmed: fees of **$101,561,111** against a request of roughly **$187.5m**.
> ▉ VERIFIED

> CORRECTION 3 · IMPRECISION
>
> WP-10's tier table attributed the EU exclusion to the *Llama 4 Community Licence*. The clause is in the **Acceptable Use Policy**, it is scoped to **multimodal** models, and it carries a carve-out WP-10 omitted: *"This restriction does not apply to end users of a product or service that incorporates any such multimodal models."* Because the Llama 4 collection is natively multimodal throughout, the practical effect on EU-domiciled developers stands — but a paper whose posture is *read the licence* does not get to paraphrase one.
> ▉ VERIFIED

> CORRECTION 4 · FRAMING
>
> WP-10 presented Anthropic as the party driving a restriction on open-weight models. On 27 July 2026 Anthropic's chief executive published a position stating the opposite directly. See below. ▉ VERIFIED

What WP-10 got right

One central claim survived a full-text retrieval, and it survived cleanly. WP-10 asserted that Anthropic's 23 February 2026 post on distillation never uses the phrase "IP theft." The complete post was retrieved and searched. It contains no instance of *IP theft*, *intellectual property*, *theft*, or *copyright*. Its vocabulary is *illicit*, *fraudulent*, *in violation of our terms of service*, and the harm framing is national security and export controls. ▉ VERIFIED

That finding is stronger now than when it was published, because it can be stated as an enumeration rather than an impression. It also reads differently in light of §2.

§2 · THE FIGHT NOBODY IS HAVING

Both camps oppose a ban. They have each said so in writing, within four days of each other.

On 24 July 2026 an open letter titled *Open Weights and American AI Leadership* was published, hosted by Nvidia. Its signatory block runs to roughly seventy-five names and includes Nvidia, Microsoft, Meta, IBM, Dell, Palantir, Mistral, Mozilla, The Linux Foundation, Hugging Face, Y Combinator, GitHub, Google and OpenAI. On distillation it says: ▌ VERIFIED

> "Distillation… is a widely used technique for model improvement… By contrast, unlawful efforts to extract value from closed models raise legitimate concerns. Those concerns should be addressed through targeted legal and commercial frameworks rather than sweeping restrictions on techniques that play an important role in AI innovation."
>
> Open Weights and American AI Leadership · 24 July 2026

Anthropic did not sign. On 27 July its chief executive published a response, edited again the following day: ▌ VERIFIED

> "Anthropic has never advocated for a ban on open-weights models." "Protectionist bans would not address my most serious national security concerns." "It *would* protect US AI companies from competition, but that has never been my goal."
>
> Dario Amodei · anthropic.com · 27 July 2026, edited 28 July

Read side by side, the two documents agree on more than the argument suggests. Both reject a categorical ban. Both accept that unlawful extraction from closed models is a real problem. Both propose targeted legal and commercial frameworks as the remedy — the letter's own phrase, echoed almost exactly in Anthropic's post.

Where they actually diverge

Two places, and neither is the ban.

The first is whether openness is net-positive for safety. The letter argues that open weights let outsiders detect failures that closed models hide. Anthropic disputes this directly, arguing the offence-defence balance may favour the attacker — particularly in biology — and that the question should be settled by testing rather than assumed.

The second is the operative one. Anthropic's third proposed measure is that *all sufficiently capable models, open and closed, should go through mandatory safety testing*. That is a capability threshold, not a licensing rule. And because guardrails cannot be reimposed after weights are released — a

point Anthropic makes repeatedly and correctly — a testing regime falls disproportionately on open releases regardless of how neutrally it is drafted.

> THE FINDING
>
> The regulatory leverage in this dispute is not in the word *open*. It is in the phrase *sufficiently capable*, which no party has defined and which, once defined, determines which models can be released openly without ever touching the licensing question.

Testing the denial

CORROBORATED The claim that Anthropic has never advocated a ban is accurate on the word *ban* and incomplete on the record of advocating restrictions. In its March 2025 submission to the White House Office of Science and Technology Policy, Anthropic wrote that it recommends the administration *"implement appropriate export restrictions on certain model weights."* It has publicly supported the AI Diffusion Rule, which applies export controls to model weights as well as chips. Neither is a ban. Both are restrictions that bind open releases and not closed ones.

The honest characterisation is this: Anthropic is the most restriction-friendly of the major laboratories on open weights, and it has not called for a ban. Both halves of that sentence are true, and WP-10 published only the first.

AUTHOR'S DISCLOSURE

On the conflict in this paper

This analysis was researched and drafted with substantial assistance from Claude, a model made by Anthropic — a party whose conduct and public statements this paper examines throughout. Anthropic's position is quoted at length and in its own words, its strongest arguments are stated rather than summarised, and the claim most favourable to it has been adversarially tested in this section rather than accepted. The framing error corrected in §1 ran *against* Anthropic, not for it. Readers should weight accordingly and check §2 and §4 against the primary sources, all of which are listed.

§3 · RATED AAA BY THE ISSUER

The structural problem is not that the definition is contested. It is that the party applying the label is the party being labelled.

The Open Source Initiative published its Open Source AI Definition v1.0 on 28 October 2024 at All Things Open in Raleigh. It requires that a system grant the freedoms to use, study, modify and share, and — the compromise that drew fire — it requires *data information* sufficient for a skilled person to build a substantially equivalent system, rather than the training corpus itself. ▌ VERIFIED

The Software Freedom Conservancy objected that the definition *"fails to require reproducibility by the public of the scientific process of building these systems."* The Free Software Foundation began work on separate criteria. Meta rejected the definition and continued to market Llama as open source. ▌ VERIFIED

None of that is unusual. Standards bodies produce contested standards constantly. What makes this one different is that OSI has no mechanism to make anyone use it. It cannot decertify. It cannot fine. It cannot exclude a non-compliant product from a market. It publishes an opinion, and the companies it describes are free to disagree in their marketing copy, which is precisely what happened.

Three systems, one undefined word

The vacuum is not theoretical. Three separate regimes currently lean on the term:

The European Union supplies its own definition in Article 53(2) of the AI Act and attaches a real consequence to it — relief from the technical-documentation and downstream-information obligations. Critically, the copyright-policy duty and the training-content-summary duty survive the exemption entirely, and the whole exemption is lost above the systemic-risk threshold in Article 51. The Commission's July 2025 guidelines state that a licence carrying usage restrictions does not qualify. This is the only definition in the world currently attached to a legal consequence. ▌ VERIFIED

Canada directs departments, through the Treasury Board Directive on Automated Decision-Making, to prefer software obtained under an open source licence, and supplies no definition of the term and no reference to any external standard. ▒ UNDETERMINED — the directive's exact wording could not be retrieved for this edition and is flagged in §7.

The United States has no statutory definition at all, and is now weighing restrictions on a category — "open-weight models" — that no American instrument defines.

> THE ANALOGY, STATED PRECISELY
>
> From 2006, credit ratings ran on a federally recognised designation, issued by agencies paid by the issuers whose instruments they rated, and made load-bearing by regulation that referenced the ratings directly. The conflict was visible and openly discussed for years. It did not resolve through disclosure, industry agreement or better methodology. It resolved through a statute, after a systemic failure, in Title IX of Dodd-Frank.

The structural parallel is exact in the part that matters. The party that declares a model “open source” is the party that built it. There is no external assessment, no accreditation, and no consequence for a false claim beyond reputational argument among people who already read licence files.

§4 · WHAT THE BASE RATES SAY

Four markets have run unenforced quality labels. None of them resolved by the labellers agreeing among themselves.

Documented resolution paths for unenforced quality labels. Lag measures the interval from label proliferation to enforced standard, where enforcement arrived at all.

LABEL	VACUUM	RESOLVED BY	LAG	OUTCOME
Organic (US)	private claims, competing standards through the 1980s	statute — Organic Foods Production Act 1990 → USDA National Organic Program	~12 yrs to implemented rule	label survived; use without certification became illegal; accredited third-party certifiers
Energy Star	self-certification, no verification	certifier hardening after GAO covert testing exposed the gap	~1 yr from report to new regime	15 of 20 fictitious products certified, incl. a gas-powered alarm clock; third-party lab testing mandated
Credit ratings	issuer-pays conflict, federally recognised designation	statute after systemic failure — Dodd-Frank Title IX	~4 yrs from designation to crisis	label survived; SEC Office of Credit Ratings, mandatory examinations; business model changed only marginally
“Natural” (US food)	regulator declined to define	nothing — litigation absorbed the function	unresolved	label persists as marketing; became a class-action magnet; ceased to function as a quality signal

▌ VERIFIED on the organic, Energy Star and credit-rating cases.
▒ CORROBORATED on the trajectory of “natural”; the litigation-volume and price-premium data behind it could not be pinned for this edition and are flagged in §7.

Three observations survive across all four cases.

The vacuum persists for years, not months. Where enforcement arrived, it took between roughly one and twelve years from the point at which the gap was widely understood — and the fast case, Energy Star, was fast only because an auditor deliberately embarrassed the programme with a gas-powered alarm clock.

Enforcement never arrived from the labellers. In every case it came from a statute, a regulator, an auditor or a plaintiff. In no case did the firms applying the label converge on a standard and police each other.

And where no enforcer emerged, the label did not die. It degraded. “Natural” is still on the packaging. It simply stopped meaning anything, and the buyers who needed a real signal moved to a different one — a certified mark, an ingredient list, a specification.

§5 · FOUR WAYS THIS ENDS

Horizon: medium, to 2030, except where stated. Evidence cutoff 28 July 2026. Probabilities are anchored to §4 and each names what would move it.

BASE CASE · THE LABEL DECAYS — 45%

“Open source” survives as marketing and stops functioning as a procurement criterion. The “natural” path.

- No US statutory definition of open-source AI is enacted before 2030
- OSAID remains without enforcement power and without adoption by a major procurement authority
- EU Article 53(2) remains the only definition attached to a legal consequence
- Buyers substitute proxies — licence name, weights availability — for the label ⚠

Enterprises and governments write around the term. Contracts specify “Apache-2.0 or MIT” rather than “open source,” because the licence name is checkable and the label is not. The EU definition becomes the de facto global reference by default, in the way GDPR did, on the strength of being the only one with a consequence attached. The word remains in every press release and in no serious contract.

If this is the future, stop writing “open source” into RFPs now and specify licences by name and version. You will be early rather than late, at no cost.

UPSIDE · A CERTIFIER GETS TEETH — 20%

Certification becomes a market gate. The organic and Energy Star path.

- An audit or covert-testing event exposes a prominent “open source” claim as false in a way that embarrasses a large buyer ⚠
- OSI or a successor body is adopted as a qualifying criterion by a major procurement authority — the European Commission, a US federal schedule, or an enterprise consortium
- Enforcement attaches to procurement eligibility rather than to statute, which is faster

Model cards become audited artifacts. Licence claims get checked before purchase rather than argued about afterwards. The tier distinctions in this paper stop being analysis and become a compliance schedule. Firms that marketed restricted licences as open source face a re-labelling exercise conducted in public.

If this is the future, audit your own licence position before somebody audits it for you, and be able to name the artifacts you actually received.

DOWNSIDE · CAPABILITY CAPTURE — 25%

The definitional fight moves from licensing to national security and the licensing question becomes irrelevant.

- “Sufficiently capable” is defined in US policy as a mandatory pre-release testing threshold
- The threshold is set low enough to capture frontier open releases
- Compliance cost falls disproportionately on open releases, because guardrails cannot be reimposed after weights ship

Openness stops being a licensing property and becomes a capability classification. Whether a model may be released openly is decided by a testing regime, not by a licence file, and the body that sets the threshold acquires more control over the open ecosystem than OSI ever sought. Frontier open releases become viable only for organisations with a compliance function.

If this is the future, plan against capability ceilings rather than licence terms, and assume frontier open weights become a regulated category rather than a free one. **This branch is also the kill-shot for this paper's thesis:** if the threshold is defined and enforced, the vacuum closes — just not through the route anyone was arguing about.

TAIL · THE CERTIFIER IS CAPTURED **10%**

Horizon open-ended; trigger could occur in any year through 2032. The credit-rating path, run to completion.

- A certifier does emerge and does acquire procurement authority
- It is funded, directly or through membership, by the laboratories whose models it certifies ⚠⚠
- A serious security or safety failure traces to a model that was certified open and audited as safe

This is the uncomfortable one, and it is uncomfortable because it is the branch where everybody behaves reasonably and the outcome is still bad. A certification regime is built in good faith. It is funded the only way such regimes are ever funded. The incentive erodes the standard slowly and invisibly, the way it did at the rating agencies, and nobody notices because the label keeps being applied and nothing goes wrong. Then something goes wrong, and the inquiry discovers that the assurance everyone relied on was purchased by the party being assured. The response is a statute written in the aftermath of a disaster instead of in advance of one, and the entire open ecosystem inherits a compliance regime designed by people who are angry.

If this is the future, keep your own independent record of what you deployed and why. A certification will not protect you, and in this branch it is specifically the thing that fails.

What is true across all branches

INVARIANT 1

In every branch, “open source” carries less information in 2030 than it did in 2024. The branches differ on what replaces it, not on whether it degrades.

INVARIANT 2

In every branch, the licence file — not the label — is the operative document. This is already true and no branch reverses it.

INVARIANT 3

In every branch, the party that declares a model open is the party that built it, unless an external body acquires enforcement power. Only two of the four branches supply such a body, and in one of those two it is captured.

Cross-branch decision: specify licences by name and version, keep your own record of the artifacts you actually received, and treat “open source” as a claim by a manufacturer rather than a fact about a model. That action is correct in all four futures, which is the strongest thing this paper can say.

What would move these numbers

The base case falls below 30% if any major procurement authority adopts a binding definition. The upside rises above 35% on a credible public audit exposing a false open-source claim at a named vendor. The downside rises above 40% the moment any US instrument attaches a numeric capability threshold to a pre-release testing requirement. The tail is not currently observable and will not be until a certifier exists; treat 10% as a deliberate under-weight held at the base rate rather than a measurement.

§6 · WHAT A FAIR READING CONCEDES

The strongest arguments against this paper's framing, stated without weakening.

The gap between MIT weights and full compliance is near zero in practice. A team self-hosting an Apache-licensed weights file can use, study, modify and share it. The missing training-data documentation matters for reproducibility, auditability and legal certainty. It does not usually matter on a Tuesday, and a paper that treats the distinction as urgent is describing a problem most deployers will never encounter.

The credit-rating analogy has a real disanalogy. Ratings were load-bearing for capital requirements by regulation; nothing today makes an "open source" claim load-bearing for anything comparable outside the EU exemption. The systemic-failure mechanism that produced Dodd-Frank requires a channel through which a false label transmits harm at scale. That channel does not currently exist in AI. If it never forms, the tail branch is materially overweighted.

Anthropic's safety argument is not obviously self-serving. It is a coherent position that predates the current commercial fight, appears in the same terms in earlier writing, and turns on an empirical claim about offence-defence balance in biology that is genuinely unresolved. Reading it purely as competitive positioning requires ignoring that the same argument would have been inconvenient for Anthropic to make at several earlier points when it made it anyway.

And the labellers might converge without being forced. The base-rate table contains four cases. Four is not many. Open-source software itself is arguably a fifth case that resolved differently — through licence proliferation, then consolidation around a small set of well-understood licences, driven by tooling and corporate legal departments rather than by statute. If AI follows the software path rather than the food-labelling path, this paper's central analogy is the wrong one.

§7 · WHAT COULD NOT BE DETERMINED

Gaps are stated, never filled. This section is longer than an author would like it to be, which is the point of having it.

> ▒ UNDETERMINED
>
> The Treasury Board Directive on Automated Decision-Making's exact wording on open-source licences, and whether any Canadian instrument defines the term by reference to any external standard.

> ▒ UNDETERMINED
>
> The position of the United States Trade Representative on distillation. Reported secondhand; no direct statement located.

> ▒ UNDETERMINED
>
> The licence under which Kimi K3's weights were released. WP-10 reported a 27 July release date; the evidence points to mid-July and the licence terms were never established. Both are corrected here to unknown.

> ▒ UNDETERMINED
>
> Whether OSAID has been revised since v1.0.

> ▒ UNDETERMINED
>
> Litigation volume and price-premium data behind the “natural” case in §4. The trajectory is corroborated; the magnitudes are not.

> ▒ UNDETERMINED
>
> The jurisdictional cascade. How the EU, China, Korea, Japan, India, the Gulf states and the United Kingdom would respond to a US restriction was not researched to a standard permitting publication, and is deliberately absent rather than speculated. It is the intended subject of WP-12.

Two notes on method. Every revenue and market-share figure circulating in podcast coverage of this dispute was excluded again this edition; the economic argument here rests on the inference-cost differential, which is separately sourced. And the share of tokens served by Chinese open-weight models is reported across sources as anywhere between the mid-forties and low-sixties percent depending on measurement window; the direction of travel is corroborated, the magnitude is contested, and this paper uses the direction only.

SOURCES

Primary sources were retrieved directly wherever the mark ▮ appears.
Where a claim rests on secondary reporting the outlet is named in the text.

1 **Anthropic.** “Detecting and preventing distillation attacks,” 23 February 2026. Full text retrieved and searched. anthropic.com/news/detecting-and-preventing-distillation-attacks

2 **Amodei, D.** “Our position on open-weights models,” Anthropic, 27 July 2026, edited 28 July 2026. anthropic.com/news/position-open-weights-models

3 **Open Weights and American AI Leadership,** open letter, 24 July 2026, hosted at images.nvidia.com

4 **Anthropic.** Response to OSTP Request for Information on the AI Action Plan, 6 March 2025; position on the AI Diffusion Rule, April 2025.

5 **Open Source Initiative.** Open Source AI Definition v1.0, 28 October 2024, All Things Open, Raleigh. opensource.org

6 **Software Freedom Conservancy;** Free Software Foundation. Published objections to OSAID v1.0, October 2024.

7 **Regulation (EU) 2024/1689 (AI Act).** Arts. 51, 53(2), 54(6); Recitals 102–104. European Commission GPAI Guidelines, July 2025.

8 **Meta.** Llama 4 Acceptable Use Policy and Llama 4 Community License Agreement. llama.com/llama4/use-policy

9 **Bartz v. Anthropic PBC.** N.D. Cal. Final approval 20 July 2026, Martínez-Olguín J. Predicate ruling June 2025, Alsup J.

10 **Thomson Reuters Enterprise Centre GmbH v. Ross Intelligence Inc.** 765 F. Supp. 3d 382 (D. Del. 2025), Bibas J.; §1292(b) certification 23 May 2025; Third Circuit No. 25-2153, argued 11 June 2026, undecided.

11 **US Government Accountability Office.** GAO-10-470, “Energy Star Program: Covert Testing Shows the Certification Process Is Vulnerable to Fraud and Abuse,” 5 March 2010.

12 **Organic Foods Production Act of 1990;** USDA National Organic Program final rule, effective 21 October 2002.

13 **Dodd-Frank Wall Street Reform and Consumer Protection Act,** Title IX Subtitle C; Credit Rating Agency Reform Act of 2006.

14 **Treasury Board of Canada Secretariat.** Directive on Automated Decision-Making, 2019, amended April 2023.

NPSI WP-11 · Rated AAA by the Issuer · 28 July 2026 · Unclassified · Open distribution.
Supersedes WP-10, which remains available unaltered with a correction notice attached.
North Pacific Strategy Initiative · Vol. I · Est. MMXXVI · Jesse James · jesse@npsi.ca · npsi.ca

Every claim carries a grade and every forecast names what would move it. If you can move one, write.