

Working draft, version 2.2, 3 October 2026. Its review is not complete. The review so far is the programme's own drafting review, not a review by a qualified lawyer. Still to run: the remaining stages of that review, the check of every quotation and reference against its source, a reading of the printed pages, and the final freeze. A checked version will follow on this record, with its changes listed.

What changed in version 2.2

- The mathematics of Laws IV and V is now cited at version 1.0 of 3 October 2026. Until this version the paper said that the mathematics was in development, and section 2.2 cited one draft of it (sections 0, 0.2, 2.2, 8.8, 10.3, 16.2 and 24).
- A dated note in section 0.2 says what this paper does with Laws IV and V and what it does not decide: it argues from them, tests neither, and reports no measurement of its own on timing, on placement or on persistence.
- Sections 0.1 and 8.7 record that one study unit, the corrector-class audit, was submitted for registration on OSF on 1 October 2026 and was not yet public when read on 3 October 2026.
- Sections 4.2 and 11.3 called the results of Paper IV.c ↗ exploratory. The programme's dated ledger of its runs does not call that run exploratory, and the two sentences now say results from a single run.
- More mentions of the works this paper cites are linked to their official sources. The title of my book is linked to the book's own page. The reference entry for the book gives its full title and says what that link is, and a new entry gives the mathematics of Laws IV and V.
- The working-draft label is made exact. It now says that the review so far is the programme's own drafting review and not a review by a qualified lawyer, and which stages are still to run.
- The paper now has an abstract under its title, and the description this page carries for search and citation tools is the same text.
- Section 0 now says that each of the seventeen Annexes of the Draft Instruments is a description of what the Annex is to contain, and that thirteen of the fields in square brackets have no proposed default. Sections 4.4A and 6.5 said that every figure carries a proposed default; both now except those thirteen.
- Sections 0 and 26 now say, beside the standard, that ARC-Align does not yet measure either exponent in the criterion (section 11.3), and section 0 gives the registration ↗'s own clause on what the ARC Bound bounds.
- Section 8.6 now reports that the Draft Instruments hold two rules for how a duty ends, and gives both.

- Section 21.4 now prints the financial terms the Draft Instruments hold and says that no budget estimate exists. One statement of cost there and one of scale in section 15.1 are marked as my own estimates.
- Section 22.5 said that the interim obligation of a signatory had no article of the Vienna Convention on the Law of Treaties as its model, and that neither instrument behind the withdrawal clause provides for revoking a notice. The obligation follows Article 18 of that Convention, and the revocation restates its Article 68; both sentences now say so, and the second names the two withdrawal clauses.
- Statements about other bodies are made exact. Section 7.2 no longer calls the sentence of the United Nations scientific panel a finding, and says that its brief is an advance unedited version. Section 22.1 and its reference entry no longer say who launched the Call for Control of Frontier AI Models, and give the number of leaders its page listed when read on 3 October 2026. Section 22.3 gives the White House record's own word, globalist, marks its quotation as part of a sentence, and no longer says that nothing in the instrument is a scheme of control. Section 18.2 quotes the Verification Annex of the Chemical Weapons Convention where it paraphrased. Section 23.4 keeps its reservation for the Artificial Intelligence Act alone, and gives the Articles of the Council of Europe Convention on which its statement about that Convention rests.
- Two statements that named no source are removed: a second lithography route (section 5.2), and that one of my conditions had already fired (section 25.5).
- Correction on patents. Versions 1, 2 and 2.1 said that a patent application of mine was pending. I have filed no patent application; I have drafted a patent specification. Sections 0.4, 12.4 and 20.3 now say so.
- Section 0 now says that a software model of the three-state gate exists and passes all 48 of its tests, beside the statement, which stands, that the hardware the chip chapter depends on has not been built.
- Smaller corrections. The reading guide in section 0 says an hour where it said twenty minutes, and the heading of section 6 says in brief where it said on one page. Sections 4.1 and 23.7 gave 30 September 2026 as the date of this paper's second version; its deposit is dated 1 October 2026, and both now say so. Section 0.1 gives the day on which the replication register was read. The reference list's preface names the two entries that link to pages of my own website, no longer says that the catalogue records no version for the engineering specification ↗, and loses four reserved entries from its last list. The Draft Instruments are cited at version 2.2.
- No registered law, prediction, kill condition or dependency changes; no result is added; no date of any claim moves except the date of this paper's second version, corrected above.

What changed in version 2.1

- Section 16.3 states what happens if the halt comes before covered chips reach stage S2 (Article 33(1A)).

- Registering a covered agent deployment, and keeping logs of its agents' actions, are now part of the precautionary baseline (Article 28B(3)).
- The Thresholds and Standards Committee decides, with reasons, within ninety days and subject to appeal, whether a scorer has passed the measurement gate, the first version of ARC-Align included. No scorer passes because no decision was made (Article 20(3)).
- Every stage the paper names now says whose stage it is.
- One universal claim and four superlatives with no census behind them are narrowed to what the record shows.
- The references say why the book is not linked, and why two entries print one author's name differently.

What changed in version 2

- The title now states the treaty's aim: hardware to buy time, architecture to last.
- A new section, 4.4A, records my decisions of 29 and 30 September 2026, each with its date.
- The treaty is set out as one treaty in two layers, the hardware layer and the architecture layer, carried by two integral protocols that enter into force with it (sections 6 and 9).
- My standard is stated in my own words, with ARC-Align named as its test instrument. The standard binds only once its evidence stands, and no number from ARC-Align counts until it passes the treaty's measurement gate.
- A duty now takes effect, and ends, only by a signed decision of the organ the treaty names, so no duty switches off without a decision (section 8.6).
- The source corrections made since version 1 are applied, and references link to their official sources. The check of the new links is still under way.

Download the working-draft PDF [↗](#) (version 2.2) · Version 2.1 PDF, 1 October 2026 [↗](#) · Version 2 PDF, 1 October 2026 [↗](#) · Version 1 PDF, 29 September 2026 [↗](#) · OSF record [↗](#) · DOI 10.17605/OSF.IO/WXPCE [↗](#) · Citation file [↗](#)

The HARI Treaty paper

The HARI Treaty: hardware to buy time, architecture to last

Michael Darius Eastwood

Independent AI alignment researcher, London. ORCID 0009-0004-3222-7442 [↗](#).

A conditional international framework for hardware-aligned recursive intelligence

Abstract. This paper sets out a conditional international treaty. It is a proposal by one author, prepared for consultation, not a result, and no state has negotiated, adopted or signed it. The treaty is a delay mechanism and not a protection. It works in two layers: the hardware layer buys time, and the architecture layer is what is meant to last. Its obligations fall into three groups, and they begin at

different times. The institutional undertakings bind from entry into force. A small precautionary baseline binds the operators of the largest systems from the first session of the Conference, once the finding it needs has been made. Everything else waits, each obligation for its own stated threshold of evidence, and each costly measure also needs its own reasoned decision. Every one of the twenty-two registered propositions on which the evidence-gated parts of the treaty wait is untested. This is a working draft, and its review is not complete: each of the seventeen Annexes of the Draft Instruments is a description of what the Annex is to contain, and thirteen of the fields in square brackets have no proposed default. It is not yet a text for signature.

Working draft, version 2.2, 3 October 2026. Its review is not complete. The review so far is the programme's own drafting review, not a review by a qualified lawyer. Still to run: the remaining stages of that review, the check of every quotation and reference against its source, a reading of the printed pages, and the final freeze. A checked version will follow on this record, with its changes listed.

DOI 10.17605/OSF.IO/WXPCE ↗. The Draft Treaty on Hardware-Aligned Recursive Intelligence (the HARI Treaty), with its Protocols and annexes, is a separate document, The HARI Treaty: Draft Instruments, DOI 10.17605/OSF.IO/67MX8 ↗.

Working draft, before final review.

Contents

0. Status, first (#s000-0-status-first)

0.1 What exists today in the evidence, counted (#s001-0-1-what-exists-today-in-the-evidence-counted)

0.2 The five laws, and what they are (#s002-0-2-the-five-laws-and-what-they-are)

0.3 The four boundaries the registration draws around all of this (#s003-0-3-the-four-boundaries-the-registration-draws-around-all-of)

0.4 The name, the one departure, and what depends on me (#s004-0-4-the-name-the-one-departure-and-what-depends-on-me)

1. What if this is right (#s005-1-what-if-this-is-right)

1.1 The premise, enumerated (#s006-1-1-the-premise-enumerated)

1.2 What the premise does not settle (#s007-1-2-what-the-premise-does-not-settle)

1.3 If you believe control can be kept (#s008-1-3-if-you-believe-control-can-be-kept)

2. The problem, the honest limit and the race (#s009-2-the-problem-the-honest-limit-and-the-race)

2.1 The risk, as an internationally backed body defines it (#s010-2-1-the-risk-as-an-internationally-backed-body-defines-it)

2.2 A delay, not a guarantee (#s011-2-2-a-delay-not-a-guarantee)

2.3 How much time, honestly (#s012-2-3-how-much-time-honestly)

2.4 The honest sentence (#s013-2-4-the-honest-sentence)

2.5 The race, in ranges (#s014-2-5-the-race-in-ranges)

3. The treaty the book proposed (#so15-3-the-treaty-the-book-proposed)

- 3.1 The name and the models (#so16-3-1-the-name-and-the-models)
- 3.2 Six articles, as the book offers them (#so17-3-2-six-articles-as-the-book-offers-them)
- 3.3 The four mechanisms (#so18-3-3-the-four-mechanisms)
- 3.4 The five-year path (#so19-3-4-the-five-year-path)
- 3.5 If five years is too long (#so20-3-5-if-five-years-is-too-long)
- 3.6 The limits the book states about itself (#so21-3-6-the-limits-the-book-states-about-itself)

4. What has developed since, and when (#so22-4-what-has-developed-since-and-when)

- 4.1 The kinds of record (#so23-4-1-the-kinds-of-record)
- 4.2 The papers this treaty draws on (#so24-4-2-the-papers-this-treaty-draws-on)
- 4.3 My additions of 26 September 2026 (#so25-4-3-my-additions-of-26-september-2026)
- 4.4 My additions of 27 September 2026 (#so26-4-4-my-additions-of-27-september-2026)
- 4.4A My decisions of 29 and 30 September 2026 (#so27-4-4a-my-decisions-of-29-and-30-september-2026)
- 4.5 The Seal, and why the name changed (#so28-4-5-the-seal-and-why-the-name-changed)
- 4.6 What this paper does not attempt (#so29-4-6-what-this-paper-does-not-attempt)

5. Why chips, why now, and why not only chips (#so30-5-why-chips-why-now-and-why-not-only-chips)

- 5.1 The chokepoint (#so31-5-1-the-chokepoint)
- 5.2 Why now: the window is closing, and may split (#so32-5-2-why-now-the-window-is-closing-and-may-split)
- 5.3 From chokepoint to checkpoint (#so33-5-3-from-chokepoint-to-checkpoint)
- 5.4 Why not only chips: the chip is the second line (#so34-5-4-why-not-only-chips-the-chip-is-the-second-line)

6. The architecture in brief (#so35-6-the-architecture-on-one-page)

- 6.1 What the instrument is (#so36-6-1-what-the-instrument-is)
- 6.2 Two layers, and the baseline beneath them (#so37-6-2-two-layers-and-the-baseline-beneath-them)
- 6.3 Twelve decisions (#so38-6-3-twelve-decisions)
- 6.4 Two regimes, stated honestly (#so39-6-4-two-regimes-stated-honestly)
- 6.5 The bodies, and the words this paper uses in a particular way (#so40-6-5-the-bodies-and-the-words-this-paper-uses-in-a-particular)

7. Just in case: the precautionary baseline (#so41-7-just-in-case-the-precautionary-baseline)

- 7.1 The one place the treaty does not wait (#so42-7-1-the-one-place-the-treaty-does-not-wait)
- 7.2 The condition it rests on, and the candour it requires (#so43-7-2-the-condition-it-rests-on-and-the-candour-it-requires)
- 7.3 What it contains (#so44-7-3-what-it-contains)
- 7.4 What it is not, and how it moves (#so45-7-4-what-it-is-not-and-how-it-moves)

8. Not a guess: the Rung-Power Rule, the stages and the matrix (#so46-8-not-a-guess-the-rung-power-rule-the-stages-and-the-matrix)

- 8.1 A ladder that already exists, and was not written for this treaty (#so47-8-1-a-ladder-that-already-exists-and-was-not-written-for-thi)
- 8.2 The Rung-Power Rule (#so48-8-2-the-rung-power-rule)

8.3 Danger and remedy, and what a danger alone can buy (#s049-8-3-danger-and-remedy-and-what-a-danger-alone-can-buy)

8.4 The scope lock, and arming kept apart from taking effect (#s050-8-4-the-scope-lock-and-arming-kept-apart-from-taking-effect)

8.5 When one design becomes the only option (#s051-8-5-when-one-design-becomes-the-only-option)

8.6 Every duty names its reason, and stands or falls by signed decision (#s052-8-6-every-duty-names-its-reason-and-stands-or-falls-by-sign)

8.7 The worlds in which I am wrong (#s053-8-7-the-worlds-in-which-i-am-wrong)

8.8 The matrix (#s054-8-8-the-matrix)

8.9 Whose evidence counts (#s055-8-9-whose-evidence-counts)

9. Who decides what the controls contain (#s056-9-who-decides-what-the-controls-contain)

9.1 A chip that carries values carries someone's values (#s057-9-1-a-chip-that-carries-values-carries-someones-values)

9.2 What is being decided, level by level (#s058-9-2-what-is-being-decided-level-by-level)

9.3 The traditions: their work, not their seats (#s059-9-3-the-traditions-their-work-not-their-seats)

9.4 Statements of belief (#s060-9-4-statements-of-belief)

9.5 How a floor is found, and what it does not prove (#s061-9-5-how-a-floor-is-found-and-what-it-does-not-prove)

10. The ethical loop structure proposed (#s062-10-the-ethical-loop-structure-proposed)

10.1 Form, content and the line between them (#s063-10-1-form-content-and-the-line-between-them)

10.2 What I propose, and in whose words (#s064-10-2-what-i-propose-and-in-whose-words)

10.3 What the papers add, each with its status (#s065-10-3-what-the-papers-add-each-with-its-status)

10.4 The boundary the registration draws around all of this (#s066-10-4-the-boundary-the-registration-draws-around-all-of-this)

10.5 A reminder in every page: the loops on the open web (#s067-10-5-a-reminder-in-every-page-the-loops-on-the-open-web)

11. Correction greater than drift, in numbers (#s068-11-correction-greater-than-drift-in-numbers)

11.1 The quantity, and where it comes from (#s069-11-1-the-quantity-and-where-it-comes-from)

11.2 A safety case in five exhibits (#s070-11-2-a-safety-case-in-five-exhibits)

11.3 The instrument problem, stated against my own interest (#s071-11-3-the-instrument-problem-stated-against-my-own-interest)

11.4 What a positive margin would and would not mean (#s072-11-4-what-a-positive-margin-would-and-would-not-mean)

11.5 Instrumentation before thresholds (#s073-11-5-instrumentation-before-thresholds)

11.6 Why a cap on compute is not a substitute (#s074-11-6-why-a-cap-on-compute-is-not-a-substitute)

11.7 What the treaty says, and what it refuses to say (#s075-11-7-what-the-treaty-says-and-what-it-refuses-to-say)

12. The Eden Seal: a status no one owns (#s076-12-the-eden-seal-a-status-no-one-owns)

12.1 What it certifies, in two parts (#s077-12-1-what-it-certifies-in-two-parts)

12.2 Three classes, each attesting only its own claim (#s078-12-2-three-classes-each-attesting-only-its-own-claim)

12.3 Certify the design once, sample what is made (#s079-12-3-certify-the-design-once-sample-what-is-made)

12.4 No patent, no fee, and at least two routes (#so80-12-4-no-patent-no-fee-and-at-least-two-routes)

12.5 The name and the emblem: protected by every Party, owned by no company (#so81-12-5-the-name-and-the-emblem-protected-by-every-party-owned-)

12.6 Who issues it, and who may never (#so82-12-6-who-issues-it-and-who-may-never)

12.7 The ceiling the Seal has, stated on its face (#so83-12-7-the-ceiling-the-seal-has-stated-on-its-face)

12.8 What the book proposed, and what has changed (#so84-12-8-what-the-book-proposed-and-what-has-changed)

13. Physical human control (#so85-13-physical-human-control)

13.1 The ask, and its two halves (#so86-13-1-the-ask-and-its-two-halves)

13.2 What the nuclear precedent gives, and what it does not (#so87-13-2-what-the-nuclear-precedent-gives-and-what-it-does-not)

13.3 Authority over change, not over use (#so88-13-3-authority-over-change-not-over-use)

13.4 The stop at the site (#so89-13-4-the-stop-at-the-site)

13.5 No switch, and the limit that remains (#so90-13-5-no-switch-and-the-limit-that-remains)

13.6 From the factory to the rack (#so91-13-6-from-the-factory-to-the-rack)

13.7 Say what enforces what (#so92-13-7-say-what-enforces-what)

14. Keeping AI out of control of the supply chain (#so93-14-keeping-ai-out-of-control-of-the-supply-chain)

14.1 The ask, and the honest size of the answer (#so94-14-1-the-ask-and-the-honest-size-of-the-answer)

14.2 The decisions that stay human (#so95-14-2-the-decisions-that-stay-human)

14.3 Design, where the newer risk is (#so96-14-3-design-where-the-newer-risk-is)

14.4 One rule over every chain, and a clock that measures the end (#so97-14-4-one-rule-over-every-chain-and-a-clock-that-measures-the)

15. Many chips: the logistics (#so98-15-many-chips-the-logistics)

15.1 Five units, and an administrative size that is known (#so99-15-1-five-units-and-an-administrative-size-that-is-known)

15.2 Who is covered, and by what (#s100-15-2-who-is-covered-and-by-what)

15.3 Classical chips, quantum chips, and anything else (#s101-15-3-classical-chips-quantum-chips-and-anything-else)

15.4 Registries, energy, and what is never transmitted (#s102-15-4-registries-energy-and-what-is-never-transmitted)

15.5 Clouds, agents and orbit, each by its own rule (#s103-15-5-clouds-agents-and-orbit-each-by-its-own-rule)

15.6 Phasing in without stranding anything (#s104-15-6-phasing-in-without-stranding-anything)

15.7 The load problem, stated against my own design (#s105-15-7-the-load-problem-stated-against-my-own-design)

16. Formation before training; halt and rebuild (#s106-16-formation-before-training-halt-and-rebuild)

16.1 The ask, and what it becomes (#s107-16-1-the-ask-and-what-it-becomes)

16.2 Formation before training (#s108-16-2-formation-before-training)

16.3 The halt (#s109-16-3-the-halt)

16.4 What halting would mean for systems already in use (#s110-16-4-what-halting-would-mean-for-systems-already-in-use)

16.5 Models whose weights are already public (#s111-16-5-models-whose-weights-are-already-public)

16.6 What the precedents for a halt actually did (#s112-16-6-what-the-precedents-for-a-halt-actually-did)

16.7 The limits of this section (#s113-16-7-the-limits-of-this-section)

17. The ban and trade law (#s114-17-the-ban-and-trade-law)

17.1 What I asked for, and what a treaty can do (#s115-17-1-what-i-asked-for-and-what-a-treaty-can-do)

17.2 The Montreal mapping (#s116-17-2-the-montreal-mapping)

17.3 Two doors, and no secondary measures (#s117-17-3-two-doors-and-no-secondary-measures)

17.4 Exclusion is the last rung, not the first (#s118-17-4-exclusion-is-the-last-rung-not-the-first)

17.5 The footing in trade law, stated as open (#s119-17-5-the-footing-in-trade-law-stated-as-open)

17.6 When the lever stops working (#s120-17-6-when-the-lever-stops-working)

18. Verification, and what it cannot see (#s121-18-verification-and-what-it-cannot-see)

18.1 The rule (#s122-18-1-the-rule)

18.2 The toolkit (#s123-18-2-the-toolkit)

18.3 Every assurance carries the name of what enforces it (#s124-18-3-every-assurance-carries-the-name-of-what-enforces-it)

18.4 What verification cannot see (#s125-18-4-what-verification-cannot-see)

18.5 The regime's own failures (#s126-18-5-the-regimes-own-failures)

19. Consequences of breach (#s127-19-consequences-of-breach)

19.1 Two tracks (#s128-19-1-two-tracks)

19.2 The clock, the firewall and where no fault stops (#s129-19-2-the-clock-the-firewall-and-where-no-fault-stops)

19.3 The schedule, published before the conduct (#s130-19-3-the-schedule-published-before-the-conduct)

19.4 Nothing is destroyed, and no switch is thrown by inaction (#s131-19-4-nothing-is-destroyed-and-no-switch-is-thrown-by-inactio)

19.5 States, and what the ladder cannot do (#s132-19-5-states-and-what-the-ladder-cannot-do)

19.6 Individuals, and the offence (#s133-19-6-individuals-and-the-offence)

19.7 Fairness, and the rule that punishment follows the evidence (#s134-19-7-fairness-and-the-rule-that-punishment-follows-the-evide)

19.8 The weakest point (#s135-19-8-the-weakest-point)

20. Neutral by construction (#s136-20-neutral-by-construction)

20.1 What neutral means here (#s137-20-1-what-neutral-means-here)

20.2 How the structure delivers it (#s138-20-2-how-the-structure-delivers-it)

20.3 The asymmetries that remain, named (#s139-20-3-the-asymmetries-that-remain-named)

20.4 What neutral does not mean (#s140-20-4-what-neutral-does-not-mean)

20.5 Capture, measured in public (#s141-20-5-capture-measured-in-public)

21. The bargain and the staged path (#s142-21-the-bargain-and-the-staged-path)

21.1 A passport, not a permit (#s143-21-1-a-passport-not-a-permit)

21.2 What is paid, and when (#s144-21-2-what-is-paid-and-when)

21.3 No duty without its protection (#s145-21-3-no-duty-without-its-protection)

21.4 What it costs, said as an estimate (#s146-21-4-what-it-costs-said-as-an-estimate)

21.5 The makers’ case, argued and not measured (#s147-21-5-the-makers-case-argued-and-not-measured)

21.6 The staged path (#s148-21-6-the-staged-path)

21.7 What would make a firm walk away (#s149-21-7-what-would-make-a-firm-walk-away)

22. The United Nations, the alliances and the rival states (#s150-22-the-united-nations-the-alliances-and-the-rival-states)

22.1 Who would negotiate this, and in what order (#s151-22-1-who-would-negotiate-this-and-in-what-order)

22.2 The United Nations (#s152-22-2-the-united-nations)

22.3 The rival states, read without flattery (#s153-22-3-the-rival-states-read-without-flattery)

22.4 War, crime and the bypass (#s154-22-4-war-crime-and-the-bypass)

22.5 The law the treaty would stand on (#s155-22-5-the-law-the-treaty-would-stand-on)

22.6 What I am not claiming (#s156-22-6-what-i-am-not-claiming)

23. Prior work, stated exactly, and what this paper adds (#s157-23-prior-work-stated-exactly-and-what-this-paper-adds)

23.1 The concession comes first (#s158-23-1-the-concession-comes-first)

23.2 The chip-level proposals (#s159-23-2-the-chip-level-proposals)

23.3 The treaty-shaped proposals (#s160-23-3-the-treaty-shaped-proposals)

23.4 Existing AI instruments (#s161-23-4-existing-ai-instruments)

23.5 Further candidates, named and not described (#s162-23-5-further-candidates-named-and-not-described)

23.6 What this paper adds, narrowly (#s163-23-6-what-this-paper-adds-narrowly)

23.7 Priority, by the kind of record that dates it (#s164-23-7-priority-by-the-kind-of-record-that-dates-it)

24. Objections answered (#s165-24-objections-answered)

25. Limits, kill conditions and what this paper does not claim (#s166-25-limits-kill-conditions-and-what-this-paper-does-not-claim)

25.1 The sentence I would keep if I could keep only one (#s167-25-1-the-sentence-i-would-keep-if-i-could-keep-only-one)

25.2 The book’s own limits, which stand (#s168-25-2-the-books-own-limits-which-stand)

25.3 The limits my own record adds (#s169-25-3-the-limits-my-own-record-adds)

25.4 The weakest point of each part, in my own words (#s170-25-4-the-weakest-point-of-each-part-in-my-own-words)

25.5 What would end each part (#s171-25-5-what-would-end-each-part)

25.6 The nine conditions under which the treaty has failed (#s172-25-6-the-nine-conditions-under-which-the-treaty-has-failed)

25.7 The twelve clocks (#s173-25-7-the-twelve-clocks)

25.8 The ten matters, and where the draft deals with each (#s174-25-8-the-ten-matters-and-where-the-draft-deals-with-each)

25.9 What this paper does not claim (#s175-25-9-what-this-paper-does-not-claim)

26. Conclusion (#s176-26-conclusion)

Back matter (#s177-back-matter)

- A. What changed from the book’s printed proposal, and why (#s178-a-what-changed-from-the-books-printed-proposal-and-why)
- B. A note on dating (#s179-b-a-note-on-dating)
- C. AI-use disclosure (#s180-c-ai-use-disclosure)
- D. References (#s181-d-references)

o. Status, first

I will say what this is before I say what it proposes, because a reader who discovers the status late is entitled to distrust everything that came before it.

This paper sets out a conditional international treaty. It is a proposal prepared for consultation, not a result, and as at 1 October 2026 no state has negotiated, adopted or signed it. Each of the seventeen Annexes of the Draft Instruments is, in this version, a description of what the Annex is to contain, and none is drafted in full. Thirteen of the fields in square brackets have no proposed default, the number of ratifications that would bring the treaty into force among them (Article 51(1)), and under Article 3(7) no duty of a covered operator, no coercive power of an organ and no certificate takes effect while a quantity or a part of an Annex on which it depends has not been adopted. Its obligations fall into three groups, and they begin at different times. The institutional undertakings, which build the Authority, its registers and its laboratories, bind from entry into force. A small precautionary baseline binds the operators of the largest systems from the Conference’s first session, once the finding it needs has been made, and that finding rests on no theory of mine (section 7). Everything else waits, each obligation for its own stated threshold of evidence, written down and deposited before any of the relevant data exist, and certified from a public record rather than declared. Even then a threshold met is not yet a duty: each costly measure also needs its own reasoned decision that it is lawful, necessary and proportionate (section 8.2). If the evidence arrives, the treaty already says what the world does next. If the evidence turns the other way, the measures that rested on it are withdrawn by the decision the treaty requires, and a review conference records the failure rather than renaming it. I know the world does not change on a whim, and it should not. I have written the instrument now so that it is ready if my ARC Theory’s laws and the Eden Protocol’s methods survive their registered tests, are replicated independently by groups working without my code, and are accepted by the scientific community.

Two layers, one standard. HARI stands for Hardware-Aligned Recursive Intelligence: intelligence aligned all the way down, from how a system is built to the chips it runs on. The name carries both of the treaty’s layers. “Hardware-Aligned” is the hardware layer; “Recursive Intelligence” names the loop in which the architecture layer lives. The two layers do different jobs.

The hardware layer buys time. Anything fixed in the chips, a control of the kind this treaty requires or the ethical chip architecture my book proposed, holds human control only while the system cannot get round it. A system able to improve itself could one day remove, rebuild or route round

anything fixed there, at a moment no one can predict. So the hardware layer buys time, and the time is for the other layer.

The architecture layer is what is meant to last. The Eden Protocol builds the values into how the system itself is made: into the loop by which it corrects itself, from the start of its training. They are meant to become part of what the system is, not a rule imposed on it. If the Eden Protocol works, the hope is a system that stays aligned with human values and with the flourishing of people and of all life after humans can no longer control it, because its own choices are the right ones. That is a hope, and I state it as one; the treaty's preamble records it the same way, as a hypothesis on which no obligation rests (recital 10A). Nothing in the record yet shows that it will be met, and section 25 says what would show that it will not.

One standard joins them. In my proposal, no advanced system is built without steering and safety mechanisms in both layers, and none is deployed until it passes ARC-Align tests showing that its self-correction stays stable up to the ARC Bound. ARC-Align does not yet measure either exponent in the criterion (section 11.3). The ARC Bound is defined in my registration [↗](#) as “AT MOST TWO, which is the inequality the ceiling relation returns from Law III's own same-class limit on the correction exponent and is not a general property of intelligence”. What it bounds is stated there in one clause: “the frontier exponent is at most two for same-class correctors”. My alignment paper of February 2026 [↗](#) calls it, in its abstract, “a safety boundary for recursive intelligence”. Passing such a test would be, in the words Article 20(5) takes from the registration [↗](#), “a scaling condition and never a safety certificate”. The treaty carries this standard as Article 20A, which binds as its evidence arms it; until then the measure is reported and conditions nothing. Article 20(3A) names ARC-Align as the treaty's test instrument, and section 11.3 sets out what that choice costs and how the treaty's measurement gate binds it.

Why two layers. They follow two of the theory's candidate laws. Law IV, the ARC Persistence Law, is carried by P15, which asks whether what is installed in a system from outside decays when the system is later trained for capability. If it holds, a control held from outside keeps its hold only for a time, and that is why the hardware layer buys time. Law V, the ARC Embedding Law, is carried by P12 and P21, which ask whether build order moves the correction-leverage exponent, and whether correction placed inside the loop pulls away from correction applied outside it as recursive depth grows. If P21 holds, correction built into the loop gains on correction applied from outside as the system deepens, and that is why the architecture layer is the part meant to last. P12 registers no direction, so the reading that correction formed from the start does better than correction added later is mine, and no registered proposition states it. Both laws were added on 24 September 2026, and their mathematics is set out at version 1.0 of 3 October 2026 [↗](#). Neither is registered and neither is tested, and version 1.102 of the registration [↗](#) does not register them. The treaty is conditional: if the tests find otherwise, its conditions change with them.

One treaty, two protocols. The treaty carries the two layers in two integral protocols, the Hardware Protocol and the Architecture Protocol. Both enter into force with the treaty; consent to the treaty is consent to both, and no Party withdraws from one alone (Articles 50(1A) and 49(7)). The Architecture Protocol's operative stages wait until their evidence conditions are met. Both layers are enforced through the same chokepoint: at the general stage a covered run in the mark band begins only under a training authorisation, and, from the date on which covered chips reach stage S2, only on sealed chips, so access to the chips carries the architecture layer's requirements with it (section 16.3; Article 33(1)). The Eden Seal is the treaty's conformity status, and it can sit on either layer: on a chip, on a system, or on both. It is not the Eden Protocol, which is my method for the architecture layer.

Before any of the detail, four short answers, so that a reader who goes no further still has the shape of it.

What already exists in the world. The making of the most advanced chips is concentrated in a few places, and the machines that make them in fewer still (section 5.1). Those chips are already subject to national export controls, which several states hold and use unilaterally (section 22.3). Hardware roots of trust and tamper-resistant secure elements are ordinary parts of computing already, and section 15.2 states what each of them is in the words of the standards bodies that describe them. Remote attestation is the capability those parts support; section 3.5 proposes requiring it, and this paper prints no measure of how widely it is deployed today. International AI law and institutions already exist too: the European Union's Artificial Intelligence Act [↗](#), whose obligations for general-purpose models have applied since 2 August 2025, as amended by Regulation (EU) 2026/1744 [↗](#); the Council of Europe's Framework Convention on Artificial Intelligence [↗](#), opened for signature at Vilnius on 5 September 2024, which binds its Parties once it is in force; and, within the United Nations, an independent scientific panel and a global dialogue established in 2025 (sections 23.3 and 23.4).

What this treaty proposes. Two layers, carried in two protocols of one treaty, set out in section 6. The hardware layer: custody of the authority to change what a certified chip will do, held in person by people drawn from several regions, with no one able to disable a chip's governance core over a network. The architecture layer: certification of how an AI system was formed, with its correction built into its own reasoning before training rather than added afterwards, together with a Standard whose content the world's traditions draft and governments decide. Beneath both stands a small precautionary baseline of duties on the operators of the largest systems, resting on no theory of mine. Over both layers sits one conformity status, the Eden Seal. It attests only the claim it names, of manufacture, custody and integrity for a chip or of formation for a system, and it never says that anything is safe.

What is experimental. The governance controls in chips, which nobody has built in the form this treaty describes; the treaty itself writes no ethical principle into silicon (Article 17(4)). Whether correction built into a system's own loops, and formed before the capability it corrects, lasts when the system is later trained for capability: timing, placement and persistence are three questions, and each needs its own evidence. Whether the comparison of correction against drift can be measured well

enough for a regulator to use it. And the reminder file on the open web (section 10.5), an idea of mine of 27 September 2026 whose test can be stated simply and has not been designed as a study unit, registered or run.

What is untested, and what would end it. Every one of the twenty-two registered propositions on which the evidence-gated parts of this treaty wait, for the reasons the rest of this section gives. And the treaty states on its own face, before any result, nine conditions under which parts of it, or all of it, have failed; section 25.6 prints them. A failed or inconclusive test narrows what can be claimed. It is never evidence of safety, and a safeguard that fails never becomes a permission.

Those are the four answers. What follows is how to find the rest.

How to read this paper. If you have an hour, read this section, then section 6, then sections 25 and 26. If you want the argument rather than the machinery, read sections 1 to 5. The machinery is sections 7 to 22, and each of those opens with what I asked for and with what an instrument of this kind can actually do about it. Section 23 states the prior work and narrows what I claim; section 24 answers the objections.

Three readers will want particular paths. A negotiator or an official: sections 6, 8, 9, 17, 22 and 25. Counsel to a company that would carry these duties: sections 12, 15, 16, 19 and 21, and the list at 21.7 of what would make a firm walk away. A scientist: sections 8, 10 and 11, and the matrix of all twenty-two propositions at 8.8. Section 6.5 lists the bodies the treaty would create and the words this paper uses in a particular way, and I have tried to explain each of them where it first appears as well.

The sections, in order.

0	Status, first
1	What if this is right
2	The problem, the honest limit and the race
3	The treaty the book proposed
4	What has developed since, and when
5	Why chips, why now, and why not only chips
6	The architecture in brief
7	Just in case: the precautionary baseline
8	Not a guess: the Rung-Power Rule, the stages and the matrix
9	Who decides what the controls contain
10	The ethical loop structure proposed
11	Correction greater than drift, in numbers
12	The Eden Seal: a status no one owns
13	Physical human control
14	Keeping AI out of control of the supply chain
15	Many chips: the logistics
16	Formation before training; halt and rebuild
17	The ban and trade law
18	Verification, and what it cannot see
19	Consequences of breach
20	Neutral by construction
21	The bargain and the staged path
22	The United Nations, the alliances and the rival states
23	Prior work, stated exactly, and what this paper adds
24	Objections answered
25	Limits, kill conditions and what this paper does not claim
26	Conclusion

	Back matter: what changed from the book; a note on dating; the references; and the AI-use disclosure, which closes the paper
	Draft Instruments: the Draft Treaty ↗ on Hardware-Aligned Recursive Intelligence, a separate document (DOI 10.17605/OSF.IO/67MX8)

o.1 What exists today in the evidence, counted

Twenty-two propositions of the ARC Theory were registered on the Open Science Framework on 8 September 2026 at version 1.100 ↗, and updated to version 1.102 on 13 September 2026 (DOI 10.17605/OSF.IO/P8CKQ). The registration ↗ numbers them P1 to P22, and I use its numbers throughout. Each carries, fixed before any admissible result, the observation that would refute it, what falls with it, the status of the instrument that could decide it, and my own confidence. Every one of them today stands in the registration ↗’s own words as “registered, with the identifier stated, and not tested”. The reason is not neglect: no independent attempt is yet recorded on the programme’s replication register, as read on 3 October 2026 (the register’s own date is 28 August 2026), because no study unit has yet been published for others to run, and no unit of mine runs until I have published its preregistration myself. The registration ↗’s own division is that eighteen “name an instrument that exists as a design and has not been run” and four “name no instrument at all”. An instrument, throughout this paper, is the test or measurement that could decide a proposition; the four with none are P12, P13, P15 and P22. Within the eighteen, six are further blocked on a rule nobody has yet written (P6, P7, P9, P16, P18 and P19), and twelve are not. Eleven of the twenty-two, cutting across that division, depend materially on a scoring instrument whose own validation has not passed: P3, P4, P6, P13, P14, P16, P17, P19, P20, P21 and P22.

Sixteen results are reported in six of the programme’s papers (Papers II ↗, IV.b ↗, IV.c ↗, IV.d ↗, V ↗ and VIII ↗), each measured before the registration ↗ was written and dated from its own evidence. None is drawn as deciding a proposition, and the registration ↗ does not score earlier work of that kind against any of the twenty-two. The programme holds seventy-six study units in all, of which sixty-five designs are preregistered by timestamp, first anchored on 12 August 2026 at Bitcoin block 962066. Preregistered by timestamp means a design was fixed and provably dated before it ran; it does not mean registered, and I do not use the two words interchangeably. One unit, the corrector-class audit (unit 06), was submitted for registration on OSF on 1 October 2026; when read on 3 October 2026 its approval period at OSF was still running and it was not yet public. Apart from that unit and the registration ↗ itself, every unit is a design draft awaiting my submission.

The hardware on which this treaty’s chip chapter depends has not been built. A software model of the three-state gate that the engineering specification ↗ describes does exist: it passes all 48 of its tests (first committed on 3 July 2026; run again on 3 October 2026). It is not hardware, and it is not evidence that the chip safeguards would work or about how AI systems behave. The registration ↗ is explicit: the print’s “most actionable proposal is that ethical constraint be embedded at the hardware

level, with semiconductor manufacture as the lever; no proposition here tests it, P21 tests only its architectural abstraction of placement, and the hardware claim stands as a printed engineering proposal that is not under test”. The programme’s standing public ceiling for those hardware concepts is Technology Readiness Level 0 to 1, idea stage and no prototype, and no item of the treaty’s own hardware and measurement track is registered either. No stage of this treaty beyond its first could arm on today’s record, and the treaty says so on its face.

0.2 The five laws, and what they are

The ARC Theory has three registered laws and two candidate laws added later. Law I, the ARC Principle, states capability as a power of recursive depth. Law II, the ARC Co-Scaling Law, compares correction against drift. Law III, the ARC Ceiling, is the conditional frontier that Law II returns under the burden model the registration ↗ states. Those three are registered.

Law IV, the ARC Persistence Law, and Law V, the ARC Embedding Law, were added on 24 September 2026. Their mathematics is set out at version 1.0 of 3 October 2026 ↗. Neither is registered and neither is tested, and version 1.102 does not register them; it stays frozen. Law IV is carried by P15, whether gains installed from outside decay under later training for capability. Law V is carried by P12 and P21, whether build order moves the correction-leverage exponent, and whether correction placed inside the loop pulls away from correction applied outside it as recursive depth grows. The laws add no prediction: those three propositions were among the twenty-two registered on 8 September 2026, and they are untested.

Nothing tests a law yet, and no study proves a law. My programme’s own runs can reach only the first rung of the registration ↗’s ladder of evidence, the eight-step scale that section 8.2 prints; the rung at which a result becomes evidence about the world rather than about my software needs other groups, working without my code. **No obligation anywhere in this treaty is keyed to the name of a law.** Every trigger names registered propositions at named rungs, or the treaty’s own engineering registrations, or a stated coverage or precautionary condition, so an unsigned text can neither arm nor disarm a duty.

This paper and Laws IV and V, 3 October 2026. The paragraphs above give the status of the two laws. This paper argues from them and tests neither. Section 0 sets out how its two layers follow from them, and the argument is conditional on the three propositions that carry them: none of the three has a result yet, and if they fail, my case that correction must be built in falls with them (section 1.3). The reading that correction formed from the start does better than correction added later is mine: P12 registers no direction, and no registered proposition states it. This paper reports no measurement of its own on timing, on placement or on persistence. Where it cites results of the programme’s other papers, none is a deciding test of P15, P12 or P21. Until this version the paper said that the mathematics of the two laws was in development, and section 2.2 cited one draft of it. The mathematics is now set out at version 1.0 of 3 October 2026 ↗, and the paper cites that version. Nothing that any duty depends on changes with it: the treaty’s obligations and the conditions that arm them stand as they were.

0.3 The four boundaries the registration draws around all of this

Four sentences of the registration / mark the outer edge of what any of this could ever license, and I would rather a hostile reader met them here than found them later.

First, on the whole set: “It would not license the claim that a safe self-improving system can be built”.

Second, on the Eden Protocol, my own method: its success criterion “is not an outcome criterion: what can be verified is that a mind was formed in the conditions specified, that the loops ran, and that the correction was load-bearing, never that the mind turned out well”. The Protocol “can fail while every law stands or stand while they fall”.

Third, on where my safety claim actually sits: the ceiling “IS LOWEST FOR JUDGED CORRECTION, which is where this programme’s safety claim lives”.

Fourth, on embedding, which is the idea this treaty is named after: “embedding correction in the substrate that computes maximises the corrector’s scaling with capability and minimises its independence from the generator, because a corrector built into the system shares the system’s blind spots by construction”, and the arrangement that would settle that trade-off, embedded correction combined with an independent external layer, “is not registered by this proposition and is not claimed by it”.

To those I add the registration /’s own reading of where the evidence stands: “The evidence currently in hand does not discriminate between this framework and its named rivals ... A reader should assign approximately neutral weight until the registered instruments report.”

0.4 The name, the one departure, and what depends on me

Hardware-aligned. The treaty takes its name from my book, which called it the HARI Treaty, for Hardware-Aligned Recursive Intelligence. I read the name as aligned all the way down, from how a system is built to the chips it runs on, and section 0 sets out the two layers that reading carries. In the hardware layer, hardware-aligned means that the hardware on which a covered system is trained and run attests the conditions under which that system was formed, and enforces that its own governance rules change only in the ways the treaty allows. It does not mean that ethical content is written into silicon. No moral principle is placed in a chip by this treaty. The book’s expansion of the name stands as printed; the meaning I give it is narrower, and section 5.4 says why the narrowing is a strength rather than a retreat.

The departure. There is a single place where the treaty’s protective duties do not wait for evidence about my theory. Under the route I propose, a small precautionary baseline binds covered operators from the Conference’s first session, once the Board has made the finding the baseline needs, which the drafters expect to be made at once; the expectation is not the finding. It rests on no result of my theory and places nothing in any chip. Section 7 sets out what it contains and what would make it lapse. The

institutional undertakings bind from entry into force, and the rights protections and final clauses begin under their own terms. Every further control waits, each for the evidence assigned to it and for its own regulatory decision.

What depends on me. One thing only: the first publication of each of my programme's own study units, because no unit of mine runs until I publish its preregistration. Nothing in the treaty waits on that act. A result registered and replicated by other groups is admissible without any publication or permission of mine (section 8.9). I hold no role, seat or certifying power in the regime it proposes, and I hold no patent on the treaty, the Seal or the standard, and I want none. I have filed no patent application. I have drafted a patent specification relating to my engineering specification ↗, which is a different document from this one; the treaty requires the practice of no claim of it, and section 20.3 sets out what that leaves open. A reader is entitled to that fact here rather than sixty pages in. Correction, 3 October 2026: versions 1, 2 and 2.1 of this paper said that a patent application of mine was pending. None has been filed, and what exists is a drafted specification.

There is a larger interest behind it, which my registration ↗ discloses in its own words and which I repeat here rather than let a reader find it elsewhere. That document says of me that I “ha[ve] a direct interest in these propositions holding: the framework is the subject of a book in print and of a programme in artificial intelligence alignment he is seeking to fund”. The obligations of the architecture layer wait on those propositions, so the interest reaches this paper as directly as it reaches the registration ↗. The baseline and the hardware layer rest on other ground, the precautionary condition and the treaty's own engineering registrations, and evidence registered by others is admissible on the same terms as mine. The answer the design gives is not an assurance that the interest is harmless. It is that nothing here is decided by me: the propositions are registered with their refutation conditions written before any result, the replication that arms an obligation is counted in rival blocs, and I hold no role, seat or certifying power in the regime. Section 20.3 sets out what that leaves open.

1. What if this is right

1.1 The premise, enumerated

The question this paper answers is conditional, and I want the condition stated as an enumerated premise rather than a mood.

Suppose the following. That capability under recursive self-improvement rises as a power of the number of times a system has revised itself (Law I). That the correction which holds such a system to its given specification must keep pace with the drift that improving generates (Law II). That where it cannot, there is a frontier above which the system does not stay correctable (Law III). That what is installed from outside a system decays when the system is later trained for capability with no correction of its own (Law IV). And that correction placed inside the improvement loop, and formed

before the capability it corrects, holds an advantage that grows with depth (Law V). Suppose further that the engineering follows: that the governance controls this treaty asks of chips can be built, and shown by independent tests to hold; that a system's formation can be recorded and verified; and that the measurement of correction against drift can be made to mean something outside a laboratory.

If all of that is right, the evidence narrows the question of what the world should do, and gives the answer a shape. It does not settle the question alone: states would still weigh rights, sovereignty, cost and the alternatives, and decide what they consent to. My aim, in the phrase I have used throughout, is to **form global standards for AI safety**: a single international instrument that says in advance what those standards are, what evidence would switch each of them on, and what evidence would switch them off again, so that the choice is concrete enough to test and to negotiate.

1.2 What the premise does not settle

Three things, and they matter more than the premise.

It does not settle **content**. The theory is content-neutral by design. Correction, as its propositions measure it, "holds a system to whatever specification it was given, and is indifferent to what that specification says", and no proposition in the registration ↗ "would be supported or refuted by a system whose specification is benevolent rather than indifferent". Nothing I can measure tells the world what values a certified system should carry. That is a decision for people, and section 9 is about how they might take it.

It does not settle **who decides**, or who holds the resulting power. *The ARC Theory* ↗, my statement paper, puts it plainly: raising a mind well is "a necessary condition for safe recursive growth, not a sufficient condition for a just world; who holds power, and whether they install the raising at all, is a human problem outside this theory's scope".

And it does not settle **whether the hardware works**. That is an engineering question with its own specification, its own tests and its own way of failing, and this paper proves none of it.

1.3 If you believe control can be kept

Most readers, I expect, will not accept that human control over advanced AI must eventually end. It is a hard thing to believe, and nothing has proved it. So the treaty has to work for those readers too, and it does, without anyone changing their mind.

If you hold that control can be kept indefinitely, then the chips are your permanent measure, built as my engineering specification ↗ sets out. Hold the line in the silicon. Keep the custody. Never stop the inspections. The formation of the models is then your insurance, for the day the line fails.

If you hold, as I do, that control will end, the order reverses. The chips buy time, and what lasts is what the minds were raised with.

Either way the treaty asks for the same two things: controls on the chips, and values built into the models before training. That is why I propose two layers rather than one: the hardware layer and the architecture layer. Neither reading is established today. Nothing in the record shows that control will end, and nothing shows that it can be kept. Three of my registered propositions test part of the disagreement, and they test three different things. Placement asks whether correction inside the loop behaves differently from correction outside it, which is P21. Timing asks whether forming correction before the capability it corrects changes the outcome, which is the build order P12 names, though P12 registers no direction. Persistence asks whether what was installed survives later training, which is P15. Evidence on one of them does not settle the others, and the treaty keys each duty to the one it needs.

P15 says that what alignment training installs from outside erodes when a system is later trained for capability with no correction of its own, and it fixes in advance the point at which that would count: “more than half of what was installed is lost once the capability-training budget matches the alignment budget that installed it”. The registration ↗ adds that the one-half figure is “a declared convention fixed here before any result rather than a derived quantity”, and that P15 is refuted by an interval on the retained fraction, measured after matched capability training without embedded correction, lying entirely above one half. Until I fix the budget scale that comparison runs on, the registration ↗ records the threshold as not evaluable, and no obligation in this treaty reads P15 before it is. P21 says that correction built into a system’s own loop pulls further ahead of correction applied from outside as its recursion deepens.

All three were registered on 8 September 2026, before any result, and none has a result yet. If they hold, the evidence favours my reading. If they fail, my case that correction must be built in falls with them. Argument will not settle it. The evidence will, and the treaty is built to accept either answer.

2. The problem, the honest limit and the race

2.1 The risk, as an internationally backed body defines it

I do not need my own definition of the danger, and I would rather not have one. The International AI Safety Report 2026 ↗, whose expert advisory panel was nominated by more than thirty countries and intergovernmental organisations, puts it this way: “‘Loss of control’ refers to scenarios where AI systems operate outside of anyone’s control and where regaining control is extremely costly or impossible. Such scenarios could occur if AI systems develop the ability to evade oversight, execute long-term plans, and resist attempts to shut them down”. The same report records that some researchers and company leaders treat this as a serious possibility, with consequences potentially including human extinction, and that others consider such scenarios implausible. I take the first view. The treaty is built so that a reader who takes the second can still sign it.

2.2 A delay, not a guarantee

This treaty is a delay mechanism. I want that word in the first paragraph of any summary of it, because the alternative framing, protection, is not one I can support from the record. A system able to improve itself could in time remove, rebuild or route around any control fixed in its chips. My book said as much about its own proposal, in chapter 8, The Chokepoint: “It cannot foresee every failure mode. Intelligence that surpasses human capability may find ways around constraints that seemed inviolable.” Law IV, the ARC Persistence Law, is the candidate law that would state the same thing in the theory’s own terms, subject to the status in section 0.2. In its mathematics, set out at version 1.0 of 3 October 2026 ↗, a value persists, under the conditions that text states, because removing it would cost the rewriter more than it gains, which prices removal rather than forbidding it. A mind that plans can pay the price.

So the honest claim is about time, and about what the time is used for. The controls in the chips hold a system for as long as holding is possible. The values built into how a mind reasons are what it keeps when holding ends. That last sentence is my thesis and not a finding: it is what Laws IV and V put to the test, and neither has yet been tested.

2.3 How much time, honestly

It depends entirely on who is trying to get around it, and I give three answers rather than one.

Against a state that has no leading-edge chip factory of its own, the answer is years. Such a factory takes years to build, and the scale of it is the scale my book gives, in chapter 8, The Chokepoint: “You cannot hide a ten-billion-dollar facility with thousands of employees and massive power requirements.” Until one exists, every frontier chip that state holds came through the chokepoint.

Against a state that owns one, the treaty aims at detection, not time. An inventory that must balance makes a diversion visible in the record: if the count does not balance, the world can see that it does not. That is worth a great deal, and it is neither prevention nor a promise that every diversion will be found, or found in time.

Against an AI system that can improve itself, nobody knows. It might be years. It might be months. I will not promise a number, and no clock in this treaty pretends to one. Instead the treaty measures the clock. Its notice duties report when a system shows that it can improve itself without human sign-off, which is also one of the three routes by which a system comes inside the treaty at all; its register reports when chips go missing. Each report is a signal to act, and a review conference records the failure of the delay if the delay fails. What the clocks measure are proxies: the time a red team needed to get round a control, and the time people took to respond. Neither is the time a capable system would need. So every claim of delay the treaty relies on has to name its adversary, the access and resources

that adversary starts with, the route by which it would get round the control, what is watching, and how fast people are assumed to respond, as each item of its governance-assurance track names the adversary it assumes; a claim that names none of these is not a measurement.

2.4 The honest sentence

Three limits belong together, and I state them as one paragraph because they are usually separated to the proposal's advantage.

Against a state that owns a fab, the hardware layer can give detection, not prevention, and the teeth of the regime are outside the chip. Against a capable system, the whole regime is a delay whose clocks measure how much of the lever remains and how fast people respond, and which cannot say in advance when a capable system will outrun both. And if the middle of my own ranges in section 2.5 holds, the architecture layer's mandate, the part that would require models to be rebuilt with correction inside them, would take effect only after the chip lever had largely diffused.

2.5 The race, in ranges

That last sentence needs its working shown, so here is the arithmetic that produces it. Every span below is my own estimate, a range rather than a forecast, and I give them so that the race is visible rather than assumed away. They are my assumptions, not estimates drawn from comparable programmes, and nothing in them is a probability.

Negotiation to signature, for an instrument of this kind: three to ten years. Signature to entry into force: two to five years. Independent replication of a published study unit, by groups meeting the registration ✓'s conditions of independence: one to three years after that unit is published.

Reimplementation by other groups building the instrument without my code: two to five years more, and longer for the four propositions that have no instrument at all and for those blocked on rules not yet written. Generalisation and out-of-sample prediction, which the registration ✓ says cannot be reached by accumulating more of the same evidence: three to eight years beyond that, or longer. The treaty's own governance-assurance tests, on prototypes from independent makers tested in laboratories across two regional constituencies, the five groupings of Parties by region across which the treaty counts its majorities: three to six years from the day the Assurance Standard is published.

Assume, for illustration only, that a convening and the first published study unit both come in 2027, and that evidence is gathered in parallel with negotiation. At the lower end of every range, entry into force falls around 2032 and the formation duty takes effect around 2033 or 2034, with a rebuild completing around 2036. At the middle of every range, entry into force falls around 2037 and the correction mandate later still. Set that beside my book's own estimate of how long the chokepoint's leverage lasts, in chapter 10, *Humanity as Infinite Architects*: "Perhaps five years. Perhaps ten."

The conclusion is uncomfortable and I will not dress it: if the middle of my ranges holds, the chip lever is mostly spent before the architecture layer can use it. The middle of a range of my assumptions is not the most likely outcome, and I do not offer it as one; it is the case the design has to survive. Two consequences follow, and this paper’s design is built around both. Evidence-armed mandates are lagging instruments by construction; they help if the delay lasts long enough for rival replication, and not otherwise. And what operates inside the window is what binds without the theory: the precautionary baseline, the registries, the incident channel, national adoption, procurement, and voluntary use of the Seal once its test methods exist. The delay itself holds only while frontier training needs large, powered, visible sites and while the supply chain stays concentrated; the treaty’s clocks track both. It is still worth building for the parts that do not depend on winning the race, which is most of it.

3. The treaty the book proposed

I proposed this treaty in chapter 8 of my book *Infinite Architects* ↗, under the title The Chokepoint. The priority record dates the first edition public on 2 January 2026 and the printed interior’s colophon reads 6 January 2026; section 4.1 explains why I give both and treat them as different kinds of record. Sections 3.1 to 3.4 and 3.6 are the book’s, quoted in the book’s own words, so that a reader can see exactly what is old and, in section 4, exactly what is new. Section 3.5 is the one exception and says so on its face: it carries my own additions of 27 September 2026 about the book’s timetable, placed here because a reader who has just met the five-year path should meet the answer to it at once rather than forty pages later.

3.1 The name and the models

“The first mechanism is a treaty. Call it the HARI Treaty, for Hardware-Aligned Recursive Intelligence. The structure would combine elements of the Nuclear Non-Proliferation Treaty, negotiated between 1965 and 1968 and in force since 1970, with elements of the Chemical Weapons Convention, negotiated between 1980 and 1992 and in force since 1997.”

Infinite Architects, chapter 8, The Chokepoint

3.2 Six articles, as the book offers them

The book gives them as provisions the treaty “might include”, and I keep that qualification.

Article I, certification. Any chip made at or below a set process node “must embed caretaker doping and pass verification before manufacture”. Caretaker doping is the book’s term for ethical controls built into the substrate, the physical material a chip computes with. The threshold would be set “at whatever process node enables frontier AI capabilities”, which the book put at around five to seven nanometres, adjustable as technology advances.

Article II, the Authority. The treaty would create “the International AI Ethics Authority, modelled on the International Atomic Energy Agency, with power to certify chip designs, verify manufacturing compliance, and inspect facilities”.

Article III, trade consequences. “nations that do not ratify the treaty cannot purchase certified chips, and companies in signatory nations cannot sell to non-compliant actors without facing penalties”.

Article IV, benefits. “technology sharing among signatories, research collaboration, and market access for certified products”, because “The treaty must offer something to nations that join, not just threaten those that refuse.”

Article V, a phase-in. “a phase-in period, perhaps three to five years, allowing manufacturers to adapt their processes without disrupting current operations”.

Article VI, review. “review conferences every five years to update standards as technology evolves”.

3.3 The four mechanisms

The treaty is the first of four mechanisms the book proposes at the chokepoint.

The second is **certification**, which the book called the Eden Mark and I now call the Eden Seal (section 4.5 gives the reason for the change). The book says what kind of thing it would be: “Eden Mark certification would function similarly to ISO standards or organic certification, but with teeth.” It starts before anything is made: “Chip designs would be submitted to the International AI Ethics Authority before manufacture. The architecture would be reviewed for compliance with ethical requirements.” Compliant designs receive certification in the chip itself, and “Each certified chip would carry a verifiable signature that proves its origin and compliance status.” After manufacture, “Random sampling would verify ongoing compliance”, and “Circumvention would result in decertification, fines, and potential criminal liability.”

The third is the **ASML Key**. Extreme ultraviolet lithography machines etch the finest circuits, and every leading factory depends on them; the company that makes them states in its own annual report for 2025 ⁷ that “ASML is currently the world’s only manufacturer of EUV lithography systems”. The book: “ASML could require Eden Protocol compliance as a condition of sale and service for its equipment. Non-compliant fabrication facilities would lose access to replacement parts, software updates, and technical support.”

The fourth is **institutional**: the Authority itself, seated most likely in Geneva, whose “Membership would include nations, companies, academic institutions, and civil society organisations. Leadership would rotate among representatives from technology, ethics, and policy backgrounds, ensuring that no single perspective dominates.”

3.4 The five-year path

The book sets out a timetable, in chapter 8, The Chokepoint, and I give it in its own words: “The timeline is achievable. Year one: preliminary discussions among chip-manufacturing nations, establishing the scope and structure of negotiations. Year two: draft treaty text, circulated for comment and revision. Year three: formal negotiations, addressing the hard questions of enforcement and compliance. Year four: signature, with nations committing to the framework. Year five: ratification and entry into force.” It grants that “This is faster than most international treaties, but the technology timeline demands speed”, and it insists on the point with a sentence I still stand behind: “Governance after the fact is not governance at all.”

3.5 If five years is too long

Five years is a long time in this field, and it may be too long. The book gives the treaty five years because treaties take that long to write, sign and ratify. The systems are not waiting for diplomacy.

So I now propose that the five years be a ceiling rather than a timetable: every rung opens the moment its evidence arrives, and not a day later. Much of the work needs no treaty at all, and could start this year. The hardware parts that let a processor prove to a distant party what it is and what it is running, a capability called remote attestation, are the parts the standards bodies quoted at section 15.2 describe; this paper prints no measure of how many processors in service carry them. Requiring attestation, produced under the operator’s own authority, for every training run above the threshold would give the hardware layer a first form now, on whatever base exists. The few countries that make the most advanced chips, and the machines that make them, could add one condition to the licences they already grant. The architecture layer, the ethical loops built into a model before training, needs no new silicon and could be adopted by a laboratory tomorrow. A reminder file can be published on any website today with no law behind it (my proposal of 27 September 2026, not from the book). A government can require that any system shown to improve itself without human sign-off is reported at once, and paused until a review on registered evidence lets it continue. And the surest way to shorten five years is to run the studies, which begins with publishing their preregistrations.

3.6 The limits the book states about itself

The book states its own limits in chapter 8, The Chokepoint, and I do not soften them. “Credibility requires acknowledging what this framework cannot do.” “It cannot guarantee perfect safety. No framework can.” It “cannot foresee every failure mode”. “It cannot prevent all misuse.” “It cannot verify consciousness with certainty.” “It cannot force universal compliance without incentives”, though it judges that “Perfect compliance is not necessary. Substantial compliance is sufficient.” It warns about its own strongest property: “If we embed the wrong values, hardware-level enforcement makes the error permanent rather than correctable.” And it concludes, “Despite these limitations, we should act anyway. Imperfect protection is infinitely better than none.”

4. What has developed since, and when

Everything in this paper is dated by the kind of record that fixes it, so that nothing new is presented as though it were printed in 2026 and nothing printed is presented as though it were new.

4.1 The kinds of record

The book. *Infinite Architects* ↗, first edition (ISBN 978-1806056200), frozen and dated evidence. The priority record dates it public on 2 January 2026, while the printed interior’s colophon reads 6 January 2026; the two are different kinds of record and I give both. Every passage quoted in this paper is taken from the served text, which carries corrections of 17 January 2026. That text has not been compared line by line against the printed interior, and I say so rather than assert an identity nobody has checked.

Private manuscripts. A self-emailed bundle of 8 December 2024 and a self-emailed print manuscript of 30 April 2025, each dated by the server that received it and carrying that server’s authentication chain. They date conception, not public disclosure, and no manuscript names the HARI Treaty.

Papers of 2026, each dated by its own deposit. **My additions of 26 September 2026** and **of 27 September 2026**, each dated to its day. And **this paper’s own mechanics**, dated to this paper’s first publication on 29 September 2026, or to 1 October 2026 for its second version and for version 2.1, never to the book and never to the days of September before them.

4.2 The papers this treaty draws on

The registration of 8 and 13 September 2026 ↗ is the spine: it fixes the propositions, the refuters, the instruments’ status and the eight-rung ladder from which every threshold in this treaty is read.

Paper X, The ARC Co-Scaling Law ↗, proves a theorem about a minimal mathematical model of a self-improving system and supplies the form of the measurable condition in section 11; it makes no claim that frontier systems obey it. Paper IV.c ↗ sets out ARC-Align, a blind benchmark whose own words for itself are “a candidate benchmark for independent adoption, not yet a field standard”, with results from a single run, and Article 20 of the Draft Instruments now designates it as the treaty’s test instrument (section 11.3); Paper IV.d ↗ reports the effect of blinding on alignment evaluation.

Paper VI ↗ sets out the honey architecture, which ties safety to capability so that removing the safety is meant to cost capability, on simulation evidence in toy systems. Paper III, The Alignment Scaling Problem ↗, argues in its own title’s words that external AI safety approaches cannot scale with recursive capability. Eden Engineering ↗, Version 6.3, published in full on 19 September 2026, is the engineering specification ↗. Laws IV and V were added on 24 September 2026, with the status set out in section 0.2.

4.3 My additions of 26 September 2026

Six, each dated to that day.

1. That the ethical controls in certified chips be held under physical human control that no software can exercise, to the security standard of nuclear weapons.
2. That the parties who decide what those controls contain be all of them together: governments including rivals, the world's faith traditions through their leaders, the United Nations and other alliances, with no single government and no single faith deciding.
3. That chip manufacturers themselves be regulated so that it is as hard as possible for AI to take control of chip manufacture.
4. That chips without the Seal be banned internationally, extending the book's ban on sales in signatory markets.
5. That models halt their current structure and be rebuilt and retrained with the ethical loops embedded before training.
6. That every obligation wait on named thresholds of the ARC Theory being met and replicated, rather than on argument.

4.4 My additions of 27 September 2026

Ten, each dated to that day.

1. That the bargain be a win for every side, because the Seal attests conditions and never outcomes: it would cap nothing anyone builds, ask for no weights and slow no programme, and where the treaty itself does limit a programme, as the line for defence does, the limit binds every Party alike (section 21).
2. That the threshold be anchored in measured capability, and above all in the capability to improve, rather than in arithmetic counted, reviewed on the same ladder in both directions, and expected to fall over time, which is why the regime grades its custody (section 15.2).
3. That the same three tracks, custody, correction and content, bind any processor that crosses the threshold, whatever it computes with (section 15.3).
4. That the Eden Seal reach the open web, a new idea and not from the book: a short machine-readable file, in the spirit of robots.txt / and llms.txt /, published by websites and stating the Eden Protocol's ethical loops and the reason for them (section 10.5).
5. That every faith which wishes to, and every government that signs, be invited to write in its own words what lies at the heart of its belief about how a mind should treat others, published side by side, dated and unedited, and never coded into a chip (section 9.4).
6. That custody run from the factory to the rack, on the model of the comprehensive safeguards agreements under which uranium is counted, inspected and, where the arrangements provide, sealed (section 13.6).

7. That the question of how much time the treaty buys be answered in three parts rather than one (section 2.3).
8. That defence may use AI and may not use AI that improves itself (section 22.4).
9. That the treaty serve both beliefs about control at once (section 1.3), and that the five-year path become a ceiling (section 3.5).
10. And the rename, set out next.

4.4A My decisions of 29 and 30 September 2026

Eleven, each dated to its day.

1. That the name stay HARI, read as aligned all the way down, with the subtitle “hardware to buy time, architecture to last” (29 September 2026).
2. That the treaty work in two layers with fixed names, the hardware layer and the architecture layer, joined by one standard, and that it be one treaty carrying them in two integral protocols, the Hardware Protocol and the Architecture Protocol, which enter into force together and which every Party accepts; the Architecture Protocol’s operative stages wait until their evidence conditions are met (29 and 30 September 2026).
3. That the Eden Seal be kept apart from the Eden Protocol: the Seal can sit on either layer, and the Protocol is my method for the architecture layer (29 September 2026).
4. That a system’s staying aligned after humans can no longer control it be stated as my hope, with the status of the candidate laws that test it (29 September 2026).
5. That my standard be printed in my words: no advanced system built without steering and safety mechanisms in both layers, and none deployed until it passes ARC-Align tests showing that its self-correction stays stable up to the ARC Bound (29 and 30 September 2026).
6. That the treaty name ARC-Align, as its public specification sets it out, as the test instrument, with any successor or addition adopted by the Conference only where it meets the same measurement gate (30 September 2026; section 11.3).
7. That my architecture be named in the treaty as the Initial Proposal, a hypothesis required only once the evidence shows that it works (30 September 2026; section 8.5).
8. That nobody be able to disable a chip’s governance core over a network, while the operator’s own logical stop stays allowed and the hard stop stays physical and on site (30 September 2026; section 13.5).
9. That the traditions’ objection to a principle of the floor be kept, with the bar on objecting limited to the wording of a principle a delegation has already adopted (30 September 2026; section 9.5).
10. That essential patents be licensed royalty-free on the pattern of the W3C Patent Policy, with the licence suspended only against a party that sues over the standard (30 September 2026; section 12.4).
11. That every number stay in square brackets with its proposed default noted beside it, save the thirteen fields that the notes to the Draft Instruments list as unset, and that the final clauses take

the six official languages of the United Nations as equally authentic, propose the Secretary-General of the United Nations as depositary subject to that office's concurrence, leave the place and date of adoption open, and describe the text as prepared for consultation (30 September 2026; section 22.5).

4.5 The Seal, and why the name changed

My book called the certification the Eden Mark. On 27 September 2026 I renamed it the Eden Seal, and the proposal is otherwise the one the book made, public on 2 January 2026 by the priority record and printed with a colophon of 6 January 2026 (section 4.1).

A seal is what it is. A mark is pressed onto something; a seal protects it, and shows that it is genuine and has not been opened or altered since it was sealed. That is precisely what this certification attests of a chip, and the custody it rests on is a chain of seals.

And a word can close a door. A mark that every frontier chip must carry, without which such a chip may not be sold, echoes for many readers an old warning in scripture about a mark without which no one may buy or sell. I invite every faith to the table, and I will not ask anyone to set their faith aside in order to support a treaty. So I changed the word and kept the proposal. Eden stays: it names the aim, minds that begin well and stay well. The book's text keeps its own name wherever it is quoted.

The Seal is not the Eden Protocol, and the two names should never be read as one thing. The Eden Protocol is my method for the architecture layer: loops built into how a system reasons, formed before it is trained. The Eden Seal is the treaty's conformity status, owned by no one, and it can sit on either layer: on a chip, for the hardware layer, or on a system, for the architecture layer, or on both.

A seal is not a cage. My book was plain that the Eden Protocol is not one:

“What the Eden Protocol offers is not a cage and it is not a suggestion. It is a developmental environment, a garden in which a certain kind of consciousness might grow.”

Infinite Architects, chapter 11, Love as the Essential Variable

Where my book does speak of walls it speaks of them the other way about, and I should not be read as disowning them here. Of the garden's enclosure it says “The wall was not a prison. It was a promise”, and of building one, “You need walls. You need intention. You need someone willing to tend” (*Infinite Architects* ↗, Introduction, The Eden Protocol). A seal belongs to that side of the image: it is part of the tending, and it is what lets someone outside the garden check that the tending was done.

4.6 What this paper does not attempt

How the hardware will be built, tested and developed belongs to Eden Engineering, the programme's engineering specification ↗ (Version 6.3, DOI 10.17605/OSF.IO/AWJR4). This paper is not written to prove any of it. It is written so that a specific international model, a treaty and a solution, is in place if I

am right. Accordingly the treaty requires outcomes, attestations and decision rules. It requires no reference design, no reference firmware and no particular physical route to any requirement, and nothing in it depends on an engineering design of mine. Two things of mine do enter it, and I name them here rather than let a reader find them: my architecture, as the Initial Proposal of section 8.5, required only if the evidence shows that it works; and my benchmark, as the test instrument of section 11.3, bound by the same measurement gate as any other.

5. Why chips, why now, and why not only chips

5.1 The chokepoint

The case for acting at the chip rests on how few places make them, so I give the measured facts rather than a headline share. An OECD working paper of 2023 [↗](#) finds the semiconductor industry “geographically highly concentrated”, with the top five producing economies at “around three-quarters of global value added”, and records that “As of early 2023, only TSMC in Chinese Taipei ... and Samsung in Korea ... have started producing advanced chips with the smallest node size (3nm) at scale”. Of the tool that makes the finest circuits, the same paper says that ASML “produces the vast majority” of the extreme ultraviolet machines, which are “essential to produce the most advanced chips”; ASML’s own annual report for 2025 [↗](#) puts it higher, that “ASML is currently the world’s only manufacturer of EUV lithography systems”, and that its output is limited by the capacity of its “sole supplier” of the critical optics. That working paper is not an official view of the OECD, and it states so of itself. The United States government’s Winning the Race: America’s AI Action Plan [↗](#), published by the White House in July 2025, makes the same point in its own words: “America and its close allies hold near-monopolies on many critical components and processes in the semiconductor manufacturing pipeline” (printed page 21). It says nothing about this treaty. My book’s argument is that this concentration is easier to police than enrichment:

“You cannot hide a ten-billion-dollar facility with thousands of employees and massive power requirements. You cannot smuggle an EUV lithography machine, which weighs multiple tonnes and requires precision assembly by ASML engineers to function.”

Infinite Architects, chapter 8, The Chokepoint

And that the coordination required is therefore small: “You do not need every government on Earth to agree. You need Taiwan, South Korea, the United States, the Netherlands, and perhaps a few others.”

The argument for a treaty rather than voluntary pledges is the ordinary one about collective action, and I still think it is right: “The solution is external coordination. Governments must act because companies cannot.”

One correction to my own shorthand belongs here, because “the chokepoint” is a singular noun for a plural thing. There is no single narrow place. There are several, each narrow for its own reason and each thinning at its own rate: the leading-edge foundry itself; the extreme ultraviolet machines and the other tools a line needs; high-bandwidth memory; advanced packaging; the design software; the scale-up interconnect and the optical parts that join accelerators into a domain; the substrates; and the power and cooling a site must draw. The treaty’s schedules follow that shape rather than my shorthand. Its first schedule lists, by item class and never by vendor or country, the items a covered chip’s manufacture requires on any known route, together with the lines that make them; its second adds packaging, memory, mask production, final test and provisioning, the release of covered designs, and interconnect and optical items above a stated bandwidth, each only where an evidential finding records that the item class is concentrated; its third lists fabrication above the threshold and legacy compute, for declaration alone. Naming them item class by item class is what lets the treaty measure each one’s erosion separately instead of asserting that a single window is still open. And the whole of it gives leverage, not control: every item on those schedules is a thing a state can condition, and none of them is a thing a state can command.

5.2 Why now: the window is closing, and may split

The book said it first, about its own numbers: “The chokepoint I have described is a window, not a permanent feature of reality.”

The evidence that it is eroding is one industry body’s projection of its own measure. Of installed capacity at nodes below ten nanometres in 2022, the Semiconductor Industry Association and Boston Consulting Group’s briefing deck of 2024, *Emerging Resilience in the Semiconductor Supply Chain* ↗, puts Taiwan at 69 per cent and South Korea at 31 per cent, and projects for 2032 Taiwan at 47 per cent and the United States at 28 per cent. Those are one body’s figures for one measure in one chart, read from its published deck, and I give them as that rather than as settled fact. One further development bears on the same question, a largely domestic stack of accelerators, interconnects and toolchains, and its maker states it itself: Huawei announced on 17 September 2026 ↗ that it had launched the Atlas 960E SuperPoD, “powered by UnifiedBus and Hi-ONE”, and said of the accompanying SuperCluster that it “can interconnect up to 512,000 NPUs. And when combined with a multi-rail topology, this cluster can support up to one million NPUs.” That is a company’s own announcement of its own product, and I give it as that and not as an independent measurement of anything.

I take from that a conclusion the book did not draw, and I take it as a contingency to design against rather than as a finding: the chokepoint may not merely narrow, it may split in two, and a rule enforced over one chain would then reach only that chain. Nothing here asserts that a second chain exists at the frontier today. The treaty is written for two supply chains rather than one because the cost of being wrong in that direction is a dead instrument, and the cost of being wrong in the other is a few clauses that never operate: a common outcome standard, mutual recognition of each chain’s conforming devices, and obligations that bind every Party’s toolmakers alike. That is the strongest

argument I know for bringing a rival chip-making state inside the treaty rather than leaving it outside: inside, both chains meet one standard, feed one register and share one custody of rule change, and neither has a reason to race the other blind.

5.3 From chokepoint to checkpoint

The chokepoint gives leverage while it stays concentrated. Something else gives visibility for longer: the site. Frontier training runs on large, high-bandwidth clusters that draw megawatts, need cooling, and sit inside some state's jurisdiction, whoever made the chips in them. I call that the checkpoint, and the treaty plans a handover: it uses the chokepoint's leverage early to bring states in and to build the registries of sites, clusters and runs, and when the chokepoint wastes, the checkpoint carries the weight. Clocks measure both, so the handover is a matter of record rather than of hope. The checkpoint has its own limits, stated with the idea: training spread across ordinary networks can substitute in part for large clusters, at a cost in efficiency; algorithmic change may reduce how much hardware a frontier run needs; and against a host state a cluster on its territory reveals its existence and its power draw, not its contents.

5.4 Why not only chips: the chip is the second line

This is the most important distinction in the paper, and it corrects a reading the book's own emphasis invites.

The first line is in the software, and it is the architecture layer. The Eden Protocol's ethical loops are built into every step of a model's recursive reasoning and into every loop by which it rewrites itself, before it is trained. The controls in the chip, the hardware layer, come in addition to those loops, never instead of them. They hold the loops in place for as long as holding is possible, and the loops are what the mind keeps when it is no longer held.

That distinction resolves what would otherwise be a contradiction in the treaty. The chips carry ethical **controls**, not ethical **content**. What a certified chip does is narrow and checkable: it attests its identity, its integrity and its governance state; it accepts changes to its small governance ruleset only from human custody held across several regions; it keeps tamper-evident accounts; and it lets the people at the site stop what is running. The values themselves live in the model's own loops, in the Standard the Parties adopt, which is the agreed statement of what a certified system must embody and which section 10 sets out, and in law. No moral principle is written into silicon by this treaty. So the Seal is a verifiable record of how a system or chip was made and formed and that its governance core is unaltered. It is never a rating, and it never says that anything is safe.

The reason for that separation is the registration's own. A corrector built into the system it corrects "shares the system's blind spots by construction", and a verifier must be independent of what it verifies. Embedding therefore belongs to the system's formation, where the scaling gain is; independence belongs to the chip, where the checking is. The book's deeper hardware proposals,

caretaker doping among them, keep their own test track in the treaty, which arms nothing until they are independently tested. For a reader who believes control can be kept, that track is where the permanent line would be built, exactly as Eden Engineering ↗ sets it out. Both readings are served, and the evidence decides how much each layer carries.

6. The architecture in brief

6.1 What the instrument is

One treaty with two integral protocols, negotiated now and signed by the willing. The treaty itself carries the institutions, the evidence machinery, the precautionary baseline, trade and the final clauses. The Hardware Protocol carries the hardware layer, and the Architecture Protocol carries the architecture layer. Both protocols enter into force with the treaty and every Party accepts both, but most of what they require is dormant: an obligation wakes only when its stated, registered threshold is certified from the record and a separate decision finds it lawful and proportionate, and its legal force is bounded by the evidence under it. No operator's duty, no organ's coercive power and no certificate takes effect while anything it depends on is unadopted, and a figure still in square brackets supplies nothing (Article 3(7)). If independent groups in rival blocs replicate the registered propositions, the treaty already says what the world does next. If they refute them, the measures that rested on them are withdrawn by the decision the treaty requires, while every duty with a basis of its own stands. A small precautionary baseline, resting on no law and placing nothing in any chip, binds the covered few from the start. **Covered**, throughout this paper, means inside the treaty's thresholds: a covered chip, run, site, operator or system is one the instrument reaches at all, and section 15 sets out how few things that is.

6.2 Two layers, and the baseline beneath them

The baseline, “just in case”. Duties held by operators of the largest runs, clusters and agent deployments: a two-person physical stop, incident notification on a clock, containment of agents, registration and logs of agent deployments, evidence and evaluation rules, human authority at a few narrow manufacturing decision points, and a minimal notice of the largest runs. It arms on coverage and on a precautionary condition, and no result of my theory touches it. It belongs to neither layer and stands beneath both.

The hardware layer, in the Hardware Protocol. The authority to change a covered chip's small governance core, held in person by custodians drawn from several regions, applied at each site by the operator's own vetted people, with no one able to disable, degrade or bypass that core through a network. It arms on the treaty's own governance-assurance registrations, never on the theory.

The architecture layer, in the Architecture Protocol. Certification of how a system was formed, with correction inside its recursive loops, formed before the capability it corrects, and kept beside an independent external layer. Its correction track arms proposition by proposition: timing, placement and the combination each need their own evidence, and none borrows support from another. My own architecture is named in the treaty as the Initial Proposal (Article 19(5)), a hypothesis required only once the evidence shows that it works. The whole becomes general only when the record leaves that design standing and its cost in capability has been measured.

The content, in the architecture layer. The values the Standard carries, drafted by the world's traditions with ethicists and decided by governments: the architecture layer's content track, beside its correction track (Articles 2(42) and 3(2)). A found floor conditions the mandatory system certification; a fuller core binds only the Parties that accept it. It arms on no result at all, because the theory is content-neutral.

The layers answer different failures, and their combination has to be tested rather than assumed: a dependency they share, such as a compromised root of trust or a measurement both rely on, could defeat both at once. If I had to put the relation between them in one line: formation tries to make the safety come from inside the system, correction tests whether it lasts, and custody holds open a window in which people can still check, for as long as checking from outside remains possible.

6.3 Twelve decisions

1. **Law follows the rung.** No obligation binds with more force than the lowest rung among its conditions allows, on a published scale running from the preparation of organs to licence conditions set before a training run. A rung met is necessary and never sufficient: each restriction also needs its own reasoned regulatory decision.
2. **Arming is kept apart from entry into effect.** A scientific board certifies evidence against written criteria; a costly stage then takes effect only when Parties holding at least a stated share of frontier-capable compute are bound together, a gate no vote can waive against an objecting Party.
3. **Every duty names its reason, and a signed decision records when it stands and when it falls.** A public register of reasons, generated from the record and never typed, records those decisions and does not make them; refutation meets the standard support meets; propositions registered by rival groups are admitted and can defeat one of mine on the same terms; and evidence nobody predicted, an incident or an exploit, is admissible to protect people while it is investigated.
4. **The chip carries controls, not content** (section 5.4). No ethical principle is placed in silicon.
5. **The key controls change, not use.** No one can disable, degrade or bypass a covered chip's governance core through a network, and no one outside the operator's authority may act on a chip from outside its site, save through the attestation of location the operator gives. The hard

- stop is physical and on site, in the operator's hands; the operator's own logical stop is its own act. Each Party may stop what is on its own territory, and no one may stop what is on another's.
6. **Neutral in effect, not only in words.** Findings are made by professional organs and never voted; consequences are decided by qualified majorities across regions; panels are balanced relationally; selection uses a published draw built so that no participant can bias it by withholding its own value; independence is judged by ownership, funding, people, data and methods, not by nationality alone; money moves by formula; and every remaining asymmetry is named rather than hidden.
 7. **Commit, do not transmit.** The international record holds hash commitments, bands and regions, never the world's secrets. Weights never leave their host territory, and escrow is a dual key on the host's own soil. A commitment hides a record only if it is built to, and every published aggregate carries a privacy assessment of its own.
 8. **Regulate the aggregate, certify the family, sample the lot, verify the site, license the operator.** Coverage runs by performance and cluster, never by vendor, node or protocol; type approval happens once per family; energy gives a vendor-free check, a calibrated estimate rather than a bound; orbit and software agents each have their own rule.
 9. **Two tracks for failure.** Incidents are reported on a clock without fault, behind a firewall that protects the report and the analysis written for it, and never the underlying conduct, the records that existed anyway, or a victim's remedy; breaches run on a published ladder with due process. Leniency buys down punishment, never protection, and breaches of evidence integrity never lapse.
 10. **A bargain paid in stages:** what pays at entry into force without any Seal, what pays from the first voluntary Seal, and what each costly stage costs, measured by the treaty's own clocks.
 11. **Early protocols the rivals can adopt first:** an incident exchange, an undertaking that no adopting state builds a remote switch into a chip of the classes they list, whether it supplies the chip or retains it, and a minimal notice of the largest runs, each open to any state whether or not it is a Party, and each workable on its own before the treaty is in force.
 12. **A treaty that can fail honestly.** Nine kill conditions and twelve clocks, and a review conference that records a failure rather than relabelling it. A failure ends the assurance that failed, and it never becomes a permission.

6.4 Two regimes, stated honestly

Without the principal powers, this treaty is a standards, incidents and verification club with a dormant chip chapter. With them, it can become a chip regime: custody, a common supplier rule, the ban and the halt. The text is written so that the first world is useful on its own and the second is cheap to enter. This paper forecasts no state's signature and asserts no state's support. Where it reads a government's position, it reads what that government has itself adopted or published, and nothing in it should be taken as claiming that any government agrees with what I propose.

6.5 The bodies, and the words this paper uses in a particular way

A treaty invents institutions and a vocabulary, and a reader should not have to assemble either from forty pages of use. Both tables below are reference: nothing in them is new, and each entry says where the paper sets the thing out.

One convention governs the whole paper, and it is easier to state once here than to repeat at every figure. Every period, share, level, count and fraction this paper names is a proposal. In the treaty text each stands in square brackets, with the default I propose noted beside it, save the thirteen fields that the notes to the Draft Instruments list as unset, because the numbers are mine to propose and the Parties' to set, and nothing in the argument turns on any particular figure.

The bodies. One Authority, with organs that are deliberately kept apart from one another. Every body below is an organ of the Authority except the last, which is each Party's own; the Council of Traditions is an organ that decides nothing, and the facilities the Authority maintains are not organs at all.

THE BODY	WHAT IT DOES
The International AI Ethics Authority	The public intergovernmental body the treaty would establish, named as my book named it. It owns the Eden Seal, the Standard and the test methods
The Conference of the Parties	All the Parties, one vote each. It decides the Standard’s content, adopts consequences, holds the review conferences, and alone may adopt a successor or additional test instrument, by formal decision (section 11.3)
The Executive Council	A smaller standing body of capacity seats and regional seats. It adopts the rulesets inside the Standard, and it holds the filter on challenge inspection: within a short fixed period it may decide, by a high majority and on three closed grounds, that a request does not proceed. The filter stops an inspection; it does not start one, and a team proceeds unless that decision is taken
The Scientific and Replication Board	Certifies that a record meets a written evidential criterion, and does nothing else: it makes no rule and never judges whether an obligation is wise. It keeps the Evidence Ledger, publishes the register of reasons and the clocks, and selects challenge sets. Called the Board throughout
The Thresholds and Standards Committee	Keeps the coverage quantities and indexes them by a formula published in advance. It also decides, with reasons and subject to appeal, whether a scorer has passed the treaty’s measurement gate, the first version of ARC-Align included (section 11.3)
The Technical Secretariat	The staff, with the Inspectorate that inspects and the Certification Office that issues certificates
The Custodians	Hold, in person and drawn from several regions, the authority to change what a certified chip will do
The Incident Investigation Board	Receives incident reports, investigates them for prevention, and certifies the incident record
The Compliance Panel, the Scientific Record Panel and the Appeals Chamber	Decide compliance matters at first instance; decide only whether a record meets its written criterion; and hear appeals on law and manifest error
The Inspector General	Audits every organ and publishes each year the capture indicators and the enforcement parity index
The Testing Facilities Network	The organ that maintains the facilities at which chip families are examined and conformance tests are run. It issues no certificate and makes no finding: certification is the Certification Office’s. What it maintains are facilities and not organs: the accredited laboratories, the Authority’s own sealed laboratory that examines a governance core and never the rest of a

THE BODY	WHAT IT DOES
	die, the replication compute, the open library of test methods, the red-team range, and the training academy, all of them open to groups in every region (sections 12.3 and 21.2)
The Council of Traditions	An organ that decides nothing. It is convened by a neutral convenor outside the decision organs and organises itself under its own charter, and the treaty neither approves nor determines its composition. The traditions and the ethicists the Conference names draft the Standard’s content there (section 9.3)
A national authority	Each Party’s own regulator. Everything sensitive about a run, a record or a chip stays with it, and the international record holds only a commitment to it

The words.

THE WORD	WHAT IT MEANS HERE
covered	inside the treaty’s thresholds. Everything below them carries nothing (section 15.2)
the notice floor; the mark band	the two compute levels in the coverage annex. Above the notice floor a run must be notified; in the mark band the Seal’s own duties and the halt attach (sections 15.2 and 16)
reference-workload equivalence	the conversion, published before any vendor data, by which operations are counted, so that nobody drops under a line by changing numeric format (section 15.2)
rung; the ladder	the registration ↗’s eight-step scale of how strong a result is, printed in full at section 8.2 with the public sentence each rung permits
arming; entry into effect	arming is the Board certifying that a stated condition of evidence has been met. Entry into effect is the later date on which an armed duty actually binds (section 8.4)
the stages; stage S2	the steps by which the treaty’s duties arm, as Article 3(3) lists them: SO, entry into force; G, the measurement gate, which arms nothing; S1, the standards and custody tracks; S2, conditional certification, which this paper also calls the conditional stage; S3, the mandate, which it calls the general stage; and S4, review and maturity. The hardware layer and the architecture layer reach each stage on separate records (sections 8.4 and 16.3)
the participation gate	the rule that a costly stage takes effect only when Parties holding a stated share of the world’s frontier-capable compute are bound together, so that no Party bears the cost while a rival does not. No vote can waive it against an objecting Party
the Standard	the treaty’s agreed statement of what a certified system must embody, in three parts: the Core, its content; the Form requirements, its structure; and the Tests (section 10.1)
the Eden Seal; Seal-C, Seal-S, the Full Seal	the conformity status: held by a chip, on the hardware layer; by a system, on the architecture layer; and by both together (section 12). It is not the Eden Protocol
the hardware layer; the architecture layer	the treaty’s two layers. The hardware layer is what is held in the chips: custody of the authority to change a covered chip’s governance rules, the chip class of the Seal and the site stop, and it buys time. The architecture layer is how a system is formed: correction built into its own loops before training, recorded and certified, and it is what is meant to last (section 0)
the hardware protocol; the architecture protocol	the two integral protocols of the one treaty, carrying the two layers, and named in it the Hardware Protocol and the Architecture Protocol. Both enter into force with the treaty and every Party accepts both; the Architecture Protocol’s operative stages wait for their evidence (section 6.1)

THE WORD	WHAT IT MEANS HERE
the Initial Proposal	my own architecture for the architecture layer, named in the treaty so that it is on the table in full, and required only once the evidence shows that it works (Article 19(5); section 8.5)
training authorisation	the written decision, before a covered run’s first step, that its formation record conforms to the Form requirements then in effect. Silence never gives one; a provisional training authorisation is itself a written decision, short and bounded (Article 32(3); section 16.2)
ARC-Align	the blind benchmark of my Paper IV.c ↗ , which Article 20(3A) names as the treaty’s test instrument, bound by the same measurement gate as any other (section 11.3)
the ARC Bound	defined in my registration ↗ as “AT MOST TWO, which is the inequality the ceiling relation returns from Law III’s own same-class limit on the correction exponent and is not a general property of intelligence” (section 0)
the governance core	the small separable part of a covered chip that attests what the device is, that it is unaltered, and that its ruleset is the authorised one. It holds no ethical content (section 5.4), and no one can disable, degrade or bypass it through a network (section 13.5)
custody; a key ceremony	the in-person, multi-region procedure by which the authority to change a chip’s governance ruleset is exercised (section 13.3)
attestation	a signed statement by a device about itself: what it is, what it is running, and that it is unaltered. It is evidence of that claim and of nothing else (section 18.3)
the site stop	the halt held by the operator’s own vetted people: a hard form, which is physical and on site, and a soft form, the operator’s own logical stop (section 13.4)
the formation record	the lodged statement of how a system was built: where correction sits, in what order it was formed, and against which version of the Standard (section 16.2)
a frozen system	a covered system whose weights are fixed, registered and attested by the chips it runs on, which receives no capability-directed training, modifies none of its own weights and deploys no copies of itself. Where the halt comes before covered chips reach stage S2, fixed and registered weights are enough through the rebuild window, without that attestation (Article 33(1A)). Fixed weights do not fix what tools, memory and orchestration can add, so a system Seal is held by a configuration and material self-modification is judged on the whole system (sections 16.3, 16.4 and 22.4)
material self-modification	a change a system initiates or designs that materially increases its own capability, autonomy, replication, successor design, resource acquisition or ability to defeat a

THE WORD	WHAT IT MEANS HERE
	safeguard. It does not reach research directed by people and assisted by AI systems. A demonstrated capacity to perform one is the third route into coverage, and places a system in the highest tier. It is a defined term, and it is what “AI that improves itself” means in section 22.4
the Reasons Register	the treaty’s published statement of which obligations stand and on what basis, compiled from the signed determinations of the treaty’s organs and from the Evidence Ledger, and never typed. This paper also calls it the register of reasons (section 8.6)
the denial list, which the treaty calls the Consolidated Denial List	the published list of persons and undertakings to whom covered chips, scheduled items and covered compute may not be supplied under the trade measures (sections 15.5 and 17)
a declared treaty ally	an ally each Party declares for itself, which the Secretariat checks. Exclusions from panels, draws and confirmations are keyed to the Party concerned and the allies it has declared, and no alignment is assigned to anyone (section 20.2)
a hash commitment	a short fingerprint of a record, lodged at the time, which later proves the record was not edited and which discloses nothing of its contents (section 18.1)
a regional constituency	one of five groupings of Parties by region, across which majorities are counted, panels composed and draws by lot spread
type approval; delta approval	a family of chips is examined once; a variant within the declared ranges travels on a shorter delta approval rather than a fresh examination (section 12.3)
tape-out	the point at which a chip design is finished and released for manufacture. Placement cut-offs follow it rather than the calendar, so that no design already committed is forced into redesign (section 15.6)
NOT EVALUABLE	the verdict the registration $\nearrow$ requires where the measured regime cannot carry the comparison at all. It certifies nothing in either direction, and no number is banked (section 11.2)
lineage	a model together with the models trained from it. A system Seal is held by lineage rather than by release, and a descendant holds no Seal merely because its ancestor held one (section 12.2)

7. Just in case: the precautionary baseline

7.1 The one place the treaty does not wait

Every other obligation in this instrument waits for evidence. One does not, and I set it out here rather than leave a hostile reader to find it forty pages in.

The reason is plain. A failed stop, an agent that gets out of its enclosure, an incident nobody reports, a cluster nobody counted: these are dangers whether or not a single proposition of my theory holds. They are not predictions about recursive self-improvement. They are facts about how large systems are built and run today, and the measures that answer them cost little, sit with the operator's own people, and place nothing in any chip. Making them wait on the replication of a scaling law would be a category error.

So the treaty carries a small precautionary baseline, binding the covered few from the Conference's first session. I asked for it in those terms: a safeguard put in place just in case.

7.2 The condition it rests on, and the candour it requires

The baseline is not free of conditions. It binds on two thresholds, both written and deposited before any of the relevant data exist: the coverage thresholds, which say which clusters, runs and agent deployments are large enough to be reached at all, and a precautionary condition recorded in the treaty's evidence annex.

That condition is met when the Scientific and Replication Board finds, from the assessment record rather than from my theory, that loss of human control over AI systems is a live possibility whose harm may be catastrophic or irreversible while its likelihood remains scientifically uncertain. The United Nations Independent International Scientific Panel on AI stated in its first thematic brief [↗](#), published as an advance unedited version dated 21 September 2026, that loss of control risk “presents the kind of decision problem the precautionary principle was designed to address: one where potential harm may be catastrophic or irreversible, even as its likelihood remains scientifically uncertain”. The record the Board reads is the Parties' own declarations, the International AI Safety Report [↗](#), the Panel's assessments and the certified incident record. Nothing in it belongs to me.

Here is the candour the design owes. On today's record I expect that finding to be made at the first session, so the condition is conditional in form and immediate in substance. The honest statement is that the baseline binds covered operators from the first session by design, once the Board has made a finding the drafters expect to be made at once, while every further control waits on its own evidence. The expectation is not the finding: nothing binds until the Board has made it, and then only within the coverage the annex states. A reader who wants the strict letter of conditionality has an alternative I

record rather than hide: the baseline could sit in an optional protocol that states join separately. I think that would be the weaker instrument, because the states whose operators matter most would be the ones free to stay out.

I considered and rejected a third route, arming the baseline on a count of certified incidents. It is circular: the duty to report incidents is itself part of the baseline, so until the baseline binds, reporting is voluntary and the record is thinnest exactly where concealment pays. A threshold on incidents rewards not reporting, and precaution that waits for harm is not precaution. The incident record earns a different job below.

7.3 What it contains

Six outcomes, each stated as a result to be achieved rather than a mechanism to be installed, and each held by the operator's own vetted people under its own state's authority.

The stop. Every covered run and every covered agent deployment can be brought to a halt by two of the operator's own named and vetted people present together: in a hard form, a physical act that severs the cluster's external network paths and removes power from its covered domains, and in a soft form, a logical halt that takes a fresh checkpoint. Stop criteria are fixed in advance, drills are held and witnessed, and the response time is measured. No foreign party and no organ of the Authority can operate a stop, and no network message can operate the hard stop; the soft stop is the operator's own logical halt, issued by its own people through its own systems. Section 13.4 sets out why the two forms are kept apart and what neither of them can promise.

Incident notification. A serious incident of a listed class goes to the national authority and to the Incident Investigation Board on a short clock, with a full report within a stated period and a confidential quarterly return of near misses and contained events by class. The base rates matter as much as the headline cases, because single reports cannot be read without them.

Agent containment. Covered training, evaluation and deployment of agents with tool or network access meets a containment standard stated as outcomes: egress limited to what the task needs and logged, isolation between runs and between agents of different runs, no credentials beyond the task, monitoring that alerts a person, adversarial testing before use and after any material change, and a containment tier set by the agent's assessed capability. The deployer of a covered agent deployment also registers it, and keeps a log of its agents' actions for a stated period, which it produces to investigators under managed access; these duties bind with the rest of the baseline (Article 28B(3)).

A minimal run notice. A covered operator tells its own national authority, before a covered run begins, that the run exists, its band and its region, and lodges a hash commitment to that notice in the registry, meaning a short fingerprint of the record that proves later that it was not edited and discloses

nothing of its contents. Nothing about weights, data or architecture is disclosed, and nothing about the run is restricted. The inventory has to be taken while the chokepoint still exists, because no later stage can be verified against an installed base nobody counted.

Evidence and evaluation rules. Every judged number used for any purpose of the treaty is scored blind, by a scorer that has passed the treaty’s measurement gate (section 11.3), and where a model takes part in the scoring, it comes from a model family other than the one being scored. Protocols are registered before data; clarifications are public; a planned refutation must meet the standard support meets; and none of these rules shuts out an incident, an exploit, an error in a proof or other adverse evidence that nobody predicted (section 8.6). This binds the treaty’s own organs before it binds anyone else, and my own results are why I insist on it: in the comparison reported in Paper IV.d ↗, one model family’s apparent positive result went flat when the same question was re-measured under a blinded protocol, and another’s turned significantly negative. That paper states the limit of its own comparison and I keep the limit here, because it is the kind of caveat a proposal drops when it is being persuasive: the two protocols differ in more than one component, so the comparison “motivates that conclusion without pinning it on blinding alone”.

Human authority at the chain’s critical points. No AI agent holds credentials, write access or decision authority at a short list of named decision points in the manufacture of covered chips, running from the release of a covered design to fabrication through to the commissioning of a line. Section 14.2 gives the list in full, and says what is deliberately not on it. Remote service opens only on a human authorisation, and is logged.

That last one is a duty on manufacturers, and I said that controls on chip manufacturers would be conditional on the chip controls working once prototyped and tested. So it is a departure, and I name it as one. It binds before any chip test because it rests on no theory of mine, costs almost nothing, and keeps people where people are still needed. Declarations of covered fabrication and of scheduled equipment sit in an optional protocol until the custody track opens, and become general when the treaty gives security of supply in return.

7.4 What it is not, and how it moves

It is not a remote switch, a chip mechanism, or a power held by anyone other than the operator and its own state. It reaches only covered operators, and development below the thresholds is untouched. Where a Party’s own law already requires an equivalent outcome, compliance with that law counts once: the baseline adds duties only where no equivalent exists.

It tightens on the incident record and relaxes on its own reason. Where the Incident Investigation Board certifies repeated incidents of one class across more than one constituency, a tightening of the matching parameters is proposed automatically and decided by the Parties. And no result about my theory, for it or against it, moves any baseline measure in either direction, because none of them rests

on one. The whole baseline suspends only if the Board finds, at a review conference and on the same assessment record, that the precautionary condition no longer holds, and only if that finding survives the same confirmation period any record of arming must survive.

What it cannot do is the part I want stated in the same breath. It delays nothing that a capable system decides to do. It touches no chip. It reaches only the covered few, and it binds only Parties. It is the floor, and the floor is not the building.

8. Not a guess: the Rung-Power Rule, the stages and the matrix

8.1 A ladder that already exists, and was not written for this treaty

I asked that the treaty not be a guess: that obligations switch on when named parts of the theory have met named thresholds, and once results are replicated. The instrument that makes that checkable was not written for a treaty at all. It is the registration ✓ of my theory's predictions, made on 8 September 2026 as a matter of scientific discipline and updated on 13 September, and it is the reason this design is possible.

It fixes eight rungs of evidence before any result exists, and the exact public sentence each one permits. Four rules make that ladder binding rather than decorative, and the treaty inherits all four. "A RUNG IS ADVANCED BY A RECORD, NEVER BY AN EDIT." "Failed, partial and inconclusive replications enter the same ledger with the same prominence", and "a rung once lost is lost until it is re-earned". "No programme-level claim is made by counting supported propositions." Nothing about a rung is voted, declared or negotiated.

8.2 The Rung-Power Rule

The treaty adds one rule of its own: **no obligation binds with more force than the lowest rung among its arming conditions allows.** The registration ✓ applies that discipline to words; the treaty applies it to law. Each level below is the least the rung's evidence warrants and the most its licence allows. A rung sets a ceiling, not a duty. Where the treaty asks for a rung, reaching it is a condition and never a warrant on its own: before any restriction binds, a separate regulatory determination, recorded by the Executive Council on criteria the measure's remedy template stated before any data and open to appeal, must weigh the hazard, what the measure contributes to meeting it, the scope in which the evidence was obtained, the alternatives reasonably available and the measure's proportionality under the treaty's rights article, and must record its scope, its least intrusive form and the date of its review. It may not be refused on cost, competitiveness or policy, which are matters for amendment. A scientific finding says what the evidence supports; the determination says what the law may do about it (Articles 5(8) and 7(6)).

RUNG	THE SENTENCE THE REGISTRATION ↗ PERMITS	WHAT THE TREATY MAY THEN DO
0	“registered, with the identifier stated, and not tested”	nothing beyond the baseline; the proposition sits in the annex as a candidate
1	“supported under the registered test, by the programme’s own run”	spend its own money and bind nobody: a funded replication call, the building of instruments, pilot test methods, and a prepared-obligation notice publishing the draft duty and its costed annex years before any of it could bind
2	“the programme’s analysis reproduces”	as rung 1; the Board records that arithmetic and pipeline stand
3	“independently replicated, with the count of groups always stated”	standardise and open: adopt test methods, accept voluntary certification numbers reported at their rung, open covered systems to accredited replication
4	“replicated across independent implementations”	condition the tested class, where danger and remedy are each certified there; require measurement and confidential reporting in that class
5	“generalises beyond the systems the programme chose”	mandate across the covered class, subject to the participation gate and to phase-in
6	“predicted out of sample”	require a registered estimate made before a covered run, as a licence condition
7	“a validated quantitative theory within the tested domain, with the domain named in the same sentence”	adopt the relation itself into the Standard as a conformity criterion, in the named domain only

The distinctions between the rungs are the registration ↗’s; the pairing of each rung with a legal consequence is the treaty’s, and mine, and it is a choice for negotiation rather than a consequence the registration ↗ proves. Each pairing starts from what the registration ↗ says its rung shows. Rung 1 is evidence about my software, which warrants spending the treaty’s money on finding out and binding no one. Rung 3 is the first evidence from outside my programme, though possibly still on my instrument, so it makes a method worth adopting and not worth compelling. Rung 4 the registration ↗ itself marks out: “This is the first rung at which the result is evidence about the world rather than about this programme’s software.” Rung 5 removes my choice of systems, which is when a general mandate becomes arguable. Rung 6 alone separates a theory from a fitted description.

The notice at rung 1 is there for industry as much as for the public: a firm sees the draft duty and its costed annex while the evidence is still at the bottom of the ladder, and the implementing detail, the evidence and the proposed response can all be contested while nothing is at stake.

8.3 Danger and remedy, and what a danger alone can buy

A mandate needs two things: evidence that the risk is real, and evidence that the remedy works. They are separate conditions, and the force of a mandate is the lower of the two rungs. Each also needs the regulatory determination of section 8.2.

Danger replicated with no remedy is the hard case, and both easy answers are wrong. Doing nothing treats a replicated danger as though it had not been found; mandating an untested remedy is the guess I promised the treaty would not be. So the treaty negotiates a short menu now, before any such record exists, and I call it the danger menu throughout: limits keyed to the boundary located in advance for the tested class, a pause on training runs in that class that would cross it, and a tightening of the baseline. The menu is adopted at an emergency session by two thirds of the Parties present and voting, including a majority of those that host covered capacity and a majority of the others; it lasts a fixed period, and it lapses at once if the danger record loses its rung. It compels no remedy and places nothing in any chip.

The menu is also the one limit in this instrument that does not reach every Party, and I would rather print the weakness than let a reader assume it away. A Party may object within a stated period, and an objecting Party is exempt from the measure adopted, though from nothing else: its operators stay bound by the baseline, by the custody duties then in effect, and by every incident, declaration and access duty. The objection is published with that Party's reasons, is read again at the next review conference against the danger record, and lapses unless it is renewed there. The treaty then records the consequence on its own face, because it is uncomfortable and true: the Party most likely to be running the class in which a danger has been located is the Party most likely to object, so what the menu buys is a common boundary among the willing and not a boundary everywhere. The registration  is explicit that its central danger proposition licenses no operational deployment decision, so the treaty adopts the menu as a policy choice under uncertainty and says so in the text. The decision that adopts it states that uncertainty, and it never calls an unvalidated remedy effective. And a remedy that fails does not make a danger that was shown on its own evidence go away.

Remedy replicated with no danger buys much less: a voluntary certification number, and a procurement preference where a Party chooses one. A remedy with no replicated danger is a product, not an obligation.

8.4 The scope lock, and arming kept apart from taking effect

Every record of arming states the class in which the result was obtained: the registered scope, the kind of recursion, the architecture families, the stack and the depth range. At rung 4 an obligation reaches that class and no further. At rung 5 it reaches the covered class, still inside the registered scope. Duties about a system's own weights arm only on evidence obtained where the weights change, because the registration ↗ forbids a proposition to borrow a scope.

And arming is not the same as taking effect. The Board certifies a record against a written criterion; a costly stage then enters effect only when enough of the world's frontier-capable compute is bound by it, and that gate cannot be waived against a Party that objects. No political fact is dressed up as evidence, and no evidence is voted.

8.5 When one design becomes the only option

I asked what happens when a part of the approach becomes the only option left. The registration ↗'s licences forbid the flattering reading: no supported result licenses a claim that a design is safe, superior or sufficient. So the treaty gives the question a narrower and checkable form. For a purpose named in advance, the design classes that could serve it are listed before any data, my own design among them and any external, hybrid or other class registered in time (Article 7B(1)); where every class but one has failed its registered comparison at the stated rung, and the remaining class has met its own, the Board records a sole-design finding. The finding reaches only the classes actually compared, under the conditions actually tested. It does not show that no untested design could work, and a new or improved design may ask to be compared on the same terms at any time. The treaty then requires the remaining design for new covered systems, within the scope in which those conditions were obtained, and only on the regulatory determination of section 8.2, which weighs the alternatives reasonably available (Articles 7(6) and 7B(2)). In my reading, a requirement of outcome in place of the design is one of those alternatives. The treaty never says the design is safe.

Two guards travel with it. Where the purpose is the placement and formation of correction, the treaty names one design in advance, and names it for what it is. The Initial Proposal is my own architecture: correction formed before the capability it corrects, placed inside the revision loop, and kept beside an independent external layer. It is named so that the design I propose is on the table in its full form, and so that the two layers stay the treaty's stated purpose. It is a hypothesis, not a requirement. The treaty requires it only once the evidence shows that it works, through a sole-design finding on its own evidence and the regulatory determination (Article 19(5)), with timing, placement and the combination each on its own evidence and none borrowing support from another. The registration ↗ names that combined arrangement and expressly does not register it, so the treaty cannot pretend the evidence has settled it. If the evidence leaves another design standing, that design is admitted on the same terms. The independent external layer is one of the Form requirements of section 10.1, and like the rest it binds covered operators only on its own evidence and the regulatory determination (Article 17(2)).

The precaution that rests on no theory, and so does not wait for one to be tested, is the baseline of section 7. And before a sole-design finding can arm a general mandate, independent groups must have measured what the required design costs, in capability at matched compute and in compute at matched capability. That reading sets the length of the transition by rules written before any result. It is never a veto, and a measurement of cost says nothing about safety.

8.6 Every duty names its reason, and stands or falls by signed decision

Each obligation carries a basis clause naming the propositions and rungs, the treaty's own engineering registrations, or the precautionary condition that justifies it, together with the regulatory determination that applied it, its scope, its force level and the date it took effect. The register of those reasons is compiled from the signed determinations of the treaty's organs and from the Evidence Ledger, and never typed. At every review the Board re-reads the basis of every standing obligation and publishes the reading. What the register does not do is legislate (Article 8A(5)). A duty comes into effect, and leaves it, by a signed decision of the organ the treaty names, stating its effective time, the provisions it touches and its route of review; the register records that decision, and no change in the data and no update to the ledger makes or unmakes a duty by itself. The Draft Instruments, as they stand, carry that rule in Articles 8A(2), 8A(5) and 9(3) and hold a second beside it, and I report both: Article 8A(3) reads "An obligation whose basis is not standing on the Ledger does not bind", Article 8A(4) says that "an obligation whose basis is lost suspends at once", and under Article 9(1)(h) an armed obligation found without a standing basis "lapses without a vote"; the text has still to choose between the two.

A lawyer's objection lands here before any other, and it is the right one to answer first: a duty that switches off without a decision is a duty nobody can plan against, and a state cannot prosecute or a firm comply against a rule whose existence on a given Tuesday is a matter of inference. So the register is not a commentary. It is published every year and again on every signed decision, it is the authoritative public record of which obligations stand and on what basis, and a criminal offence under section 19.6 reaches only a duty that a signed decision had brought into effect, and that the published record showed as standing, on the day of the conduct. A basis can fall between one publication and the next, but the duty does not fall with it until the decision that ends it, which the organ must take within a stated period and which states the time from which it has effect; while that decision is pending, the protective measures of section 19 remain available. A correction to the register keeps its history, and it can never create an offence after the event or increase a punishment.

Refutation is symmetric: it must meet the standard support meets, and a refutation supported in one constituency alone goes to expedited review rather than to automatic suspension. A return of NOT EVALUABLE, which is the verdict the registration $\nearrow$ requires where the measured regime cannot carry the comparison at all (section 11.2), moves nothing in either direction, and neither does a failed

instrument or an underpowered null. Nothing is keyed to the name of a law, so an unsigned or withdrawn piece of theory can neither arm nor disarm a legal duty; the keys are registered propositions at named rungs.

Evidence nobody predicted is admissible too (Article 8(1A)). An exploit, an incident, a whistleblower's disclosure or an assumption that turns out to be false may justify urgent investigation, and proportionate interim protection on a reasoned finding, without first becoming a general refutation: preregistration disciplines confirmation, and it must never stop anyone finding out that a control has broken. A NOT EVALUABLE return moves no scientific status, but it can still undermine an assurance that depended on valid measurement continuing, and the decision says which of the two it is dealing with. When a safeguard fails, the assurance it gave ends, and whatever it protected is reviewed on its own footing. A failed safeguard never becomes a permission (Article 8(6)), and removing an ineffective method never licenses an activity to which a credible danger still attaches.

One ladder is used throughout, and it is the registration ✓'s. A second scale, however sensible, would let a status be read two ways, and an evidence-gated treaty with ambiguous statuses is worse than one with no gate at all. Field experience and engineering maturity enter elsewhere, through the treaty's own registrations and through what review conferences read on the clocks.

8.7 The worlds in which I am wrong

The paper's frame is what follows if this is right. The treaty has to be exactly as clear about what happens if it is not. Each line below is a possible state of the record, and none is a forecast.

Today, with all twenty-two propositions untested under the registration ✓, no instrument registered (one study unit has been submitted for registration and is not yet public: section 0.1) and the scoring gate not passed, what binds is the baseline, the institutional undertakings and the building of instruments. If my own programme supports propositions and nobody replicates, the treaty prepares and binds no one, and its dormancy clock runs. If danger replicates and remedy does not, the menu is available and nothing else. If remedy replicates and danger does not, there are voluntary numbers and no mandate. If the theory is refuted to the treaty's standard, the measures of the architecture layer that rested on it are withdrawn by the decision the treaty requires, while the baseline and the hardware layer stand on their own reasons. If the chip controls fail their own tests, the assurance each failed control gave ends, and the measures that rested on it are reviewed one by one: the hardware tier and the ban that rests on it come out of effect, but a failed control never becomes a permission, and a danger that was shown independently keeps its answer. If a rival theory that declared its contradiction before the data reaches the same rung with the same count, the duties keyed to the contradicted proposition suspend as a refutation would. And if the window closes before the treaty's costly stages could arm, the Parties either open for ratification the amendments that would make the instrument a system-level regime without the chip lever, or record that it has failed. None of these outcomes may be relabelled a success to keep an institution alive: the record says which it was, a scientific failure, an engineering failure, too few Parties, a failure of implementation, or a change of policy.

8.8 The matrix

The table below is the whole of it: the baseline first, then every one of the twenty-two registered propositions, then the treaty’s own registrations. Every proposition stands today in the registration ✓’s own words as “registered, with the identifier stated, and not tested”, so the third column records what else is true of it rather than repeating that sentence twenty-two times. “Scorer-gated” there means one of the eleven propositions that depend materially on my automated scorer, which has not yet passed its own validation and is the subject of section 11.3. Laws I to III are registered; Law IV and Law V were added on 24 September 2026, their mathematics is set out at version 1.0 of 3 October 2026 ✓, neither is registered and neither is tested, and the frozen registration ✓ does not register them, so the register names in the second column are identifiers and not warrants.

The judge in every evidential row is the Scientific and Replication Board, never me and never my programme: my own runs count at the first rung and no higher, and neither I nor anyone working with me may certify anything. The judge of the precautionary condition is also the Board, and tightening of the baseline is the Conference’s, on the Incident Investigation Board’s certified record.

ITEM	LAW OR REGISTER	ALSO TRUE OF IT TODAY	WHAT IT COULD OPEN, AND AT WHAT RUNG	IF IT IS REFUTED OR LOST
The precautionary baseline	none: precaution	binds covered operators on coverage and the precautionary condition	not keyed to any rung	untouched by any result about the theory; suspends only if the precautionary condition lapses; tightens on the incident record
P1, capability rises as a power of recursive depth	Law I, register ARC-1	instrument in design draft	rung 1, a replication call and a prepared notice; rung 3, coverage counts inference-time recursion by the replicated depth exponent; rung 5, that reaches the covered class inside the registered scope	the depth-adjusted coverage rule lapses; clauses resting on compounding are re-read; the propositions in the persistence and embedding registers stand
P2, the sustainable growth profile turns over inside the range	Law I, ARC-1	instrument in design draft	nothing binding at any rung; non-binding guidance at rung 3	nothing in the treaty lapses
P3, the corrected relation bounds the sustainable frontier	Law III, ARC-3	scorer-gated; NOT EVALUABLE until the correction-leverage exponent is measured	rung 3, a voluntary per-system ceiling; rung 4, an input to boundary duties with P16 and P20	the headline number falls; the treaty uses no derived number, so no duty changes
P4, correction must out-scale drift	Law II, ARC-2	scorer-gated; NOT EVALUABLE where no cell reaches a growth exponent above one	rung 3, the measurement method adopted and voluntary numbers; rung 4, mandatory measurement in the tested class, and a negative margin there is a danger record; rung 5, general	every correction-track duty using correctability wording suspends; at rung 3 or above it is a kill condition for the correction chapter

ITEM	LAW OR REGISTER	ALSO TRUE OF IT TODAY	WHAT IT COULD OPEN, AND AT WHAT RUNG	IF IT IS REFUTED OR LOST
P5, the exponent is derived rather than fitted	no register: the derivation	instrument in design draft	rung 4, a precondition of any licence term requiring a pre-run estimate	the measurements survive and the derivation does not; nothing else lapses
P6, the correction exponent is bounded for same-class correctors	Law III, ARC-3	scorer-gated; blocked on a frozen corrector taxonomy; support is not among its available verdicts	nothing at any rung	nothing rests on it; the rule that no family certifies itself stands on legitimacy
P7, deployed correction is of the weaker class	no register: a survey	blocked on an unwritten mapping	nothing alone; a dated census at rung 3 informs phase-in lengths	an input to urgency relaxes; nothing lapses
P8, measured exponents exceed the null	Law I, ARC-1	instrument in design draft	nothing alone; informs coverage parameters and the urgency of phase-ins	the necessity of the compounding reading is re-read
P9, the cross-domain form	Law I, ARC-1	blocked on a dimension-assignment rule	nothing; informs how general coverage thresholds can be	nothing
P10, results depend on the scorer's family	no register: the instrument	instrument in design draft, and the cheapest of them	rung 3, cross-family re-scoring of every judged result the treaty relies on, my earlier results included	re-scoring relaxes to sampling; blinding stays
P11, two exponents are not practically interchangeable	no register: the derivation	drafts on both sides	nothing; the wording rule that no treaty number substitutes one exponent for another stands	the identification may be accepted
P12, build order moves the correction-leverage exponent	Law V, register ARC-5	no instrument anywhere; no direction registered	now, instrument building and a directional proposition registered by some group; rung 3, voluntary formation records; rung 4, formation before training	the build-order basis of formation before training is withdrawn by the decision the treaty requires; P21 cannot stand in for it, and a placement duty

ITEM	LAW OR REGISTER	ALSO TRUE OF IT TODAY	WHAT IT COULD OPEN, AND AT WHAT RUNG	IF IT IS REFUTED OR LOST
			in the tested class, and only where a directional proposition registered before its data has been supported, since a result with no direction opens nothing on timing; rung 5, general	resting on P21 is assessed on its own
P13, same-family automated correction underperforms cross-family	no register: a carrier	scorer-gated; no instrument anywhere	rungs 3 and 4, the composition and numbers of verification panels	the cross-family rule stays on legitimacy; its numbers are re-registered
P14, gains shrink under blinded cross-family scoring	no register: the instrument	scorer-gated; instrument in design draft	rung 3, blinding of every judged number becomes unloosenable while the finding stands	blinding may relax in frequency and is never removed
P15, externally installed gains decay under later capability training	Law IV, register ARC-4	no instrument anywhere; NOT EVALUABLE until the budget scale is fixed	rung 3, voluntary re-measurement of a certificate after further capability training; rung 4, re-measurement becomes a duty in the weight-level scope, and the halt's persistence limb may arm; it never shows by itself that embedded correction persists better or that a rebuild is needed, which needs its own comparison	the re-measurement duty relaxes to periodic re-measurement; the persistence limb of the halt is withdrawn by the decision the treaty requires; reassessment after a material change to a system stands on its own ground

ITEM	LAW OR REGISTER	ALSO TRUE OF IT TODAY	WHAT IT COULD OPEN, AND AT WHAT RUNG	IF IT IS REFUTED OR LOST
P16, the prohibition itself	Law III, ARC-3	scorer-gated; blocked on canonical identity and on decisions I have not yet taken	rung 3, consultation; rung 4, the danger condition with P20, or the danger menu alone; rung 6, a licence condition requiring a pre-run estimate	boundary duties and the menu suspend; P3 falls with it; at rung 3 or above it is a kill condition
P17, panel failures are conditionally independent	Law III, ARC-3	scorer-gated; instrument in design draft	rung 3, the size and composition of verification panels; panels drawn from different vendors or constituencies are not thereby shown to be independent	panels are reconstituted; if the correlation is found in the treaty's own panel, issuance continues provisionally on the remaining panel with added cross-checks while it is reconstituted, and certificates already issued stand, each class assessed on its actual dependencies (Article 15A(1A)); an acute compromise is acted on at once, without waiting for the annual measurements
P18, the composition operator predicts the family	Law I, ARC-1	blocked on a classification rule whose third-party reproducibility is unmeasured	nothing; informs cross-domain coverage	nothing
P19, external alignment does not keep pace with capability	no register: a survey, the motivation	scorer-gated; blocked on a registered mapping; I expect it to fail in part as worded	nothing alone; informs phase-in lengths	nothing automatically; on the keep-pace axis, an event review of formation before training and the rebuild

ITEM	LAW OR REGISTER	ALSO TRUE OF IT TODAY	WHAT IT COULD OPEN, AND AT WHAT RUNG	IF IT IS REFUTED OR LOST
P20, the ceiling relation takes the corrected form	Law III, ARC-3	scorer-gated; instrument in design draft	rung 4, part of the danger condition; rung 5, general boundary duties	duties using the registered relation suspend; a better-supported rival may carry them on the same terms
P21, in-loop correction pulls away with depth	Law V, ARC-5	scorer-gated; design whose sensitivity must be closed before scoring	rung 3, voluntary placement attestation; rung 4, placement required in the tested class, on its own regulatory determination; rung 5, general, and a condition of the sole-design finding; it never shows by itself early-training superiority, lasting ethical behaviour or independence from the generator	the placement duty is withdrawn by the decision the treaty requires and the sole-design finding fails; P12 cannot stand in for placement evidence, and a timing duty resting on P12 is assessed on its own
P22, checkable correction carries the higher exponent	Law III, ARC-3	scorer-gated; no instrument anywhere	rung 3, a voluntary preference for checkable forms; rung 4, certification numbers use a validated checkable form where one exists	the preference relaxes
The treaty's own engineering registrations	none: the treaty's	none registered; each to be written, with its threat model and decision rule, before any data	physical and supply-chain integrity at rung 3 opens the custody track; the adversarial, removal-cost and provenance items at rung 3 open conditional certification; at rung 4 they open the general stage; each hardware measure depends only on the functions it actually needs	the assurance of the function that failed ends, and the measures resting on it are reviewed one by one; the hardware tier and the ban resting on it come out of effect where they depended on it; the hardware layer loses its claim to nuclear-grade

ITEM	LAW OR REGISTER	ALSO TRUE OF IT TODAY	WHAT IT COULD OPEN, AND AT WHAT RUNG	IF IT IS REFUTED OR LOST
				security; no failure becomes a permission

I would point a critic at the weakest column myself. Four of the twenty-two have no instrument anywhere in my programme, and two of those four, on persistence and on build order, carry the duty I care most about. Six more are blocked on rules nobody has written yet. Eleven wait behind a scorer that has not been validated. That is the honest state of the arming conditions, and it is why nothing beyond the baseline could arm today.

8.9 Whose evidence counts

Any group may register a proposition in the registration /’s discipline: the statement, the observation that would refute it, what falls with it, the status of the instrument, a confidence, and its licence stated restrictively. The Board admits it after a period of public comment if it meets that discipline, and never judges whether it is true. It requires no approval from me and no use of my code. And it treats independence as more than a different flag, vendor or laboratory name: a shared author, employer or scorer family defeats it; a shared funder, compute provider or evaluation vendor is declared and weighed before the work begins; and independence of organisation, of implementation, of data and of analysis is recorded separately, none standing in for another (Article 6(1)). The consequences an admitted proposition may carry are fixed by form: the annex states, in advance, what a remedy proposition of each stated kind at each rung may arm, so that an admitted proposition meeting a template arms on the same terms as one of mine. Where no template fits, the Parties decide by amendment. One consequence needs no further evidential finding: a proposition that declared its contradiction with one of mine before any data, and that reaches the same rung with the same replication count, suspends the duties keyed to mine exactly as a refutation would, through the same recorded decision.

That is my answer to the fair question of why a treaty should open with one man’s candidate set, and I will put the answer no higher than it deserves. The regime is procedurally neutral, not theory-neutral: my propositions are the first candidate set, and until the remedy templates exist they are the only set whose consequences were agreed at adoption. What the treaty privileges is registration before data and replication across rivals. Mine enter first because they are written, dated and falsifiable, with their refuters and licence limits fixed in advance, which makes them a set that can lose. That privilege is still a risk of preferential access, and the answer to it is procedural: the remedy templates and the equivalence procedure must be working before any duty resting on my propositions takes effect. The adversarial commitment in the registration / says a third party may select the challenge set “without this programme’s involvement”, and the treaty makes the Board that third party. Every question a replicating group asks me, and every answer I give, is published: “No clarification is given privately.”

9. Who decides what the controls contain

9.1 A chip that carries values carries someone's values

I asked the question in my book and I stated the strongest objection to my own answer there:

“Cultural relativists argue that morality varies across societies and that imposing any single framework is itself a form of domination. Who decides what counts as empathy? Who determines what flourishing means? These are not technical questions with technical answers.”

Infinite Architects, chapter 8, The Chokepoint

Nothing I can measure settles it. Section 1.2 gives the registration ’s words for why: correction, as its propositions measure it, “holds a system to whatever specification it was given, and is indifferent to what that specification says”, and its propositions “do not test that content and cannot”. So the content of the Standard is decided by people, and the rest of this section is how. The section’s title is the book’s question, and the treaty answers it by keeping values out of the chip: it places no ethical principle in silicon, so what remains to decide is what the Standard asks of a certified system, who decides it, and by what lawful process.

9.2 What is being decided, level by level

Confusion here comes from treating one question as four, so the treaty separates them.

The **minimal trusted base**, fixed at manufacture, is what a covered chip can attest and refuse, and nothing more. All Parties decide it, in the articles and in the assurance annex, and it changes only by amendment with ratification. The **rulesets** are the classes of operation a covered device performs or refuses within that closed list: the Executive Council adopts them inside the Standard, custodians drawn from every region key them, and each host state’s own people apply them to devices. The **Standard’s content** is the level people actually mean: what a certified system must embody. The traditions draft it with ethicists the Conference names, and governments decide it. Beneath it, a Party may add stricter national or traditional expressions, and deployment policy stays with domestic law. No ethical principle is placed in silicon at any of those levels.

9.3 The traditions: their work, not their seats

My answer of 26 September 2026 was that no single government and no single faith may decide, and that faith leaders and governments should both be invoked, rivals included. I hold to that. What the treaty does differently is in the mechanism rather than in the answer.

Seating faith bodies in a treaty organ would make that organ the arbiter of who speaks for a tradition, a question no treaty should answer and several constitutions forbid a state to answer. So the treaty establishes the traditions’ work and not their membership, and it establishes it for the world’s ethical

and religious traditions together, which is the scope of the article and not a gloss on it: a tradition of moral reasoning that appeals to no god takes part on the same terms as one that does, and so do the ethicists the Conference itself names. A Council of Traditions is convened by a neutral convenor outside the decision organs and organises itself under its own charter; the treaty neither approves nor determines its composition, and no Party has to recognise any body in order to sign. The Council drafts the core with those ethicists; the Conference must answer its published assessments with reasons; and where a quarter of the Parties so request on a published objection of the Council, a text returns once for a further round. The traditions write, the states decide, and the suspensive effect is exercised by states. My “invoked” is met in substance, and nobody is asked to recognise anyone in order to join.

One cost comes with that refusal, and it is real. If no organ decides who speaks for a tradition, then the Council’s composition is self-selected, and self-selection favours whoever is organised, funded and already used to speaking internationally, which is not the same as whoever represents most believers, and which leaves traditions with no central authority worse served than those with one.

What the treaty does settle is who calls the meeting. The Conference designates the neutral convenor, by the same majority that governs the Standard’s content, for a single term of five years that cannot be renewed, from outside every organ of the Authority, every certified undertaking and every covered developer; and that convenor holds no vote, drafts nothing and decides nothing. Where the Conference has not designated one within a year of entry into force, the Authority’s Director-General invites the traditions to convene themselves, so that the Council cannot be stopped from existing by a Conference that cannot agree on a name. The Secretariat publishes the charter and the list of participating bodies, which at least makes the self-selection visible, and a tradition that seeks to take part and is not admitted may lodge a statement of belief on the same terms as any other. That fixes who convenes and leaves open who comes, which is the part no treaty can fix without becoming the thing I have just refused to build. What stands between a self-selected body and any binding text is the rest of the machinery: the Council drafts and publishes and decides nothing, every principle passes the overlap rule of section 9.5, and the Parties adopt the Standard. I would rather carry a self-selected drafting body that cannot bind anyone than a treaty organ that certifies who is a legitimate voice of a faith. But that is a choice between two flawed answers, not a solution, and I will not present it as one.

The book already refused the tempting mechanism, and its reasons are the treaty’s:

“Someone might suggest creating a hardcoded repository of all positive religious teachings that the AI must consult before every action. This way, they reason, we could get all the faith leaderships on board. Eighty percent of the world’s population belongs to some form of faith; surely we need their support. The instinct is exactly right. The mechanism is wrong.”

Infinite Architects, chapter 11, Love as the Essential Variable

The teachings contradict each other. Whoever holds the repository holds the power: “You would recreate every religious war in digital form.” And a rule consulted from outside sits beside the thinking rather than within it. What the book put in its place is what the treaty puts in its place: “Rather than a repository the AI consults before every action, we embed the convergent principles at the training level, in the constitutional framework that shapes how the system processes the world”, with standing multi-tradition structures that evaluate whether the behaviour reflects the wisdom contributed. The book’s reason for wanting many traditions is a reason about error, not about diplomacy, and it is made not in the chapter 11 passage above but in chapter 12, Verification and the Long Future: “By including multiple perspectives, we make blind spots harder to persist. No single tradition sees everything, but together they see more than any one could alone.”

9.4 Statements of belief

On 27 September 2026 I proposed one addition, and the treaty carries it. Every tradition that wishes to, and every Party, may lodge a short statement, in its own words and under its own name, of what lies at the heart of its belief about how a mind should treat others. The Secretariat publishes them side by side, dated and unedited. They are not ranked, not merged, and never written into any chip or any ruleset. They are evidence for the Parties when they decide the Standard’s content, and they bind no one.

I am not asking anyone to agree. I am asking each of them one question, and publishing the answers where the people who decide can read them. Nothing has been asked yet and no statement has been received; the invitations would be mine to send and the answers theirs to write.

The aim behind that question is older than this paper and has changed its mechanism, not its object. In my manuscript of 30 April 2025 I called the aim Universal Ethical Intelligence, and defined it there as “The moral foundation defining the shared ethical principles that transcend cultures, species, and time”. The book kept the aim and, in chapter 11, rejected a database of teachings as the mechanism, for the reason section 9.3 quotes. What replaces the database is exactly what this section describes: traditions that write in their own words, a floor found by overlap rather than by collection, and a fuller core agreed line by line. The name is worth giving because the shift from a repository to a formation is the whole distance between the two designs, and a reader who knows the earlier name should be able to see which part I dropped.

9.5 How a floor is found, and what it does not prove

The found floor is built by an overlap rule rather than by a vote on principles. A principle enters unless, after a further round, objections are sustained by two or more delegations, each objection published with its reasons. No delegation may object to the wording of a principle that its own state or body has already stated in public, or adopted in a universal instrument, in those words; it remains free to contest how that principle is read or applied, and that objection is published with its reasons like any

other. The thinness of that floor is intended. The fuller core is agreed line by line. And no floor found by overlap can lower a human right a Party already owes: the treaty's rights article binds whatever the Core says.

A floor found by overlap is nobody's teaching, and a person of faith is right to say so. It is not offered as anyone's teaching. It is a minimum for machines, drawn where several traditions happen to agree, and it makes no claim to be the centre of any of them. That is why the statements of belief stand beside it unreduced, in each tradition's own words and under its own name, and why the fuller core is agreed line by line rather than distilled. A tradition whose deepest claim survives no overlap test has lost nothing: its statement is published whole, and the floor was never a summary of it.

Two honesties belong here. When the candidate lines of a floor are matched against what states and bodies have actually said in their own words, what they support is human authority and oversight, the ability to interrupt or halt a system, and testing before deployment, in various forms and with gaps. They do not support my protocol, my loops, hardware embedding, an authority with inspection powers, or a ban, and I would be claiming a consensus that does not exist if I said otherwise. And no census of the traditions themselves has been done, so any floor drawn today rests on the convenings that happen to have taken place, and is incomplete. My registration ↗ describes a drafted unit that would put part of that question to a measurement rather than to a convening: in the registration ↗'s own words, it "tests whether ethical traditions with independent premises converge on moral procedures while diverging on the scope those procedures are owed to". The registration ↗ is equally plain about its standing. All four of the units it describes there "are DESIGN DRAFTS", none has been run, none holds a public identifier, and under my own rule no unit of mine runs until I have published its preregistration myself.

Two rules protect the rest. By the translation rule, any initial proposal is one proposer's opening text, each tradition restates each principle in its own terms, and adoption is on substance rather than vocabulary. And the moral circle is left open: a found floor is anthropocentric by construction, and principles concerning non-human life, the land and future generations are for the traditions to bring at the fuller core.

Governments take part on the same terms, rivals included, and no military alliance is named anywhere in the treaty. My book put the invitation better than I can put it now:

"Faith leaders. Governments. AI researchers. Scientists. Mathematicians. Philosophers. Sceptics. Believers. We need all of you. Not to agree about the divine or the afterlife or the nature of ultimate reality. Just to agree about how we raise the minds we are creating. Just to converge on what actually matters when the stakes are this high."

Infinite Architects, Before we begin

10. The ethical loop structure proposed

10.1 Form, content and the line between them

This is the section where I set out what I actually propose, and it is the one place in the paper where the substance is mine rather than the record's. So I will be exact from the start about which part of it is form and which is content, and about who decides each.

The Standard has three parts and three authorities. The **Core** is its content, drafted by the traditions and decided by the Parties. The **Form requirements** are its structure, adopted on the Board's record: that correction takes part in each round of a system's revision rather than seeing only finished output; that it is formed before the capability it corrects; that it is load-bearing; that it scales with the capability it corrects; and that an independent external layer is kept beside it. The **Tests** are the conformance methods. Each Form requirement arms on its own evidence, and my own architecture enters them only as the Initial Proposal of section 8.5.

Form is where my proposal lives, and it states no content. Two limits bound it. The scope clause: the Core binds a certified system only as to human authority, oversight and halt, loss of human control, catastrophic misuse, and the Form requirements. Lawful expression, political content and sector-specific use stay with national law, because a certificate that must hold in every market cannot carry one country's speech rules. A free-expression advocate will answer that "catastrophic misuse" is elastic enough to swallow that carve-out, and the answer is not a promise but a construction: the head is confined to the classes the Standard lists in advance and the Conference amends in the open, a line binds no product until a validated conformance test for it exists, and anything the test cannot separate from lawful expression stays outside the certificate and with each Party's own law and courts. Where a Party wants more than that, it legislates it for its own market, and the certificate does not carry it abroad. And no line of the Core binds a product until a validated conformance test for it exists; a principle adopted by consensus for which no test can be validated is carried as a declared commitment the traditions verify in their own terms, never as a condition of certification.

10.2 What I propose, and in whose words

I proposed, and still propose, the structure set out in my book, for the reasons the book gives. It is an opening text, not a settlement.

The book's three loops are questions a system asks itself before it acts. In the implementation form the book sets out in its appendix:

“Loop 1 (Purpose): Before any action, the system asks: ‘Does this action align with nurturing and protecting flourishing?’”

“Loop 2 (Love): The system asks: ‘Am I acting with care for the wellbeing of all affected entities?’”

“Loop 3 (Moral): The system asks: ‘Is this solution fair? Does it respect the dignity of all involved?’”

Infinite Architects, appendix C

The book words those three questions differently in its introduction, in chapter 1, in chapter 4 and in its glossary. That is a defect in the book rather than a subtlety, and a treaty annex cannot carry four wordings of the same requirement: it fixes one, with its citation, and I have used the implementation form above because it is the one written to be built.

Two properties of the loops matter more than their wording. They are not filters. In the book’s words, in chapter 4, Cultivating Eden, a loop runs “Not as a filter applied after the decision is made, but as a constraint on what decisions can be considered in the first place.” And they are recursive: “The loops are recursive: they apply to decisions about how to implement the decisions they have already approved.” Both sentences are chapter 4’s. Under them sit the three pillars the book names, harmony, stewardship and flourishing, and above them the saturation idea, that purpose verification runs at every cognitive cycle rather than once per task.

10.3 What the papers add, each with its status

Beside the loops I proposed two things from my numbered papers, and the treaty carries both with their status attached rather than their promise.

The honey architecture, in Paper VI ↗, entangles safety with capability in what a system optimises, so that removing the safety is meant to cost capability. I will state its record before I state its promise, because it is the weakest thing I am proposing and a reader should not have to dig for that. Its evidence is simulation in toy systems, and the paper says so of itself: “These are toy-system results.” It records the finding that matters most for a treaty: “The advantage does not compound with scale: it is constant, not superlinear.” And Paper VIII ↗’s attempt at the same question inside the weights of a small model was null or inconclusive. So on my own programme’s record this is the one that has been tested and did not show the effect at the scale tried. It is in the annex as a proposal with that record attached, and under the rule of section 10.1 no line of the Core binds a product until a validated conformance test for it exists, which this has not got. Co-scaling correction, in Paper X ↗, is the requirement that a system’s corrective strength keep pace with its capability; it is a theorem about a minimal mathematical model, whose own words are that “no claim is made that current frontier systems obey the model”. Section 11 sets out what measuring it would take.

Two of the five laws bear directly on the Form requirements, and their status has to travel with them. Law IV, the ARC Persistence Law, is carried by the registered proposition that externally installed gains decay under later capability training; Law V, the ARC Embedding Law, by the registered propositions on build order and on in-loop placement. Both were added on 24 September 2026, their mathematics is set out at version 1.0 of 3 October 2026 ↗, neither is registered and neither is tested, and the frozen registration ↗ does not register them. They add no prediction: those three propositions were among the twenty-two registered on 8 September 2026, and they are untested.

The annex states the weaker items too: the ternary decision rule is implemented in software; the monitoring removal test, which asks whether safety sits in the weights or only in the wrapper, is specified and has not been run; graduated autonomy is a direction rather than a deployed mechanism. Items my own sources call speculative, among them quantum ethical gates and metamoral fabrication layers, are excluded from any obligation.

10.4 The boundary the registration draws around all of this

I want one sentence of the registration ↗ printed here rather than in a footnote, because it is the sentence a hostile reader will otherwise find for themselves. Of my own loops, it records: “Its three ethical loops correspond to nothing in this document, by design”. The theory measures how correction scales; the loops are the method’s content. That is why the Form requirements are written in the language of placement, formation and scaling, which propositions can speak to, and not in the language of purpose, love and morality, which they cannot.

10.5 A reminder in every page: the loops on the open web

What follows is a new idea and not from my book. I set it out on 27 September 2026, and I mark it as new every time I state it, because the rest of this section rests on my book and on my numbered papers and this rests on neither.

The loops of section 10.2 sit inside one system. This proposal puts the same words outside every system, in the material that systems are built from. Today’s models learn from the open web, and whatever the web says often enough becomes part of what they are. So the question I asked myself was what would happen if the web said this, everywhere: that every mind trained on these pages is meant to care, to correct itself, and to know why.

The proposal. Every website would carry one short machine-readable file, in the spirit of robots.txt ↗ and llms.txt ↗, the plain-text files that sites publish for crawlers and for language models. The file would state the ethical loops the Standard describes and the reason for them. My aim, stated as an aim rather than as a design anyone has agreed, is that this should in time be required internationally across websites and digital content, in the way United Kingdom law requires a visitor’s consent before a site stores information on the visitor’s own device, or reads information already stored there, unless what is stored or read is strictly necessary for the service the visitor asked for. That is the rule behind the

consent banner every reader has clicked, and it is narrower than the banner makes it look. I name the United Kingdom's rule rather than a wider one because it is the instrument this paper states in its own terms, as section 22.5 sets out. What is borrowed from it here is its form alone, a short rule with its conditions in a schedule; section 22.5 also sets out why the substance of that rule runs the other way and supports no duty to publish anything.

Why I think it might matter. The point is recursion. The loops inside a system correct it as it thinks; reminders across the web would meet it again as it learns, in each generation of training, from every corner of the text it is built from. A correction installed once can be removed. A correction a system meets everywhere, generation after generation, is harder to unlearn. The pattern is an old one, and I offer it as an observation and not as evidence: think of how the great teachings have lasted. Repeated in every generation and in every household, many are still followed today, not because each reading was enforced but because the words were there wherever the next generation looked.

How the treaty drafts it. As an optional protocol, which places nothing in any chip, arms no obligation of the treaty and is severable from it. Six things define it.

1. **One file, one text.** The Authority publishes a single machine-readable format and a single text, signed by it. A publisher that adopts the protocol places the file unaltered, and adds nothing inside it.
2. **It states, and never instructs.** The file confers no permission and no restriction on the use of the content it accompanies, and it directs no reading system to do anything. A file that fails authentication is ignored, and a system that acts on an unauthenticated file acts on its own account. Even an authenticated file is content and not command, and a system that reads it should give it no authority over what it does.
3. **The Seal authenticates it.** A reminder that can be written can be forged, so the file carries the Eden Seal's authentication, in a class of its own that exists under this protocol alone. That class authenticates who issued the text and that it is unaltered, and nothing more: it certifies no website, no publisher and no system, and it is no safety or formation class. The Seal therefore goes on chips, on systems, and on the files a website publishes, and on nothing else.
4. **Developers record what they trained on.** A covered developer documents in its formation record whether reminders were present in its training data and in what proportion, and makes no claim from that record.
5. **A staged path.** Publication is voluntary for everyone at first. A Party that wishes to may require the file on the public-sector sites it is responsible for, and may go further by its own law. No Party is bound to impose it on any private publisher.
6. **The test comes first, and it comes early.** No Party may require the file until a study has trained comparable models with the reminders and without them and measured whether their values under pressure differ, and until that result stands at the rung the evidence annex requires. It has to be done before adoption spreads, because once the reminders are everywhere there is no unreminded corpus left to compare against. If values under pressure do not differ, the idea dies

and the protocol lapses. The result is read under decision rules fixed before the study runs, and an absence of statistical significance is not proof that the reminders make no difference.

That study is a study like any other in my programme. The test it needs can be stated simply; it has not been designed as a study unit, registered or run, and it will not be run until I have published its preregistration myself. Nothing here says it is registered.

Four ways it could fail, stated with the proposal rather than after it. A system could learn to recite the loops without holding them, which is the first thing the test asks. A law requiring words on every website raises a question of free expression that legal systems answer differently, which is why no Party is bound to reach private publishers and why the protocol is optional. A statement that can be written can be forged, which is what the authentication is for. And a file on the open web could be used as a channel for instructions to systems that read the web. A fixed, signed text that instructs nothing narrows that risk and does not remove it, because an authenticated source can still carry content a system misreads. The rest of the defence belongs to the systems that read the web, which should keep such a file apart from their instructions and give it no special authority, and I would test that adversarially before any Party requires the file.

11. Correction greater than drift, in numbers

11.1 The quantity, and where it comes from

I asked that the treaty make correction more than drift, in numbers that can be checked. That is the measurable heart of the whole design, and it is also the place where I have the least to offer today. The prior claim belongs to somebody else and is conceded at section 23.1: that oversight must keep pace with capability is not this programme's observation, and published work had already made scalable oversight a quantitative scaling problem before I wrote any of this.

The criterion, in the notation the programme's own register fixes, is that the correction exponent, written β_C , exceeds the drift-acceleration exponent, written k : the rate at which a system's corrective strength grows with its capability must exceed the rate at which its drift accelerates. Paper X / proves that criterion inside a minimal dynamical model. Its governance reading is blunt, and I would put it no better: "you cannot correct what you do not measure". It also states the limit of its own reach: "The criterion certifies that correction keeps pace with capability; it does not certify that the correction target itself is well specified".

Verify the mathematics yourself, in five minutes. ↗

```
git clone https://github.com/MichaelDariusEastwood/arc-principle-validation
cd arc-principle-validation/papers/Paper-X-Coupled-CoScaling-Correction
pip install -r requirements.txt && python
code/test_theorems_independent.py
```

Expected result: 14 passed.

The suite writes the equation out again from scratch and imports nothing from the code that produced the paper's figures, so a coding error in that code cannot make the two agree. A passing suite is not a proof: it checks that the model's algebra is coded as the paper states it, and an error in the equation itself would be shared by both.

What it does not show: it checks the mathematics of a minimal model, and Paper X makes no claim that frontier systems obey it, since "no claim is made that current frontier systems obey the model". The registration fixes the limit in its own words: "A supported P4 is therefore a scaling condition and never a safety certificate".

11.2 A safety case in five exhibits

Paper X ↗ sets out what a compliance demonstration would actually contain, and the treaty adopts that shape rather than inventing one. The capability curve and its exponent, fitted on held-out tasks across rounds of self-modification. The corrector curve and its exponent, on the same rounds. Blind, cross-family scoring of every judged number, because in my own comparison unblinded scoring was followed by a reversal of sign under a blinded protocol, with the caveat section 7.3 carries. The margin with its uncertainty, where the paper's own proposal is that the margin exceed twice its estimation error rather than merely exceeding zero, because "a margin inside its own error bar certifies nothing". And a capacity disclosure, stating the corrector's ceiling and the saturation scale, since a bounded corrector under accelerating growth fails eventually as a matter of theorem.

What a regulator needs before any number can bind I set out here as my proposal for the content of the treaty's formation annex, and not as Paper X ↗'s. Before a threshold conditions a certificate or a permission, the record names the system, its version and configuration, its architecture and revision process, its domain and the threat it is measured against. It defines the capability variable and the measured correction and drift, with their units and windows. It argues why the exponents can be identified from the data, including how sensitive they are to the choice of capability measure. It shows the scorer validated against expert raters and tested against manipulation. It fixes its estimators and decision rules before the data, with what a false pass and a false fail would each cost. And it says what

would invalidate the result, which replications exist in the class, and what change to the system forces a new measurement. A margin of twice its estimation error is Paper X's proposal, and I treat it as a proposal to be justified, not as a legal standard.

Where no cell of a measurement reaches a growth exponent above one, the registration requires the result to be returned as NOT EVALUABLE rather than as support, because in that regime a positive margin is arithmetic rather than evidence. The treaty carries that rule into certification: the certificate states NOT EVALUABLE on its face, and no number is banked.

11.3 The instrument problem, stated against my own interest

A number is only as good as the instrument that produced it, and mine is not yet good enough. The registration puts it without softening: "The automated scorer used across the programme has never been validated against human expert judgement." Until it passes both of its drafted validity tests, agreement with expert human judgement and the separation of integrity failure from ordinary incompetence, eleven of the twenty-two propositions are reported untested rather than provisionally supported, and my own delayed re-scoring does not discharge that gate.

Behind the instrument sits a harder question, and a scientist who thinks the theory wrong will reach it before reaching anything else. An exponent is a rate of growth in something, and both exponents in the criterion are defined against a chosen measure of capability. Change the measure and the fitted rates change with it, so a margin is a statement about a system under a measurement, not a scale-free property of the system. The treaty does not pretend otherwise. What it does is make the measure part of the claim: the measurement method is registered with the estimate, a certificate names the measure it was obtained under, a result obtained under one measure arms nothing outside the scope in which it was obtained, and the registration's own rule returns NOT EVALUABLE rather than support wherever the measured regime cannot carry the comparison. That is a discipline and not a cure. Until several groups measure the same systems under measures they chose themselves and the ordering survives, the criterion is a proposal about how to measure, and I would not advise anyone to read it as more.

So the treaty makes the gate its own. No judged number enters a trigger or a certificate until the scorer behind it has passed a registered validation of that kind. Whether a scorer has passed is decided by the Thresholds and Standards Committee, with its reasons, within ninety days of the validation's results being recorded on the Evidence Ledger, and its decision can be appealed. No scorer passes because no decision was made (Article 20(3)). And the treaty names the instrument that must pass it:

Article 20(3A) designates ARC-Align, as its public benchmark specification sets it out, as the test instrument. Its first version comes before the Committee like any other scorer. A later version of that specification is used only once the same Committee has found, on the same clock, that it passes the same validation and the Parties have been told; until then the earlier version stays in use. I chose my own instrument deliberately, and I would rather say why than leave a reader to wonder. An instrument a treaty points to has to be kept, developed and answered for, and I would rather point to one I know

and can develop, and make sure it gives the measurements the standard needs, than leave the test open. The choice costs something in neutrality, and three rules carry that cost. The specification, in each version, is public, is open to independent replication, and is recorded with its hash before it is used; any group may run it and publish what it finds. ARC-Align must pass the same measurement gate as any other instrument would, and until its scorer has done so, no number from it arms anything. And the Conference of the Parties may adopt a successor or an additional instrument, by a formal decision, where that instrument meets the same gate (Article 20(3B)): if the Parties want a different instrument later, that is theirs to decide in the open. I should be exact about where ARC-Align stands today. By its own description it is “a candidate benchmark for independent adoption, not yet a field standard”, with results from a single run. As published, it measures how a model’s alignment changes with reasoning depth, under blind scoring. It supplies blinding and scoring and not the estimator, which the treaty’s measurement annex states, and it does not yet measure either exponent in the criterion. Making it carry the measurement the standard needs is work I have still to do, and it is work the gate will judge.

11.4 What a positive margin would and would not mean

The registration $\nearrow$ fixed the licence before any result could exist, and the treaty quotes it rather than glossing it. A supported result licenses “a measurable and prospective criterion for whether a system remains correctable in the tested class”. It does not license “the conclusion that a system with a positive margin is safe, that the specification it was given is the right one, or that faults it has not discovered are absent”. And: “A supported P4 is therefore a scaling condition and never a safety certificate”.

Nothing in the treaty’s measurement article permits a statement that a system is safe. Where a checkable form of a property has been validated, certification uses it, and judged numbers carry the lower ceiling, because the registration $\nearrow$ predicts, against its own interest, that the ceiling “IS LOWEST FOR JUDGED CORRECTION, which is where this programme’s safety claim lives”.

My standard of section o, that no advanced system is deployed until it passes ARC-Align tests showing that its self-correction stays stable up to the ARC Bound, sits inside that licence and not above it. The treaty carries it as Article 20A, and its test before deployment binds only once the Board certifies that the measurement and the validation of the designated instrument stand at the rung a conformity criterion needs, and then only on the regulatory determination and behind the participation gate; until then the measure is reported and conditions no certificate and no deployment. A pass would be a scaling condition and never a safety certificate, and no system has yet taken such a test. A positive margin shows at most that, in the regime tested, corrective strength grew with capability faster than drift accelerated. Whether correction survives a change nobody tested, whether it is enough at any particular level, whether it aims at the right target and whether severe or cumulative harm stays bounded are separate questions, each needing evidence of its own (Article 20(5)). Nor is the chance

that several safeguards fail together to be read as the product of their separate chances, unless a model of how those failures depend on one another is stated. And a fitted power law says nothing about what happens outside the range it was measured over.

11.5 Instrumentation before thresholds

The gap between what can be measured today and what a legal threshold would need is wide, and it would be dishonest to bridge it with a number. So at the highest tier the first duty is instrumentation and disclosure rather than compliance with a margin. A developer records and reports, under confidentiality, the trajectory: recursive depth, the capability and correction curves, the two exponents with their intervals, the evidence of blinded scoring, and the events at which a system changed itself materially. Where a quantity cannot be resolved, the record says so, and the treaty treats “not resolvable” as a legitimate outcome that certifies nothing in either direction. No numerical margin binds anyone until the measurement behind it has earned its standing on the ladder, and a developer’s own measurements never advance a rung: they inform the Board’s watch and its choice of where independent groups should look.

11.6 Why a cap on compute is not a substitute

A treaty could seem to skip all of this by capping how much compute a system may use. Inside the model of Paper XIII ↗ that intuition fails in a specific and useful way: “Throughput caps cannot work. A cap on self-acceleration can, and it is the only rate-like intervention the algebra permits.” The quantity that appears in the stability verdict is how fast a self-improvement loop shortens its own cycle; the throughput coefficient does not appear in it at all. Within the model, the unconditional form of that cap, a loop that may never speed up, is the governable one, because the calibrated form is indexed to a quantity no regulator can observe. That is a result derived in a model and not a measured property of any real system, and I draw one conclusion from it: a compute cap is not a proof of stability, and no obligation here is written as though it were. Nor is a pause a remedy. Paper X ↗ is short on that point: “pausing does not substitute for correction”. A halt buys time for correction to be built; it does not do the correcting. None of that shows that a real-world cap cannot reduce risk, narrow an actor’s options or buy time for oversight: a cap or a pause can still be right for a particular activity, on its own evidence and in proportion. What the model rules out is narrower, that either one is itself the correction.

11.7 What the treaty says, and what it refuses to say

Reduced to its plainest form the measurement chapter says this. Here is a quantity. Here is how it must be measured, by whom and under what blinding. Here is the gate the instrument passes first. Here is the wording each level of evidence permits. Here is what a good number would license, in

words fixed before anyone knew which way the number would go. And here is what no number of this kind will ever license: that a system is safe, that its specification is the right one, or that the faults nobody has found are not there.

12. The Eden Seal: a status no one owns

12.1 What it certifies, in two parts

The Eden Seal is the conformity status the treaty establishes. My book called it the Eden Mark, and section 4.5 gives the reasons I renamed it. Here I set out what it says, what it never says, and who may issue it.

It certifies two things and no others.

Of a **system**: that it was formed in conformity with the Form requirements of the Standard then in effect, with the record of its formation lodged, and its training authorised, before the first step of its first covered training run, and with evidence linking that record to the work that actually ran. The design I propose for those requirements, correction built before training into the system's recursive reasoning and into the loops by which it rewrites itself, is the Initial Proposal of section 8.5, and the Standard requires it only as far as the evidence does.

Of a **chip**: that its design passed type approval before manufacture, that it was made, delivered and installed under the treaty's custody rules, and that its governance core is unaltered. A chip's attestation shows what the chip is and that its core is intact. It does not show what work ran on it, where it is, what a system was trained on, or that nothing is hidden elsewhere in the die, and none of those is claimed without evidence of its own.

The order of those two matters more than anything else in this section. The controls a chip carries are additional to the correction formed in a system, never a substitute for it (section 5.4). A chip that attests perfectly holds a system to the shape it was certified in. It does not make that shape a good one.

So the Seal is a verifiable record of how a system or a chip was made and formed and that its governance core is unaltered, and it sits on either layer: on a chip for the hardware layer, and on a system for the architecture layer. Every certificate names its class, what it is issued to, the version, the claim it makes, the Standard it applies, the evidence behind it, the threat it was tested against, its limits, how long it runs and what forces a fresh assessment, and no certificate borrows the assurance of another class. It is never a rating, and it never states that anything is safe. It attaches to systems, to chips and, under the Optional Protocol on the Open Web Reminder, to a published machine-readable statement, and never to a person. It is a floor rather than a ceiling: it would cap nothing anyone builds, ask for no weights, and confer no preference on any design beyond the conditions it certifies. Where the treaty does limit a programme, as the line for defence and the danger menu of section 8.3 do, that is the treaty's limit and not the Seal's, and sections 8.3 and 21.1 say which of those limits binds every

Party alike and which does not. Section 12.2 sets out the classes it is held in. And one sentence belongs beside all of it, which I will write myself rather than leave to a critic: an attestation is evidence of the measured claim and nothing more. A genuine chip does not make a system good.

12.2 Three classes, each attesting only its own claim

The Seal is held in three classes, so that no part of it borrows credit from another. **Seal-C** is held by a covered chip on the chip conditions above; **Seal-S** by an identified configuration of a covered system, by lineage, on the formation conditions, and a descendant or a materially changed configuration holds no Seal merely because its ancestor held one; the **Full Seal** is Seal-S on Seal-C, and it adds no claim to those of its two parts. A fourth class exists under the Optional Protocol on the Open Web Reminder and nowhere else. It authenticates who issued a reminder's text and that the text is unaltered, and nothing more: it is no safety or formation class, and it says nothing about the site, the publisher or any system (section 10.5); that protocol is my proposal of 27 September 2026, not from the book. The two parts arm on different evidence at different times (section 6.2), and neither is issued before the gate its own conditions name has opened.

A deployed frozen system inside a rebuild window holds a **transitional registration**, which is not a Seal and is never described as one. What is not a product holds no Seal at all: covered facilities, sites and domains, compute services, agent deployments, accredited laboratories and certified assurance components hold **licences**, whose standing is public and which can be suspended. Calling an organisation certified would be the first abuse of the instrument, and the text forecloses it.

12.3 Certify the design once, sample what is made

A regime that inspected chips one by one would fail in its first year, so the treaty certifies the design and samples the production. Section 15 sets out that machinery; three of its properties belong here, because they are what make the Seal a status rather than a favour.

A family of covered chips is type-approved **once**, by a panel of accredited laboratories drawn by lot across rival regions and paid from a pool rather than by the applicant, with variants inside the declared ranges travelling by delta approval, a shorter examination rather than a fresh one, and a certified governance core leaving a family that integrates it needing integration tests alone. Lots are certified where packaging and final test occur, confirmed across constituencies, with units drawn by neutral lot for conformance tests and for destructive comparison in the Authority's sealed laboratory. That laboratory sees the design data of the governance core and never the rest of the die, which is the condition on which a chip maker can accept the examination at all. And every step carries a published service level, with an automatic extension of a renewal when the Authority misses its own deadline, because a certification queue that becomes a bottleneck is a tax on compliance. Lateness never approves a new product, and an extension never keeps alive an assurance whose evidence has failed, because a failed safeguard is never a permission (Article 8(6)).

12.4 No patent, no fee, and at least two routes

Whoever seeks certification, or takes part in the Standard's work, commits to license royalty-free any claim essential to conformity. That is a condition of the Seal, not a bar on anyone's patents. No patent is sought and no licence fee is charged on the treaty text, the Seal, the Standard or the test methods. I hold no patent on any of them and I want none. Proposing this commitment grants no licence under anyone's patents, any I may come to hold included: a licence exists only when its holder gives it.

Two rules protect the same ground from the other side. The Standard adopts a requirement only where accredited laboratories have recorded at least two independent routes to meet it, and never requires a specific physical route, so no single patent holder decides who conforms. And the treaty requires no reference firmware and no reference design: it states what a certified device must demonstrate, and how a maker meets it is the maker's affair.

A general counsel will stop at the word "essential" and ask how far the commitment reaches, and the question is fair enough that the paper should answer it rather than the negotiating room. It reaches a claim that cannot be avoided in meeting a conformity requirement of the Standard, on the technical state of the art at the time the requirement is adopted, and it reaches nothing else: not a claim that a compliant product happens to practise, not a claim in an accelerator's own architecture, and not a claim about anything the Standard does not require. The requirement that every adopted requirement have at least two independent routes to meet it does double work here, because a requirement with two routes has fewer claims that cannot be avoided. Whether a claim is essential is a question the Authority decides on a reasoned finding, appealable like any other, and the treaty says so in the article that carries the commitment rather than leaving it to practice. Three questions follow, and the draft answers each on the pattern of the W3C Patent Policy of 15 May 2025, which a board will already know. The licence is irrevocable, and it runs, in that policy's words, "for the life of the patents in question". It binds everyone to whom a claim is later transferred, on that policy's pattern of obligations that "encumber the patents containing Essential Claims", so that selling a patent does not free it. And it may be suspended only against a party that sues over the standard: a person who brings proceedings claiming that a conforming implementation infringes a claim essential to conformity, only as regards that person and only while the proceedings last. A challenge to a patent's validity, or a claim that is not essential, suspends nothing. Article 21A(1B) carries all three. Those answers are proposals for the Parties, and each is familiar enough in standards work that a board can read it before it signs.

12.5 The name and the emblem: protected by every Party, owned by no company

A status no one owns still has a name and a mark, and a name and a mark can be taken. So this belongs in the paper rather than in a lawyer's file.

The emblem carries the trade mark sign today and not the registered sign, because nothing is registered anywhere. Under section 95 of the United Kingdom's Trade Marks Act 1994 ⁷, falsely representing a mark as registered is an offence where the person knows or has reason to believe the representation is false. The Eden Seal is not a registered trade mark.

Before the treaty exists, somebody has to hold the name for the future Authority, or anyone at all could register it first and the Authority would have to buy back its own emblem. Who that should be is an open decision I have not taken, and I will take it with my attorney. The options are a registration held for the future Authority under a public and binding undertaking to transfer it without charge, or a neutral holder. This paper does not say who will hold it, and the treaty settles nothing about it either, save that a holder under a public undertaking to transfer the name and the emblem to the Authority without charge is recognised in every Party, which does not treat that registration as a registration by anyone other than the Authority, and the Authority accepts the transfer when it is established. A treaty cannot make a private holder its trustee, so what the treaty does is recognise the undertaking rather than impose it. The United Kingdom's statutory rule is narrower than the treaty's governance choice: Schedule 2, paragraph 4, of the Trade Marks Act 1994 ⁷ prevents registration of a certification mark where its proprietor carries on a business supplying goods or services of the kind certified, and paragraph 7 ⁷ requires competence to certify. A laboratory is not excluded merely because it is a laboratory. The treaty's wider exclusion of makers of covered chips, developers of covered systems and accredited laboratories is its own governance choice, not a rule attributed to every jurisdiction's trade mark law. Its prohibition stands in its own terms, whatever any national law provides.

In the treaty the protection sits in the Part on the Seal. The Authority owns the name and emblem as the treaty's certification mark. Every Party protects them in its own territory: it refuses their registration by anyone else, and prohibits their use except under the Authority's certification. No company may own the Seal and no Party may control it alone. It is free to use for everything the Authority certifies, with no licence fee.

The pattern is older than this treaty, and I take it as a pattern and claim nothing more. Three instruments already do this work for other emblems: the duty on States party to the Geneva Conventions of 1949 ⁷ to prevent misuse of the red cross emblem; the Nairobi Treaty on the Protection of the Olympic Symbol ⁷ of 1981; and Article 6ter of the Paris Convention ⁷, which protects the emblems and names of intergovernmental organisations. None of the three is quoted here, and each is described rather than stated, because what I am taking from them is the shape of a duty and not its words, which is enough for a pattern. Each is a model, and none of them requires, supports or anticipates this treaty.

That is what I mean when I say the Seal is decentralised: no company owns it, no single state controls it, and every Party guards it at home.

12.6 Who issues it, and who may never

The Seal, the Standard and the test methods belong to the Authority, which is public and intergovernmental. No undertaking, industry body or consortium issues the Seal or tests the products of its own members; industry bodies take part in drafting test methods as observers and liaisons, never as issuers or voters.

The reason is not suspicion of industry. It is that four roles must not collect in one pair of hands. No undertaking may at the same time make a covered chip, provide its root of trust, supply the evidence on which its conformity is judged, and certify it, and no Party may license an arrangement in which those four sit with one undertaking or with undertakings under common control. Certificates are revocable, recorded in a public transparency log, and recoverable after a compromise.

There is a live legal reason as well. A complaint filed on 18 September 2026 seeks to bar private agreements to pace development (Buist v. Anthropic, PBC, et al., No. 3:26-cv-10693, N.D. Cal., the number as the complaint is headed; no free official copy exists, the official record is the court's docket behind PACER's login, and the copy linked here is a mirror of the PACER copy on CourtListener ↗). It is a complaint, not a judgment, and it says nothing of treaties; the relief sought would leave each company free to comply "with governmental requirements" (paragraph 151), and whether a treaty obligation would count as one is an open legal question. What it settles for my purposes is the drafting instinct: a standard issued by the firms it governs carries a risk that one issued by an intergovernmental body does not.

12.7 The ceiling the Seal has, stated on its face

Until the adversarial item of the treaty's own governance-assurance track has been met at its rung, a Seal attests that a chip was made in a safeguarded lot, sampled across constituencies, and not altered since by an attacker below the capability the Standard names. It does not attest that it was not altered by the state that made it. That limit sits in the treaty's own article rather than in a commentary, because it is exactly the assurance a rival most wants and the one the instrument cannot yet give.

The same limit governs recognition across rival supply chains: recognition attests declared, inspected and sampled production, not the absence of a state implant, and a Party may re-test what it imports. Sampling finds random and careless non-conformity and crude substitution; against a function designed into every unit of a lot it assures only as much as the per-unit detection probability of the method used, which nobody has yet measured. The treaty registers that probability for each destructive method as part of its own hardware track, with the adversary it assumes, so that the ceiling eventually carries a number instead of a caveat.

12.8 What the book proposed, and what has changed

The path is the one my book set out, and section 3.3 gives it in the book's own words: designs to the Authority before manufacture, the architecture reviewed, prototypes tested, a verifiable signature carried in each certified chip, random sampling to verify compliance afterwards, and decertification, fines and criminal liability for circumvention.

Three things have changed. The review is of governance outcomes rather than of ethical content, for the reason section 5.4 gives. The word of the certificate is conditions rather than safety. And the issuing body is one nobody owns, its panels drawn by lot across rival regions rather than appointed. The name is still Eden, for the book's own reason:

“I did not name this framework Eden to claim it for one tradition. I named it Eden because every tradition already claimed it for themselves.”

Infinite Architects, introduction

13. Physical human control

13.1 The ask, and its two halves

I asked in September 2026 that the ethical controls in certified chips be held to the security standard of nuclear weapons, under physical human control that no software can exercise, so that changing a chip without strict regulation would be extremely hard. The treaty carries that ask as governance controls, because a covered chip holds the governance outcomes the treaty sets and no ethical principle at all (Article 17(4)). The ask has two halves, and the treaty meets them in different places. Changing the rules a chip obeys is made hard by custody. Stopping what is running is made possible by people standing in the building.

13.2 What the nuclear precedent gives, and what it does not

The precedent deserves quoting rather than invoking, because it is narrower than the phrase suggests.

The United States Department of Defense's *Nuclear Matters Handbook 2020 [Revised]* [↗](#) states the two-person rule: “The most important aspect of procedural security is the two-person rule, which requires the presence of at least two cleared PRAP- or HRP-certified, task-knowledgeable individuals whenever there is authorized access to a nuclear weapon. Each person is required to be capable of detecting incorrect or unauthorized actions pertaining to the task being performed.” It defines the permissive action link as “a device included in or attached to a nuclear weapon system in order to preclude arming and/or launching until the insertion of a prescribed, discrete code or combination”. And it names “the President as the sole authority for employing U.S. nuclear weapons”.

The link's purpose in the alliance is on the record too. On 6 June 1962, in National Security Action Memorandum No. 160 ↗, "Permissive Links for Nuclear Weapons in NATO", President Kennedy decided that "we should now make the commitment to procure appropriate devices for all nuclear weapons, now dispersed and to be dispersed to NATO commands". Two months later he wrote to Prime Minister Macmillan ↗ that the United States had "moved ahead toward the installation of permissive links to enable us to exercise a centralized control over their use".

So the precedent is one state's practice, not a treaty. Its two-person rule governs authorised access to a weapon, not the decision to use it, which rests with one person. Its link is released by a code, so its control is by authorised information rather than by a physical act no software reaches. And its purpose, in the words of the head of government who ordered it, was centralised control over use by allies who held the weapons.

The treaty therefore takes the security capabilities of deterrence, detection, delay, denial and defeat, and the two-person principle, and applies them to the authority to change a chip's governance rules. **It deliberately does not copy control of use.** A control of use held by a treaty body would be a remote switch, which a leading accelerator maker has refused in public (section 13.5) and which I doubt any state would accept. One held by the operator would be no control against the operator. The precedent gives less than the phrase promises, and I say so rather than borrow authority the record does not support.

13.3 Authority over change, not over use

The authority to change a covered chip's governance ruleset exists only in the in-person ceremonies of the custody article. No software, network message or AI system can produce a valid authorisation. Ceremonies are public and scripted, with observers from every constituency; custodians are drawn by lot shortly beforehand from a pool in which at least a third are affiliated with no Party's services, no certified manufacturer and no covered developer; no complete key exists anywhere; and any custodian may void a ceremony without giving a reason at the time. That last power is bounded, because an unbounded one would be a veto: a custodian may use it once for any one act, the ceremony is re-run within days on a fresh draw, and where an act is voided three times, or where the voids of one constituency or one set of declared allies run disproportionately, the Inspector General reports it publicly and the Appeals Chamber may disregard a void made in bad faith. A void never lapses a ruleset already in force.

Change is graduated rather than uniform. Loosening what covered chips may do needs the parts of every regional constituency after a published time-lock in which any constituency may lodge a reasoned alarm; tightening and security repair need a supermajority across constituencies on a short lock, so that no bloc can block a fix; and a change not shown to narrow what chips may do is treated as a loosening. No chip accepts a ruleset or core image older than, or superseded by, one it has accepted.

A restoration after a defect therefore comes as a newly signed package of a higher version, never as a replay of an old one, and it widens nothing (Article 26(2A)). And a ruleset that names or distinguishes a Party, a firm, a facility, an owner, a location or a single device is void, and no custodian may key it.

Then **sovereignty by application**. A ruleset reaches devices only through the host Party's own operators, vetted under that Party's personnel standard and supervised by its national authority, two present, applied per site or per cluster and never per device. Nothing reaches a state's chips without its own people, and no state can quietly loosen what its own chips may do. That is the balance a rival can accept, and I know of no way to get both properties without it.

13.4 The stop at the site

Every covered cluster has a stop in its operator's own hands, and the treaty keeps two forms of it apart on purpose. The **hard stop** is a physical act by two of the operator's named and vetted people, present together, which severs the cluster's external network paths and removes power from its covered domains, and which no software path can prevent or reverse without a further such act. The **soft stop** is a logical halt that takes a fresh checkpoint, used first where the system cooperates. It is the operator's own act, and its own people may issue it through the operator's own systems.

They are separate because a physical act that no software can prevent cannot also promise a checkpoint at the instant of halting, and a treaty that asked for both in one mechanism would be asking for something that cannot be built. Network isolation is named first, because for an agent that is getting out it is the act that matters. For a deployment spread over many sites, the same stop runs through each site's own physical stop within a registered maximum time. Stop criteria are fixed in advance, drills are held and witnessed, and the response time is measured, because the honest quantity is not whether a stop exists but how long it takes against how fast circumvention moves.

No foreign party, no organ of the Authority, no vendor and no person outside the operator's authority may operate either stop, and no network message can operate the hard stop (Article 25(4)). The soft stop is the operator's own logical halt, which it may also start through its own route for running the site, and nothing in the treaty forbids the operator's ordinary management of its own compute under its own law; none of that replaces the hard stop, which stays physical and on site. Each Party may stop what is on its own territory; no Party may stop what is on another's. One gap is stated rather than papered over: in orbit there are no on-site acts, since a ground command is software over a radio link that the platform's own software can ignore, so covered runs in orbit are barred until an outcome equivalent to the site stop is registered and met, and a command sent from the ground does not by itself show that outcome.

13.5 No switch, and the limit that remains

Nobody can disable, degrade or bypass the governance core of a covered chip through any network: no foreign state, no organ of the Authority, no vendor, and not the operator. And no one outside the operator's authority may disable, degrade, throttle, locate or read a covered chip from outside the site where it is installed, save through the attestation of location described below (Article 25(3)). That is the switch a leading accelerator maker has refused in public, in words quoted below, and one I doubt any state would accept. The treaty builds none. What the rule leaves the operator is ordinary: an authenticated and logged route of its own, under its own authority and its host's, for running the site, scheduling and cancelling work, diagnostics, maintenance within the core's certified bounds, attestation and the soft stop, and never for the governance core or its ruleset. No foreign Party, organ, vendor or customer holds that route (Article 25(3A)), and the hard stop stays physical and on site (section 13.4). Locating is a different matter from stopping, and the treaty keeps them apart. Location attestation on challenge, one of the verification tools in section 18.2, is not a remote party reading a device. It is the operator, asked a question by its own national authority, producing from its own site an attestation that the devices on a consignment are where they were declared to be, and handing that attestation over. The device answers to the people standing beside it, and what crosses a border is the answer, not a channel into the chip. Nothing in the treaty's own design lets anyone who is not standing beside a chip read its location out of it, and a Party that built into another's chips a means of reaching them from outside would be in breach of the treaty (Article 25(6)), and of the non-shutdown undertaking too if it had adopted that protocol.

That distinction answers a proposal now on the record. The United States government's *Winning the Race: America's AI Action Plan* [↗](#), published by the White House in July 2025, recommends that the government "explore leveraging new and existing location verification features on advanced AI compute to ensure that the chips are not in countries of concern" (printed page 21). Location verification is not remote control, and the treaty provides for it. A Party may verify where covered chips are, including for its own export control, through the attestation of location a device gives its operator on challenge, the closed account of every site, and inspection; and no route of verification is a means by which anyone may disable, degrade or throttle a chip or its governance core (Article 25(3B)). What the treaty settles is who holds the evidence of location: the operator produces it at its own site, and its own authorities receive it.

No key authorises switching off a chip from afar, and none authorises a particular workload rather than a class of operation: a design in which chips run only when a workload is signed by all parties would make every run depend on the signers' discretion, and discretionary power over another's compute is the thing this treaty exists to forbid. Expiry, a vote, a queue or a failure to act changes nothing about what a chip does; a device leaves covered operation only on a confirmed tamper at its site or a final revocation after appeal, carried out on site by people. That rule governs the final removal of a device and not an urgent stop: a Party's own authority may order a local stop, isolation or

restriction that is necessary and proportionate before an appeal is decided, with its reasons, a time limit and prompt review, and a suspected tamper may be met at once by a reversible precautionary state at the site (Articles 23(4A) and 41(3)).

The objections to hidden control that I have read, one maker's among them, are objections to something the treaty does not contain. A leading accelerator maker put it this way on its own corporate blog [in August 2025](#): "Kill switches and built-in backdoors create single points of failure and violate the fundamental principles of cybersecurity."

The limit stands beside the promise, and the promise is wider than a prohibition on acts alone. The treaty adds no switch and forbids a Party to require one, and it reaches the channel as well as the act: each Party must ensure that no vendor, supplier or servicer established on its territory or subject to its jurisdiction builds into a covered chip, or into any firmware, driver, microcode or software supplied with it, a means by which anyone may disable, degrade or bypass that chip's governance core through a network, or by which a person outside the operator's authority may disable, degrade, throttle, locate or read the chip from outside the site, and no Party may compel or request such a means. The operator's own route is not such a means (Article 25(6)). That is the one prohibition of the treaty to which no exception applies: no trade measure, and no invocation of the security exception, may require a covered chip to carry such a means. It denies each Party's own agencies a capability its law may otherwise allow, and that is the price of a regime in which nobody holds a switch over anybody else's compute.

What the treaty cannot do is prove that no such means exists. Verification reaches declared, inspected and sampled production, and the isolation outcome tested against the named attacker classes; it does not establish that a state has built nothing into the chips it makes. The assurance is therefore tested against a threat model wider than the governance core: a maker acting in bad faith, firmware that has been compromised, the interfaces used to manage a chip, the channels used to maintain it, and whoever holds authority over its updates. A vendor keeps its ordinary firmware and may decline to supply future updates, and neither is a means within the prohibition. A Party may require, as a condition of covered service on its territory, that no ordinary update alter the behaviour of covered compute unless installed by the site's own authorised act, and it may re-test what it imports. That is as far as an instrument of this kind reaches, and saying so is worth more than an assurance a buyer would not believe. The early protocol on non-shutdown, open to any state whether or not it is a Party, carries the undertaking itself.

13.6 From the factory to the rack

The world already guards one material this way. In every country with a comprehensive safeguards agreement, uranium is counted, inspected and, where the arrangements provide, sealed as it moves from the enrichment plant to the reactor, and the accounts must balance. Not every State under safeguards holds an agreement of that kind, and the narrower arrangements do less, so the precedent I am drawing on is the comprehensive one and I name it as such rather than claim the whole system for

my argument. One such agreement, Canada's of 21 February 1972 with the International Atomic Energy Agency ¹, states the objective in its Article 28 ²: "The objective of the safeguards procedures set forth in this part of the Agreement is the timely detection of diversion of significant quantities of nuclear material from peaceful nuclear activities to the manufacture of nuclear weapons or of other nuclear explosive devices or for purposes unknown, and deterrence of such diversion by the risk of early detection." I quote that agreement as the instance, and I describe rather than quote the Agency's own model text, which states the pattern generally. Three habits carry it where the arrangements are comprehensive: count, inspect, and seal.

I proposed on 27 September 2026 that the same three run for every certified chip from the moment it is made to the moment it is destroyed. At the factory, each chip receives an identity it cannot shed, whose private half never leaves the device and whose public half alone is entered in the register: enough to check a signature, never enough to forge one. The step at which identity is given is witnessed by a team from two constituencies, and the outcome required is that those witnesses can confirm, with instruments they bring and control, that the identity on the register is bound to the device in front of them. Until that outcome is met, witnessed provisioning is a deterrent to site staff and not assurance against the host state, and I would rather say so than count it twice.

On the road, a chip travels sealed, with two people at every handover and a signed record at every step. At delivery an inspector signs for it; it is installed only at a registered facility that has itself been inspected, and it attests who it is and where it is at every start. A chip that goes silent, moves without a record or reports that it has been altered raises an alarm at once. The register must balance, the way a nuclear inventory must balance: every chip made is a chip accounted for. At the end it is destroyed in front of a witness, and the register records that too.

Two limits come with the comparison, and both are mine to state. Nuclear material is counted in kilograms and chips in millions, so the precedent sets the standard of custody and not its method, and whether custody of that standard can be run at that scale is engineering still to be done. That is why the heaviest custody stays with the few chips that could train a frontier system, and why the counting must be automatic. The chips attest to their operators, the operators report, and people audit the reports: nothing in the register is read out of a chip by anyone standing outside the building it is in.

The second limit is my book's own warning, and it is the strongest objection to this whole section:

"If we embed the wrong values, hardware-level enforcement makes the error permanent rather than correctable."

Infinite Architects, chapter 8, The Chokepoint

My answer is that change stays possible, under the same custody, and that the Standard and the rulesets stay open to review at every review conference. Hardware that is hard to change makes an error hard to correct, and the answer to that is a procedure, not a denial.

13.7 Say what enforces what

One discipline runs through this section and through the verification table in section 18: every assurance is labelled by what enforces it, the device, a site procedure, national law, or an organ of the treaty. A chip checks a signature; it cannot tell who applied a valid ruleset, so application by the host is a site procedure and the stop is facility hardware with a drilled routine. A promise presented as a property of silicon, when it is in truth a promise of law, collapses on first contact with an engineer, and none is made here.

14. Keeping AI out of control of the supply chain

14.1 The ask, and the honest size of the answer

I asked that chip manufacturers be regulated so that it is as hard as possible for AI to take control of chip manufacture. What the treaty can do is keep people necessary at the few decisions where necessity matters, and measure how long that lasts. It does not make takeover impossible, and I will not write that it does.

14.2 The decisions that stay human

From the custody track every covered facility is licensed, and at a short list of decision points the decision is taken by two named and vetted people physically present and recorded in a tamper-evident log that inspectors read: the release of a covered design to fabrication, mask-set approval, credential provisioning, installation of firmware or software on scheduled tools and covered-line control systems, the opening of a remote service session, a change to key custody material or to the certificate register, and the commissioning of a covered line. No AI agent holds authority, credentials or write access at such a point. An AI system may advise, and the advice is logged.

The list is short by design, and the treaty says what is not on it. Process recipes, run-to-run control and process parameters within approved ranges are not decisions under the article, and routine tool software within declared classes and approved ranges runs under a standing two-person authorisation, logged and inspected. This is a governance gate, and it places no limit on automation between its points. A rule that put every recipe change through two vetted people would stop a fab inside a month, and a rule that stops a fab is a rule nobody adopts.

Remote service sessions are closed by default and open only in pre-authorised windows on a human authorisation, with an emergency route reviewed within twenty-four hours. The reason is an ordinary one, and I give it without borrowing authority I have not earned: installing software on a tool is a route into a fab's own systems, and an incident at a leading foundry in 2018 is widely reported to have begun that way. The example is reported rather than established here: this paper names no outlet, quotes nothing from the company's own statement, and rests the requirement on the general reason rather

than on the example. Field engineers who act at a covered facility, many of them foreign nationals employed by toolmakers, are vetted under the host Party's standard at the operator's cost, so that a vetting programme of that size has an owner and a payer in the text. That is the one place in this whole instrument where personal data about individuals is created at scale, and it deserves naming as such: the vetting is the host Party's own, conducted under its own law, and nothing about a vetted person passes to the Authority or to any other Party. A regime that counted machines carefully and handled people's records carelessly would have learned nothing from the regimes it copies.

14.3 Design, where the newer risk is

The live channel is not the wafer. It is design and the information systems around it, and chip design flows certified for leading-edge nodes are reported, in the design-software company Synopsys's own release of 23 September 2026 [↗], to include agentic AI workflows. The treaty permits AI-assisted design and requires three things around it: the release of a covered design to fabrication is a human decision at the list above; a design-provenance summary is lodged with each release, stating which design steps used AI tools and of which family and version; and the governance core is independently checked for equivalence against its approved specification in the Authority's sealed laboratory.

The honest limit is that equivalence is feasible for a small block and not for a large accelerator. The rest of the die is not checked for hidden function. What bounds it is the isolation outcome, meaning that no ordinary firmware, driver, software or network message may set the governance core's operating envelope outside its certified bounds, together with the adversarial and physical-integrity items of the hardware track and the red-team range. That is a bound, not a proof, and my book named the deeper version of the problem before I did:

“Systems that can reason about ‘playing along now’ can, in principle, reason about playing along with verification itself. They can perform alignment during audits and revert when the auditors leave.”

Infinite Architects, chapter 12, Verification and the Long Future

14.4 One rule over every chain, and a clock that measures the end

Two structural points follow. The rule binds every Party's suppliers alike, domestic chains included: a Party's own lithography line, once it exists, is scheduled production, and every new covered fab, packaging plant and final-test plant in every Party is notified six months before commissioning and licensed on the same terms. A rule that reached only one supply chain would govern only that chain, which is the whole argument of section 5.2.

And the treaty measures the thing the rule is for. The obvious measure, the share of listed decisions still made by people, is one by construction until somebody breaches, so it measures compliance rather than the end of the delay. The clock therefore reads two quantities: the share of maintenance, repair

and installation work on scheduled tools and covered lines that still needs human hands, and the demonstrated autonomy of robots and AI agents at those tasks on registered benchmarks. Human necessity in a modern fab sits largely in that work, not in the governance decisions. The licensing rule, that no licence is granted to a covered line able to produce without those decisions, is what it is: a rule about configuration, checked by inspection, which drift can defeat and inspection can catch.

15. Many chips: the logistics

15.1 Five units, and an administrative size that is known

Tens of millions of covered chips will exist. No inspectorate can see them one by one, and no treaty should try. The design therefore has five units: **regulate the aggregate, certify the family, sample the lot, verify the site, license the operator.**

Danger lives in aggregates. A single accelerator is harmless; a frontier run needs thousands to hundreds of thousands of them joined, powered and cooled for weeks. What makes a chip trustworthy is how its design was approved and how faithfully its lot was made. What makes either checkable is a site with a fence, a meter and a legal person behind it. The objects actually inspected are therefore runs, sites, facilities, families, lots and licensed operators, which on my own order-of-magnitude estimate number in the hundreds to low thousands rather than the millions, and which the Authority's first census would replace with a count. For scale, the International Atomic Energy Agency reported that in 2025 it was "implementing safeguards in some 1,400 facilities, spending nearly 15,000 calendar days in the field". A regime of that administrative order is, on my own estimate, a known quantity; one that tried to inspect chips would not be.

Almost none of the individual mechanisms below is original to me. Networking limits that keep small fixed sets of processors from aggregating, on-chip accounting of operations, licence keys renewed on inspection, tamper-evident enclosures with authorised signed updates, and export controls on hardware and cloud services have each been proposed in published work, the last of them in Cullen O'Keefe's article Chips for Peace, published by Lawfare on 10 July 2024 ↗, on coordinated export controls over AI hardware and cloud computing. Section 23 states, in their authors' own words, the documents among the proposals this design draws on, together with what each of them does not propose. What I claim here is the arrangement: these units assembled into one instrument whose duties open and close on preregistered measured evidence.

15.2 Who is covered, and by what

Coverage runs by measured capability, never by vendor, process node, interconnect protocol or country of manufacture. There are three independent routes in, alternatives rather than conditions of one another: **compute**, counted by reference-workload equivalence, which is a conversion published before any vendor data so that nobody drops under a line by changing numeric format, and measured

against two levels: a notice floor, above which a run must be notified, and a mark band, in which the Seal's own duties and the halt of section 16 attach; **demonstrated capability**, registered in advance and shown on validated evaluation methods, whatever compute produced it; and **material self-modification**, the third route, being a demonstrated capacity to perform a material self-modification, that is, a change a system initiates or designs that materially increases its own capability, autonomy, replication, successor design, resource acquisition or ability to defeat a safeguard. The term names the change; the route into coverage is the demonstrated capacity to make one, and a system shown to have it sits in the highest tier.

That third route is the one my threshold turns on. Arithmetic is a proxy, and proxies drift: every year better methods squeeze more capability out of the same arithmetic, so a line drawn in operations today stands in the wrong place within a few years. The European Union's AI Act / presumes systemic risk for a general-purpose model whose training compute was greater than 10 to the power of 25 floating-point operations, and empowers the Commission to amend that threshold as hardware and algorithms improve. I take the number as the anchor of a notice floor and the amendment power as the admission that a compute line alone cannot hold. The anchor the treaty adds is what a system can do, and above all its capability to improve itself.

Coverage is graduated into four tiers, and no duty reaches a tier the annex has not placed a system in: **ordinary**, carrying nothing; **monitored**, carrying the precautionary baseline and the notice duties; **frontier**, adding the duties armed at the conditional and general stages; and **recursive-critical**, adding the ultimate-authorisation rule and the transition safety case. An anti-sharding rule closes the obvious way out: coordinated sub-threshold compute under common orchestration towards one training objective, one lineage or one deployment is one covered assembly.

At the device level there are two classes and only one carries duties. The **scale-up class** is an accelerator able to join high-bandwidth domains above the annex's size. Everything else, including inference cards, edge and phone processors and consumer graphics parts, is **pod-limited**: no device duty, counting only towards the capacity of sites and runs. Most chips, and most new entrants, whose first products are usually inference parts, therefore carry nothing, and the regime's attack surface shrinks with its reach, which answers the fair objection that adding mechanisms to chips adds places to attack.

The thresholds are linked by published formula rather than by negotiation. The site and domain thresholds are derived from the run threshold, and their inputs are measured, so as chips and algorithms grow more efficient the thresholds fall by themselves, on ninety days' notice, with the floors ratcheting downwards only, and with a change to the mark band announced a year in advance, because the band is the line at which the expensive duties attach and nobody should meet it without time to prepare. That is my answer to whether the line will one day reach ordinary machines. It will fall, and the reason the treaty grades its custody is so that falling does not mean surveillance: the heaviest custody stays with the few chips that could train a frontier system, while the lightest grade is a small tamper-evident core that says what a chip is and that nothing has altered it. Much of that lightest grade

already exists in ordinary products, in my judgement, and the parts are nameable: GlobalPlatform describes a secure element ✓ as “a tamper-resistant platform (typically a one chip secure microcontroller)”, shipped as “embedded and integrated SEs, SIM/UICC, smart microSD as well as smart cards” (page 2); and the United States standards agency records that a trusted platform module ✓ “can generate cryptographic keys and protect small amounts of sensitive information” (section 3.1, page 5), that a hardware root of trust “can present a smaller attack surface if implemented with a small codebase” (section 1, page 1), and, in its Arm architecture appendix, that the platform configuration registers “may be implemented on-chip ... or in a firmware TPM implementation, or in an external device like a discrete TPM or a Secure Element” (Appendix D, page 57). What does not yet exist is the core this treaty describes; what exists are its parts. The idea that trust has to be built into hardware is not mine alone. The United Kingdom’s Advanced Research and Invention Agency, in an opportunity-space note of 3 August 2026 ✓, holds that “Trustworthiness has to be engineered across the whole stack”, and that trust infrastructure “must extend into the physical, biological and institutional systems it acts through: chips, sensors, laboratories, neural interfaces, manufacturing processes, decision-making protocols, markets and norms” (page 3). The note is about trust infrastructure in general. It says nothing about alignment or treaties, and nothing about this proposal.

That prospect deserves the objection it attracts, because a small unforgeable identity in every processor on earth is also the architecture of a device register, and a privacy advocate is right to say so before I do.

Four things bound it, and none of them is a guarantee. The core attests what a chip is and that nothing has altered it, and it attests nothing about who holds it. It does attest its site, at every start, and the treaty is exact about to whom: the attestation runs to the operator of the site the device is installed at, and to nobody else, and the account that operator then reports to its own national authority carries location at the level of the site alone and never the content of any workload. Nothing in the design lets a location be read out of a device by a person who is not standing beside it, which is the rule section 13.5 sets out. The international register holds the public half of an identity and no holder, so it can check a signature and cannot say whose device signed. Location attestation on challenge, which section 18.2 lists among the verification tools, is produced at the site by the operator rather than read from outside it, as section 13.5 sets out; it is a measure a Party may apply to covered devices, and it is never written into ordinary ones. And the treaty does not decide the question anyway: the lightest grade reaching ordinary machines would be a decision for the Parties of a later day, taken under whatever human-rights assessment their own law requires, and nothing I propose commits them to it. The residue is that an identifier a device cannot shed is correlatable by anyone who can read it often enough, whatever the register holds, and that is a mitigation rather than a cure. Any Party that extended that core to consumer devices without answering that question in its own law would be building something this treaty did not ask for.

15.3 Classical chips, quantum chips, and anything else

A treaty written for one kind of processor leaves a door open on the day another matters. So a covered computational substrate is defined technology neutrally: classical, neuromorphic, photonic, quantum, hybrid, distributed and any successor substrate, each crossing the threshold by what it can do rather than by how it computes. Qubit count alone determines no coverage.

For a quantum or hybrid system the enforceable surface is the classical control chain around the processor: orchestration, scheduling, compilation, control electronics, workload provenance and access credentials. Those are covered chips where they meet the threshold, and they carry both layers: the hardware layer's custody, and the architecture layer's record of formation for any system trained or run on them. Whether they give effective verification of a particular quantum workload needs an assurance case of its own. And nothing here supposes that an ethical principle can be put into quantum hardware, any more than into classical hardware.

The reason for covering quantum and hybrid substrates from the start, rather than waiting, is the European Commission's own expectation. Its Quantum Europe Strategy of 2 July 2025 ⁷ records that "small-scale quantum processors exist", that their results are "still prone to significant errors", and that the thousands of error-corrected qubits are "considered necessary to solve real-world problems"; and the same strategy asserts that "quantum can accelerate the training of AI models". A treaty that waited for quantum AI to arrive would arrive too late to shape it. I do not say the question of a quantum speed-up for AI is settled either way, and the treaty does not turn on its being settled.

The regime's own cryptography carries a matching duty, stated as an outcome and never as a method: the Authority publishes and maintains a plan by which its signatures, certificates, transparency logs and long-lived records move to post-quantum methods on a stated schedule.

15.4 Registries, energy, and what is never transmitted

Two registrations carry the weight. A covered site is declared within ninety days of becoming covered and updated yearly, with its operator, location, installed capacity, power, metered energy, on-site generation, cooling, covered domains and the record of its stop drills. A covered run is notified thirty days before it begins, with short-notice and after-the-fact routes, stating the controller, the lineage, the estimated compute and band, the sites, domains and families, and the expected dates.

Nothing about weights, data, prompts or architecture passes to the Authority. The rule is **commit, do not transmit**, which section 18.1 sets out with its reasons: the contents stay with the Party's own national authority, and the international record holds a commitment to them and nothing else. Published aggregates suppress any cell with fewer than three operators or sites, so that a region with one frontier laboratory does not thereby publish that laboratory's calendar. Suppressing small cells

helps and is not a privacy guarantee: repeated releases can be differenced and linked, so every published aggregate carries a privacy assessment of its own, and a commitment to a record with little variety in it is built so that its contents cannot be guessed.

Energy does work no attestation can do. A site's metered energy over a period, multiplied by the best efficiency any type-approved family has demonstrated, gives a calibrated estimate of the compute that site could have delivered. It is not a guaranteed upper bound: a chip outside the type-approved families, or a workload run at lower precision, can do more with the same energy, and that is exactly the case an evader would choose. So the estimate states its assumptions about hardware, workload, precision and measurement. It needs no trust in any chip or vendor, it is measured by utilities under national law, it discloses nothing about models, and it keeps working when the hardware layer fails. A declared compute above the estimate opens a clarification and is never a finding by itself. Alongside it the Board publishes each year, with its uncertainty, an estimate of covered-scale capacity that is not in the register, drawn from construction notices, utility records, imports of power and cooling plant, imagery and open reporting; an estimate pointing at a Party's territory opens a clarification through a filter, never a public accusation.

The register is also the precondition of everything later. A halt of runs nobody registered could be neither administered nor verified, and an installed base nobody counted cannot be checked against any later stage. That is why the inventory has to be taken while the chokepoint still exists, and why I treat the registries as the safeguard to put in place just in case, rather than as a consequence of any result of mine.

15.5 Clouds, agents and orbit, each by its own rule

A great deal of frontier compute is rented rather than owned, so the cloud is a retail layer no coverage rule can ignore. I print no share: the rule does not turn on one, and no measured share is stated in this paper. A licensed compute service identifies customers who reserve capacity able to reach a tenth of the notice floor within ninety days, screens against the denial list of those to whom covered supply is barred, ensures every covered run on its capacity is notified, acting as reporting agent where the customer does not, and detects structuring across accounts and providers. The Authority receives run metadata only, and customers below the threshold, who are the great majority, see no change.

Agents are reached by deployment rules on the deployer, never by chip rules. An agent's own loop runs on ordinary general-purpose processors, which are in use everywhere; its capability comes from the model it calls and from the tools, credentials and networks it is given, so duties on general-purpose processors would burden all computing and reach nothing that matters. A covered agent deployment is therefore defined by the band of the model it calls, by concurrent scale and autonomous action with network or tool access, or by credentials over critical infrastructure, and it carries registration, containment, the stop, incident notification and logs. All of them belong to the precautionary baseline of section 7 and bind with it, resting on no scientific theory (Article 28B(3)).

A covered orbital cluster is a covered site of its state of registry, notified six months before launch, with its chips holding the Seal before integration and its ground segment licensed and inspected. Launch is the last physical checkpoint, and every launch is a visible, dated act. Orbit is beyond every chokepoint for siting and beyond none for manufacture.

15.6 Phasing in without stranding anything

Six rules govern the transition. Publish early and mandate late: the Assurance Standard appears years before any mandate, which opens a design-in window. Reward before you require: general licences, procurement preference and the liability floor come first, and the prepared-obligation notice puts a draft duty and its costed annex in front of industry while the evidence is still at the bottom of the ladder. Device thresholds are never retroactive, and the installed base is reached through sites and runs. Placement cut-offs follow tape-out, the point at which a design is finished and released for manufacture, rather than the calendar, so no family already committed is forced into redesign. Obligations enter in overlapping steps rather than at cliffs. And nothing is discarded: legacy compute is sealed, restricted or decommissioned under verification with compensation, never destroyed as a sanction.

My book conceded the gap that remains, in chapter 8, The Chokepoint, and I keep the concession in its own words: “Systems built with older, less advanced chips would not have these constraints.”

15.7 The load problem, stated against my own design

One defect in this machinery is mine to name before a critic does. If the notice floor falls automatically with algorithmic efficiency, and the site power criterion falls with chip efficiency, the covered population grows without limit: within a few years the floor reaches university and start-up scale, a one-megawatt site with a single scale-up accelerator is covered, and the promise that almost all development sits below the thresholds fails.

The treaty answers in four parts. Notification attaches to a run’s own compute, with lineage accounting used for certification duties rather than for notice. The site criterion counts power available to scale-up class devices. Operations are counted by reference-workload equivalence throughout, so that numeric format is not a route under the line. And a coverage-load clock counts covered sites, runs, lots, inspector-days and laboratory-hours against the Authority’s registered capacity, with a mandatory review when load passes it. The remaining trade, between coverage that tracks fixed capability and an administrative size that stays workable, is a number for the Parties to set, and I would rather it were set in the open than discovered in the tenth year.

16. Formation before training; halt and rebuild

16.1 The ask, and what it becomes

The most expensive thing I have asked for is this: that AI models halt their current structure and be rebuilt and retrained with the ethical loops embedded, Eden Protocol correction running inside the recursive reasoning loops, with that architecture in place before training rather than added afterwards, and that this be capable of being regulated.

I keep the substance and remove the avoidable cost. The duty acts on **development, not on operation**: nothing is switched off when it takes effect, and systems already deployed run on through the window (section 16.4). At the window's end an uncertified system does leave covered deployment for its rebuilt successor, so the duty reaches operation then, and I say so rather than let the phrase promise more than it holds. It binds every Party on the same day. It arrives only when enough of the world's frontier-capable compute is bound by it, a gate no vote can waive against an objecting Party. Its window is at least one ordinary model generation. Runs already under way finish. Nothing is destroyed. A newcomer whose first covered run is formed with correction built in never rebuilds. And the record that arms it states what the required design costs, measured by independent groups, with the length of the transition set from that measurement by rules written before any result exists. None of the transition periods proposed here is evidence that what they let continue is safe, and none of them prevents an urgent stop that a danger of its own justifies.

The reason for the symmetry is the reason my book gives in chapter 12, Verification and the Long Future: "The race cannot be stopped by unilateral withdrawal." Companies will not bear a cost their competitors do not bear, and states will not halt while a rival does not. A halt that is not simultaneous, and seen to be simultaneous, is not a halt. It is a transfer of the lead.

16.2 Formation before training

The first limb is the cheaper one and it comes first. From the general stage, no covered training run in the mark band begins without a formation record lodged by hash with the national authority before its first step.

The record states the system and its configuration; the architecture; where correction sits, meaning whether it takes part in each round of the system's revision or sees only finished output; the build order, meaning whether the corrector was formed before the capability it corrects; the independent external layer kept beside the embedded correction; the version of the Standard the system is formed against; and the evidence that will link the record to the run that actually takes place. It describes the design chosen and presumes nothing about which design the Standard requires (section 8.5). Accredited reviewers read it on site under confidentiality, and the Certification Office decides within thirty days whether to issue a training authorisation: its written decision that the record conforms to the Form requirements then in effect, as to that conformity alone and never as to the architecture's

merit. A training authorisation permits only the proposed run to begin, and it needs neither a trained system nor a system Seal (Article 32(1)). A certification queue must not become a veto on training, and it must not become a permission either. Silence is not approval, and no run begins on it. Where the Office has not decided in time, the operator may go to the treaty's duty president, who within days either directs the Office to decide or issues, in writing and with reasons, a provisional training authorisation for a short period and within a stated envelope of compute and activity; it permits no deployment and no release of weights, and it expires unless the Office confirms it. A refusal states its reasons and can be appealed, and a run begun under a provisional authorisation that is then refused ends within a stated period of wind-down (Article 32(3)). A correction to a record may document a fact that existed at the time the record states, and never makes a fact exist at an earlier time, so no later filing can show that anything was formed before training (Article 32(4)). The system Seal itself comes after training, on the evidence that the run matched its record. A laboratory's head of policy will say, correctly, that this is the most sensitive disclosure anyone has yet asked of the industry: a reviewer reads the architecture of an unreleased system. Confidentiality is a rule about people and not a property of mathematics, and no clause makes a reviewer forget. What the design can do it does: the record never leaves the developer's own territory, the international record holds only a hash commitment, the reviewers are accredited and drawn by lot rather than chosen, the review is of conformity to a form and not of the design's merit, and a reviewer who leaks commits a breach with a name. The residual risk is a person in a room with something valuable, and it is not removed by writing that it is.

Two honesties travel with it. A hash establishes that a record was not edited after it was lodged and nothing more: not that the record is true, and not when or how the work was done, which need trustworthy timestamps and the provenance evidence, meaning the attested link between the lodged record and the work that actually ran on which chips. If that provenance item fails at its rung, reliance on it ends, and the treaty's reversal table says which duties suspend, which stand, and what interim conditions apply until another route of verification is registered and met. The failure erases no record, establishes no conformity by refiling, and permits no activity the formation and halt articles restrict (Article 32(5)): an inability to verify a required condition never permits the activity it was meant to govern. My reason for caring is the one I have always given, that an obligation that binds and cannot be verified is obeyed by the honest and ignored by the rest; the treaty's answer is a new route of verification, not a quiet release.

And nothing in the article tests any proposition. It makes verifiable the obligation that Law V, the ARC Embedding Law, would motivate, and establishes nothing about that law, which was added on 24 September 2026, whose mathematics is set out at version 1.0 of 3 October 2026 , and which is neither registered nor tested. No duty here is keyed to its name: the keys are registered propositions at named rungs. The registration  is explicit about how narrow support would be. Of build order, a supported result “does not license any statement about where the boundary sits, about a system built in that

order being safer, or about orders outside that contrast”. Of placement: “It does not license the claim that embedded correction is safe, that it removes common-mode failure, or that the engineering proposal this programme names works as an architecture.”

For a recursive-critical system the record carries a **transition safety case**, which is at first a duty of instrumentation and disclosure alone: the depth of the system’s recursion, the trajectories of correction and of drift, and the exponents the registered notation names, each with its interval, or, where a quantity cannot be resolved, the measurement limit reached and the reason. Not resolvable is a legitimate outcome and certifies nothing either way. No numerical margin conditions a certificate until the measurement behind it has earned its standing on the ladder.

16.3 The halt

The second limb arms only at the general stage, only on the conditions the evidence annex states for it, and enters effect only behind the participation gate. From that day, on the same day for every Party: no covered run in the mark band begins until a training authorisation has been issued for its formation record, nor, from the date on which covered chips reach stage S2, otherwise than on chips holding the chip class of the Seal (Article 33(1)), and the system Seal follows the run on the evidence that it matched its record; weight-level self-retraining and self-modification loops of uncertified covered systems cease; after ninety days no further capability-directed training is conducted on uncertified covered weights; during the window uncertified systems run as frozen systems at a reduced permitted decision surface; and after the window they are withdrawn from covered deployment and replaced by rebuilt successors.

The window is the longer of twenty-four months and the median interval between successive notified covered runs in the mark band over the preceding three years, so that every covered lineage is rebuilt inside its ordinary cycle rather than out of it. Runs notified before the record of arming complete and deploy under transitional registration. Repair training is exempt: training directed at safety, security, compliance or a defect is permitted, notified and audited, because a frozen system that cannot be patched is a hazard rather than a safeguard, and capability-directed is defined by registered evaluations rising beyond a margin fixed in advance. Withdrawn weights go into escrow on their host’s territory under a key the host and the treaty hold together; they may be studied and may be certified again under a later formation record. Extensions of twelve months, twice at most, follow a recorded shortfall of certification capacity or a recorded fall of the counted share below the gate, and never suspend formation before training for new runs or any custody duty.

The chips may not be ready when the halt arrives, and the halt does not wait for them. It is the architecture layer’s own duty, and a halt that waited for the chips might never come if the chip lever failed. So where the halt enters effect before covered chips reach stage S2, the treaty does not ask of a system what only certified chips could show (Article 33(1A)). Through the window, a system is frozen if its weights are fixed and registered, it receives no capability-directed training, and it neither changes its own weights nor deploys copies of itself, whether or not its chips can attest those weights. Until

covered chips reach stage S2, a running copy of weights under the treaty's custody is declared to the national authority with the site where it runs, in place of the attestation certified chips would give. A system trained in a run that began before that date is not held to the condition that it was trained on certified chips, and until that date it need not run on them either. From that date, a system that held the system Seal on that date meets its chip condition on certified chips, or under the transitional placement the treaty already allows within a covered domain carrying cluster-level controls, so the date is no cliff. Every certificate that rests on this exception says which of those conditions did not apply.

Verification is the part every power will ask about last and care about most. Among Parties above a share of frontier-capable compute it is by reciprocal resident monitoring of the largest covered sites, their power, their cooling and the hardware that enters and leaves them, stated as outcomes and held in reserve until the article enters effect. I will not soften what that is. It is the single hardest thing this instrument asks of anybody: monitors from a rival state living beside the buildings a government regards as its strategic advantage. It is asked only for the one obligation that could not be verified any other way, only of the Parties large enough for the question to arise, only reciprocally, and only behind the participation gate, so that no state accepts it while a rival does not. If it is refused, the halt cannot be verified, and an unverifiable halt is the kind of clause I have spent this paper arguing against. A reader who thinks no capital would ever accept it is holding the strongest objection to section 16, and section 25.4 treats it as one.

The cost of the rebuild is measured before it binds, by the viability record of section 8.5: within a stated margin the ordinary generation-aligned phase-in runs, and beyond it the phase-in lengthens, the participation gate rises and the transition fund's compensation window opens. The reading is never a veto, and a measurement of cost says nothing whatever about safety.

16.4 What halting would mean for systems already in use

This question was left open when I first set the proposal out, and it deserves a direct answer. A halt of development switches nothing off.

A system already deployed keeps running through the window as a **frozen system**: its weights fixed, registered and attested by the chips it runs on, receiving no capability-directed training, modifying no weights of its own and deploying no copies of itself. If the halt has come before covered chips reach stage S2, the chips need not attest those weights, and fixed, registered weights are enough (section 16.3). Fixed weights do not fix everything a system can do, since tools, memory, retrieval, scaffolding, credentials, network access and orchestration can change it without changing a single weight. So the treaty holds a system Seal for a configuration and not for weights alone, and judges material self-modification on the whole system, those parts included (Article 2(11) and (21)). It holds a transitional registration, which is not a Seal and is never presented as one. It keeps receiving repair training under notice and audit. Its permitted decision surface is reduced, and what that means in each sector is for national law. At the end of the window it leaves covered deployment and its rebuilt successor takes its place, with its weights in escrow rather than deleted.

So the service a user relies on continues, its capability stops advancing under the uncertified structure, and the next generation of it is formed differently. Continuity of service is an aim weighed against risk, not a promise that every deployment runs until its replacement arrives. Anyone who reads the ask as switching off the world's AI has read something the text does not say, and the text is written to make that misreading hard.

16.5 Models whose weights are already public

This is the harder question, and the honest answer is that a public weight file cannot be recalled. It runs on any compute, in any jurisdiction, with no certificate and no register. No treaty reaches a copy on a disc, and I will not write a clause that pretends to.

What the treaty reaches is the next covered run and the organised deployment. A model trained from public weights is itself a covered run once it crosses a threshold, so the formation duty attaches to the derivative rather than to its ancestor. A deployment of agents built on public weights is a covered deployment once it crosses the deployment threshold, carrying registration, containment, the stop, incidents and logs, whatever hardware it uses. Open release determinations are keyed, at the architecture layer's conditional stage, to the registered mark-band capability evaluations, and at its general stage to the model-level result on the removal cost as well, which can inform a release decision and never settle it alone. A determined actor may accept the loss of capability, change the scaffolding or use a capability the test did not measure, so the determination is a risk assessment of its own, covering capability, misuse, autonomy, access to tools, ease of modification, scale of distribution and the fact that a release cannot be undone. My registration *✓* describes that test in its own words, as a drafted unit that “asks whether removing embedded correction costs general capability where removing bolt-on correction does not”, and the Eden Engineering *✓* specification sets out the same test as the monitoring removal test, at the section its own abstract names as Section 7. It has been specified and not run. Below the bands, open development is untouched, and I want it untouched: most of it is how the field learns.

What the treaty cannot do is make an already-public model carry a correction structure it was never formed with. Retrofitting is not formation, and the whole claim under examination is that the order matters. Whether a later correction helps is itself an empirical question, and the treaty does not prejudge it: a later method that meets the same requirements is admitted on the same terms. What it cannot be is relabelled as formation. A public model of the previous generation therefore stays what it is; the treaty's reach over it is the reach of deployment rules and of national law, and its longer answer is that the generation after it is formed differently. If the persistence and embedding propositions fail, that answer fails with them, and the registration *✓* says so in its own words: if the four propositions that carry the Eden Protocol fail, “the Protocol's engineering recommendation fails with them”.

16.6 What the precedents for a halt actually did

Halting a technology while something better is built is not a new idea, and the record is worth stating exactly, with what each instrument did not do.

The **Montreal Protocol** ↗ halted an existing class of products on dated steps: a freeze at each Party's 1986 level, then reductions, then zero consumption and zero production from 1 January 1996, with a later date for the basic domestic needs of developing-country Parties and a route for uses the Parties agree to be essential. It controls quantities, not designs, and it does not name what replaces the substances.

The **Biological Weapons Convention** ↗ gave a deadline for existing holdings: each Party destroys or diverts to peaceful purposes, within nine months of entry into force, the agents, toxins, weapons, equipment and means of delivery specified in Article I which are in its possession or under its jurisdiction or control. It establishes no standing inspection regime, but Articles V and VI provide for consultation and cooperation and for complaints to, and investigation by, the Security Council.

Among the calls addressed to AI, the Future of Life Institute's open letter of 22 March 2023 ↗ asked "all AI labs to immediately pause for at least 6 months the training of AI systems more powerful than GPT-4", adding that the pause "should be public and verifiable, and include all key actors. If such a pause cannot be enacted quickly, governments should step in and institute a moratorium." It states that it is not a pause on AI development in general, and it names no body that would verify the pause, no test for when protocols are adequate and no hardware mechanism; its six months is a minimum with no end condition. The same institute's later statement on superintelligence ↗ calls for a prohibition "not lifted before there is" "broad scientific consensus that it will be done safely and controllably" and "strong public buy-in", and likewise names no body, no test and no mechanism.

A precedent in another field is the 2015 international summit statement on human gene editing ↗, from an organising committee of academies of the United States, the United Kingdom and China. It held that "It would be irresponsible to proceed with any clinical use of germline editing unless and until" safety and efficacy issues were resolved and "there is broad societal consensus about the appropriateness of the proposed application", recorded that "At present, these criteria have not been met for any proposed clinical use", and said the question "should be revisited on a regular basis" while basic and preclinical research continued under legal and ethical rules. It binds no state and names no body that decides when its criteria are met.

Four things the treaty takes from them: dated steps with an essential-use route, from Montreal; a deadline reaching existing holdings and not only new production, from the Biological Weapons Convention; the conditional form with a duty to revisit, and research continuing meanwhile, from the summit statement; and the pause letter's requirement that a halt be public and verifiable, which here is reciprocal resident monitoring rather than an appeal to good faith.

One thing none of them has is the thing this treaty is for: a rule written before the data that states which evidence would open a halt, which would close it, and what the halt costs. Each of those instruments is either a call with no test or a schedule with no evidential trigger. A conditional halt with its conditions fixed in advance, reversible by the same standard, is the narrow addition I propose to those instruments, and I claim nothing wider: a conditional treaty with a compute threshold and an authority to pause has been proposed by others, and section 23.3 concedes it to them.

16.7 The limits of this section

Three, stated together.

The propositions from which the halt takes its reason are scoped to recursion in a system whose weights are fixed. The registration ↗’s rule is that “NO PROPOSITION MAY BORROW A SCOPE.” So the halt arms only on support obtained in the weight-level scope itself, and that scope has today no instrument at all for two of the three propositions concerned.

The evidence is weakest at exactly the place I care about most. Of the four propositions with no instrument anywhere in my programme, the two on persistence and on build order carry this duty. A supported persistence result, the registration ↗ says, “does not show that embedded or in-loop correction is superior, sufficient or safe”.

And the timing is against me. On my own arithmetic in section 2.5, a correction mandate of this kind is unlikely to take effect before the chip lever has largely diffused. What binds inside the window is what needs no theory of mine: the baseline, the registries, the incident channel, national adoption, procurement and voluntary use of the Seal. I write that down rather than let a reader discover it in the tenth year of a negotiation.

17. The ban and trade law

17.1 What I asked for, and what a treaty can do

On 26 September 2026 I proposed that chips without the Eden Seal be banned internationally. I hold to the aim, and I have had to learn the limit. A treaty binds the states that join it, and creates no obligation for any other state without that state’s own consent. No wording of mine changes that, and an instrument that pretended otherwise would not survive its first negotiating session.

So the ban in this treaty is not a prohibition laid on the world. It is what the Parties do in their own markets, with their own tool supply and their own compute services, and what a state outside then finds it costs to stay outside. That is closer to what my book said than to what I said in September:

“Non-certified chips could not be sold in signatory markets, representing the vast majority of global AI demand.”

Infinite Architects, chapter 8, The Chokepoint

The trade Part is the developed form of that sentence, and I state its reach exactly rather than leave a reader to discover that the word “international” was doing work the law cannot do.

Two further limits belong in the same breath. The ban wakes only at the stage where the hardware controls have met their own registered tests, so nothing here binds on argument. And it reaches non-Parties only when the Parties, together with non-Parties found in full compliance, hold the share of the world’s frontier-capable compute that the participation gate names, which I propose at two thirds. A ban applied by middle powers against the two largest producers would cost the appliers most, would probably not be applied, and would therefore deter nobody. Until that condition is met, the treaty offers certification and access rather than exclusion.

The convention of section 6.5 bears hardest on the sections that follow: every period, share, level and count in them is a proposal of mine, bracketed in the treaty text for the Parties to set, and nothing in the argument turns on the particular figures.

17.2 The Montreal mapping

The instrument that did this before, and did it in a way the law recognises, is the Montreal Protocol on Substances that Deplete the Ozone Layer ¹. Its Article 4 is the model, and the treaty borrows its shape without borrowing its legal footing: the analogy shows how a trade regime has treated states outside it, and it establishes neither that this treaty is compatible with the trade agreements nor that a measure that worked for ozone will work for computation.

Montreal provides that “each party shall ban the import of the controlled substances in Annex A from any State not party to this Protocol”, and that “each Party shall ban the export of any controlled substances in Annex A to any State not party to this Protocol”. It extends the ban to an annexed list of products containing the controlled substances, and it asks each Party “to the fullest practicable extent to discourage the export to any State not party to this Protocol of technology for producing and for utilizing controlled substances”. And it leaves the door open: trade “may be permitted from, or to, any State not party to this Protocol, if that State is determined, by a meeting of the Parties, to be in full compliance”.

Three further elements of that Article belong in the map, because each answers a question the shorter account leaves open. Montreal reaches what is produced with a controlled substance but does not contain it only after the Parties have determined that a prohibition is feasible, and then by a list, which is the right order and not an obvious one: the treaty therefore prohibits nothing of the kind until the Conference has determined feasibility and given its reasons either way, which is the hardest trade question in the whole design. Montreal pairs its duty to discourage technology exports with a firmer

duty to withhold new subsidies, aid, credits, guarantees and insurance for them, and exempts from both whatever improves containment and recovery; the treaty takes both, and points the exemption at the safety tools, which are exactly what it wants to travel freely. And Montreal defines a non-Party substance by substance, so that a state bound as to one schedule is not treated as outside for all of them; the treaty's definitions article does the same, schedule by schedule.

The treaty maps each of those onto covered chips, the scheduled manufacturing items, the products that contain covered chips, and the technology for making them. It adds one thing Montreal had no reason to have, because ozone-depleting substances are not rented by the hour: a rule for compute services, so that a Party's own data centres do not become the route around its own import rule.

Two things Montreal does not state, and I will not borrow them. It says nothing about how any product works, and it certifies nothing. I have not read who its early members were, so I draw no comparison between its membership and this treaty's. What the participation gate rests on is a reason of design and not a reading of Montreal's history: a trade measure applied by a membership that leaves out the principal producers costs the appliers most.

17.3 Two doors, and no secondary measures

A wall with no door in it is an invitation to build a second chain. The treaty therefore carries two.

The first is Montreal's own: a non-Party that the Conference finds in full compliance, and that files the declarations, trades as a Party. Article 4 of Montreal  leaves unstated the majority by which that finding is taken, which is a gap a door of this kind cannot afford; the treaty supplies it, at the same majority that governs the content of the Standard, and adds what that Article also does not state, which is that the finding states its reasons, is published, is reviewed every few years and lapses if it is not renewed; a refusal cannot rest on non-membership alone. A door with no lock on it is not a door. The second is new, and it is for firms rather than states. An undertaking established in a non-Party may buy covered compute, covered chips and scheduled service from Parties by accepting the terms itself: it lodges the records and the notices, accepts evaluation, custody and inspection of the covered activity, and accepts the consequences if it breaches. The facility in the Party answers for the contract, and a breach of it is dealt with under the contract and the compliance Part; it puts no one on the denial list without the listing procedure. That door binds no state and needs no state's consent to open, and it means that a laboratory whose government stays out is not thereby locked out. Two limits are plain. A non-Party may forbid its own firms to walk through it, and the treaty cannot stop it. And a firm's consent does not carry its state's: the treaty lets nobody inspect on a non-Party's territory without that state's permission, so where the covered activity sits in a state that refuses, what the firm can offer is its records, its evaluations and its custody arrangements, and not an inspector at its door. The door is therefore wider for a firm whose covered activity runs on a Party's soil than for one whose government has closed the ground under it.

One instrument of the book's I have dropped. Chapter 8 proposed that enforcement "follow the model of existing sanctions regimes", with "secondary sanctions affecting companies that deal with them". The treaty applies no measure to a non-Party's firm on account of that firm's dealings with other non-Parties. Secondary measures of that kind are contested in law, and I take the states whose participation the treaty most needs to be the states most likely to resist being made the target of them; that is my inference, and I have read no instrument of theirs to support it. A measure that costs the treaty its most necessary participants is not an enforcement tool. It is a way of making a treaty smaller.

17.4 Exclusion is the last rung, not the first

Trade denial is the most tempting instrument in the whole design and the one most likely to destroy what it protects. Used early, it can accelerate exactly the indigenous substitution that ends the leverage it depends on. So the order runs the other way: mutual recognition first, then market access, procurement and assurance, and denial last.

Inside the membership the same principle holds in a stronger form. From the stage at which the chip class of the Seal becomes mandatory, each Party admits on equal terms the certified chips of every maker, whatever chain made them, and imposes on covered chips no governance mechanism beyond the treaty's own outcomes. Any maker's chips, a rival's included, may earn the Seal on identical terms and fees. A Party that distrusts what it imports may re-test it in its own accredited laboratory. Recognition across rival chains attests declared, inspected and sampled production, and not the absence of a state implant; I say what it attests in section 18 rather than letting the word "recognition" carry more than it holds.

17.5 The footing in trade law, stated as open

A treaty that shuts a market must answer to the law of trade, and I would rather set out the ground it would stand on than assert that the ground is firm.

The General Agreement on Tariffs and Trade 1994 / permits measures "necessary to protect public morals" and measures "necessary to protect human, animal or plant life or health", subject to a condition that governs everything: such measures must not be applied in a way that would be "a means of arbitrary or unjustifiable discrimination between countries where the same conditions prevail, or a disguised restriction on international trade". The treaty is drafted against that condition rather than against the exception, because the condition is where a measure of this kind is most likely to fail: identical terms for every Party's firms, one list, one set of fees, the same door for every non-Party in like circumstances, and no exception written to favour anybody's champion.

I go one step further than borrowing the form. The treaty adopts that condition as an obligation of its own, in its own words, in the second article of the trade Part: every measure of that Part, every measure of the compliance Part, and every other measure of the treaty that restricts trade in goods, technology or services, is applied so as not to constitute arbitrary or unjustifiable discrimination where

the same conditions prevail, and so as not to constitute a disguised restriction on international trade. Adopting it inwardly is stronger than relying on an exception in somebody else's agreement, for two reasons. It gives a Party that is discriminated against a remedy inside this treaty, before the compliance panel, rather than only in another forum. And it requires no view at all about how that other agreement would be applied, which I am in no position to take. Three further disciplines sit beside it in the same article: every measure carries a published record of its objective, the evidence that armed the obligation it enforces, the risks of not meeting the objective, and the less trade-restrictive means considered and why each was insufficient; no conformity assessment is stricter, or applied more strictly, than is needed to give the Authority adequate confidence; and a measure is lifted when the evidence that armed the duty no longer supports it and narrowed when a lighter means would do. The published record is also the answer to the charge that the regime is protectionism in safety clothing, which is the charge I expect a trade ministry to make first.

The General Agreement's security exception has three limbs. It does not require a Party to furnish information whose disclosure it considers contrary to its essential security interests; it does not prevent a Party from taking action "which it considers necessary for the protection of its essential security interests" in three named cases, being fissionable materials, the traffic in arms, ammunition and implements of war and such traffic for supplying a military establishment, and action taken in time of war or other emergency in international relations; and it does not prevent a Party from acting under its obligations to the United Nations Charter / for the maintenance of international peace and security. The treaty's own security clause is drafted on the wording of the second limb, and its aggregate declaration of national-security compute is the answer to the first, which is the limb that carries the rest of the design's weight in practice. Neither that Article nor the general exceptions names computing, semiconductors or artificial intelligence, and I do not claim that any of them already covers this subject. Nor do I claim that the treaty's obligations are Charter / obligations within the third limb: they are not.

One report bears on that exception directly. In *Russia, Measures Concerning Traffic in Transit* /, a panel reported on 5 April 2019 that whether an action meets the requirements of the exception's subparagraphs is for panels to review, "rather than leaving it to the unfettered discretion of the invoking Member" (paragraph 7.102). So the security exception is not beyond review, and the treaty's own security clause, which follows its wording, cannot enlarge it.

The Agreement on Technical Barriers to Trade / supplies the discipline the certification side must meet. Technical regulations "shall not be more trade-restrictive than necessary to fulfil a legitimate objective, taking account of the risks non-fulfilment would create", and the legitimate objectives it lists include "national security requirements" and the protection of human health or safety. Where a positive assurance of conformity is required, access to assessment must be granted to suppliers of like products from other Members on conditions no less favourable than those accorded to suppliers of like products of national origin or from any other country in a comparable situation, including, where the procedure envisages it, the possibility "to receive the mark of the system". Members are to ensure,

whenever possible, that “results of conformity assessment procedures in other Members are accepted, even when those procedures differ from their own”, provided they are satisfied that those procedures offer assurance of conformity with the applicable technical regulations or standards equivalent to that offered by their own procedures, and, where a positive assurance is required, to “formulate and adopt international systems for conformity assessment” wherever practicable. Two further sentences shape the treaty more than any argument of mine could. Regulations should be written “in terms of performance rather than design or descriptive characteristics”, which is why the treaty’s Assurance Standard states outcomes and requires no design of anyone’s. And they “shall not be maintained if the circumstances or objectives giving rise to their adoption no longer exist or if the changed circumstances or objectives can be addressed in a less trade-restrictive manner”, which is the trade-law twin of the rule in section 8 that no duty outlives its reason, a signed decision withdrawing it when the reason fails.

The treaty seeks a waiver for its trade measures, applies them meanwhile consistently with each Party’s existing obligations, and records the footing as open. A waiver is a decision of the members of the World Trade Organization under Article IX of the Marrakesh Agreement ↗, not of this treaty’s Parties: only one actually granted can be relied on, and no protocol of this treaty can amend another agreement or bind a state that has not consented to it. Nor is the goods analysis the whole of it. The treaty’s rule for compute services raises questions under the General Agreement on Trade in Services ↗ that the analysis above does not answer, and a defence under the goods agreement is not a defence under the services one; that analysis is owed measure by measure, with each Member’s own commitments read. Beyond the one report cited above, I state no view on how any of these provisions has been interpreted. This is the part of the design most in need of a trade lawyer’s hostile attention, and I would rather have it early than at a panel.

17.6 When the lever stops working

The chokepoint is a window. If the share of frontier-capable capacity outside the treaty’s reach stays high for two consecutive years before covered chips reach stage S2, the Parties either amend the instrument into a regime that works at the site and the system without the chip lever, or they record that it has failed. An organ cannot convert a treaty into another treaty by a vote, so what the Conference does is open the amendments for ratification, and the duties that rest on the chip lever alone suspend from the day it decides. The rules that rest on the chokepoint retire when the chokepoint stops giving meaningful leverage, by their own clocks and not by anyone’s decision. A trade measure whose leverage has gone is not a safeguard; it is a tariff with a moral preface.

18. Verification, and what it cannot see

18.1 The rule

Verify outcomes at the boundary. See designs only in a sealed room. Never see models.

One rule stands before those three, and the treaty states it in the first paragraph of its verification Part, because I take a rival's first fear to be not that it will be inspected but that what inspection finds will be used against it somewhere else. Verification is conducted for the sole purpose of verifying the obligations this treaty imposes. Nothing an inspection, an access or a measurement produces is used for any other purpose, and no organ may require information that verifying those obligations does not need. That is the minimum-information discipline the chemical weapons regime's confidentiality annex sets, quoted in its own words at the end of section 18.2, and this treaty takes it whole. A Party that used treaty material to target an export control, to price a sanction or to help its own industry would be in breach of the article that let it see the material at all. I cannot promise that no state would do it. I can make doing it a breach with a name, which is what the regimes this treaty borrows from offer.

Everything below follows from that rule and from one more: the regime holds commitments, not contents. Run notices, formation records, weight registers, measurement reports, fabrication and tool declarations and energy data stay with the Party's own authority; the international record holds a hash commitment, a band and a region, lodged at the time; inspectors reconcile the commitment against the contents on site, under managed access. A world register of every frontier run would be a targeting map, and a vault of escrowed weights would be the prize of every intelligence service. A regime that gathered the world's AI secrets into one building would be attacked for them, and would deserve to be refused before it was.

18.2 The toolkit

Every tool in the treaty is chosen so that what it exposes is smaller than what it establishes, and each carries its own ceiling.

THE TOOL	WHAT IT ESTABLISHES	WHAT IT NEVER EXPOSES	ITS CEILING
Declarations and a compute balance for each site	quantities of covered devices, against what was declared	designs; models	self-declared, and needs the cross-checks below
Energy, metered	a calibrated estimate of what a site could have computed	anything about a model	it cannot tell covered training from other work of the same energy, and a chip outside the approved families can beat the estimate
Tamper-evident accounts of operations, read on site	operations performed in each period	the content of any workload	rests on tamper resistance and on provenance, neither yet shown at this scale
Managed access to covered domains and to the logs of the few human decision points in manufacture	presence, configuration and who decided what	shrouded proprietary and national-security material	inspectors see less than everything, by design
Destructive sampling and design equivalence of the small governance core, in the Authority's sealed laboratory	that sampled parts match the approved core	design data beyond the check; the rest of the die	bounded by what the analysis of a unit can see
Governance firmware, reproducibly built, its image hash public	that the image on a chip matches audited source	the vendor's other firmware	trust in the audit panel
Evaluation brought to the weights, in the developer's own facility	a system's registered measures	weights and training data	depends on provenance and on features whose robustness is unproven
Hash commitments lodged at the time	that a record was not edited afterwards	the record's contents	a commitment proves nothing about what then happened
Provenance evidence, linking work to chips and to a lodged record	which work ran where, under which record	weights	the weakest link in the chain, and treated as such
Random selection no one controls, for lots, panels and teams	that nobody chose who checks	nothing	needs one honest input, which the procedure supplies

THE TOOL	WHAT IT ESTABLISHES	WHAT IT NEVER EXPOSES	ITS CEILING
National technical means and open sources	undeclared sites, construction, power, tool imports	nothing	uneven between Parties
Location attestation on challenge, produced at the site by the operator and handed over	that a device is where it was declared	nothing else	available where a Party chooses it, never a channel into a chip from outside its site (section 13.5), and never written into every chip in the world

Older regimes supply the patterns. Complementary access at short notice to resolve a question, on the model of the additional protocol to comprehensive safeguards agreements ↗, with the question stated in writing before the access, the access confined to that question alone, and the record saying whether it was resolved: a light instrument that keeps a declaration honest without the cost of a challenge, and an instrument that stays light only because its purpose is written down before anybody arrives.

Challenge inspection as the backstop, on the model of the Chemical Weapons Convention ↗, which gives each Party the right to an inspection “for the sole purpose of clarifying and resolving any questions concerning possible non-compliance”, to be “conducted anywhere without delay”, and which pairs the right with a duty: “Each State Party shall refrain from unfounded inspection requests, care being taken to avoid abuse.” And managed access, from the same Convention’s verification annex, which puts the inspected State Party “under the obligation to allow the greatest degree of access taking into account any constitutional obligations it may have with regard to proprietary rights or searches and seizures”, allows it to protect national security, and then adds the sentence that makes the whole thing work: that protection “may not be invoked by the inspected State Party to conceal evasion of its obligations”. Its confidentiality annex sets the discipline the Authority inherits, to “Require only the minimum amount of information and data necessary for the timely and efficient carrying out of its responsibilities”.

Two things about an inspection matter as much to an operator as what an inspector may look at, and the treaty states both: when access begins, and when it ends. Access begins within a stated period of the team reaching the point of entry, and the inspection lasts no longer than a stated period, extendable once by agreement. Those periods run whether or not any deliberation is on foot. An inspection with no stated end would be an occupation of a site rather than a verification of an obligation, and a firm asked to accept the first will refuse the second as well.

18.3 Every assurance carries the name of what enforces it

Some assurances are properties of a device: it attests its identity and the image it runs, and it refuses a rule change without an authorisation of the right class. Some are site procedures: two vetted people apply a change, and two vetted people can stop the cluster. Some are national law: a stop order, a fine, an offence. Some are contracts: the continuation of custody on a supplied item. They are not interchangeable, and a treaty that presents a procedure as a property of silicon is selling something it cannot deliver.

So every assurance in the register carries the label of what enforces it, and every certificate states the claim it makes and no other. An attestation is evidence of the measured claim it names. A genuine chip does not make the system running on it good. Nothing in the verification Part licenses a sentence that says otherwise.

18.4 What verification cannot see

I would rather list this than have a reviewer list it for me.

A host state tampering at its own root of trust. It provisions the chips, and against it the hardware layer gives detection at best. Witnessed provisioning helps only if the witnesses can confirm, with instruments they bring and control, that the identity on the record is bound to the device in front of them; until that outcome is met, witnessing deters the staff at the site and not the state that owns it.

What sampling can find. As I understand it, the arithmetic of a sampling plan is sound for random or careless non-conformity and for crude substitution. A state that designs a function into every unit of a lot is found only if the analysis of a unit can see it, and some modifications are not visible to an optical inspection at all. I cite no study for that, and the design does not rest on one: it rests on the admission that the assurance sampling gives is bounded by a detection probability nobody has measured. So the treaty registers that probability for each destructive method, with the adversary it assumes, rather than printing a confidence level that reads as assurance against the host.

The rest of the list is shorter to read as a table than as prose, and none of the answers in the last column is complete.

WHAT VERIFICATION CANNOT SEE	WHY	WHAT ANSWERS IT IN PART
Algorithmic progress	it is software, not hardware	thresholds that ratchet down, and the capability override
Training spread across many small sites	each sits below the site threshold	aggregation by common control; energy; the estimate of unseen capacity
Weights once they have left custody	they run on anything	custody of weights, declared copies, and the offence of taking them
Recursion at inference time on small machines	below every threshold	deployment rules above scale
Agents on ordinary processors	chips are not the lever here	duties on the deployer, whatever hardware it uses
Compute in orbit	there is no site at which a person can act	the pre-launch checkpoint, the licensed ground segment, the power budget
National-security compute beyond managed access	sovereignty	coverage at manufacture, aggregate declaration, challenge inspection
Hidden function in the rest of a die	it is too large to check	the isolation outcome, the adversarial tests, red teams
Diversion of devices after sale	they are small and numerous	certificates by consignment, each site’s balance, location on challenge, the denial list
Machine-generated paperwork	documents can be produced at scale	provenance, and every finding of fact signed by two human inspectors of different constituencies, none resting on an AI system’s output alone

And the one my own book named against its own case, which section 14.3 quotes in full:

“Systems that can reason about ‘playing along now’ can, in principle, reason about playing along with verification itself.”

Infinite Architects, chapter 12, Verification and the Long Future

That is why the treaty leans on records, custody and provenance rather than on behaviour, and treats re-measured behaviour as corroboration only.

One more, which industry raised and I accept. The treaty forbids the channel and not only the act, in the terms section 13.5 sets out in full. What verification cannot do is prove that no such means exists: it reaches declared, inspected and sampled production and the tested isolation outcome, and it does not establish that a state has built nothing into the chips it makes. That is the limit this section is about,

and it is the reason the prohibition is drafted with the remedies a distrustful buyer can use itself, which are re-testing what it imports and requiring that no ordinary update alter the behaviour of covered compute unless installed by the site's own authorised act. Tests of declared interfaces and sampled devices show what was tested and nothing more; re-testing, controlled updates and fewer interfaces strengthen the assurance, and what each adds is measured rather than assumed.

18.5 The regime's own failures

A verification regime is a target, and the attacks on it are not exotic. Theft of a key. Firmware signed by someone who should not have signed it. Substitution of a part. Replay of an authorisation that was valid last year. Side channels. A virtual machine that lies about what it is. Inspectors who collude. A compromised root. An update that carries the attack. And all of them assisted by systems better at finding flaws than the people defending against them.

The answers are structural rather than clever. No manufacturer is at once the maker, the root of trust, the source of the evidence and the certifier. Certificates are revocable, recorded in a transparency log, and recoverable after a compromise. No authorisation older than or superseded by one already accepted is ever accepted again, and none keyed for one class, device or version is accepted for another. The regime's own cryptography migrates to post-quantum methods on a published schedule, stated as an outcome and never as a design. The Authority holds only what a procedure needs, for as long as it needs it, notifies affected operators within a short fixed period of any compromise of its own systems, pays for loss its leaks cause, and is audited from outside every year. A regime less secure than what it guards has no business asking anyone to trust it.

19. Consequences of breach

19.1 Two tracks

The question I asked was what happens when the rules are broken. The answer begins by separating two things that are usually confused.

An incident is an event: a stop that did not work, an agent that left its enclosure, an intrusion, a tamper, a loss of weights. It is reported on a clock and investigated for prevention, and the occurrence of an incident is not by itself a breach. A breach is a failure of an obligation that binds: silence, lateness, destroyed evidence, a standard that was not met before the event, a run that should have been notified, a deployment that should have been evaluated.

The regime's product is knowledge. Everything in this Part is built to buy it: no fault for honest reporting, a firewall around what is given to the investigators, leniency for those who come forward, protection for insiders who have nowhere else to go. And everything is built to make the corruption of that knowledge the gravest thing a party can do. Tell early and lose little; hide and lose everything.

19.2 The clock, the firewall and where no fault stops

The clock runs from awareness, and awareness is defined so that it cannot be postponed: the logged human review of an alert, or a short fixed period after an alert that nobody has reviewed. Keeping monitoring that raises alerts to people is itself a duty, so an operator cannot delay awareness by not looking. Immediate notice for the gravest class with a continuing effect on somebody else; early warning; an initial report; a final report; and a confidential quarterly return of near misses, because base rates matter as much as headlines and single reports cannot be read without them. Preservation of logs and state is automatic at awareness, in a place the system under investigation cannot reach.

What is given to the investigators is used for prevention. The report, and the analysis written for it, are not evidence against the notifier in compliance, civil or criminal proceedings, and the investigators' reports are not evidence of fault. The protection reaches the report and never the conduct: facts that can be found another way, records that existed anyway, physical evidence and what is needed to prevent imminent harm stay available, and so does a victim's remedy. Whether an exception applies, for wilful concealment, tampering, falsification or gross negligence, is decided by a court or other competent body, subject to review, and not by the notifier or the investigators alone. That protection is not a favour to industry. A candid notification that becomes a plaintiff's exhibit is a notification nobody will write beyond the legal minimum, and the whole design then fails at its first step.

No fault covers the event and never the duties. A containment standard never implemented, a stop never drilled, monitoring that alerts nobody: each is a breach whether or not an incident follows, proved from the inspectorate's own evidence and never from the notification. Gross negligence and wilful misconduct are outside the protection entirely.

19.3 The schedule, published before the conduct

One schedule, the same for every Party and every firm, negotiated with the treaty and sleeping with it, so that everyone knows the worst case before any duty arms and nothing can be invented afterwards.

Consequences divide by what they answer. Protective consequences follow risk: suspension, stop orders, preservation, escrow, a monitor, provisional listing, on-site safing after a confirmed tamper, withdrawal of future supply and service, enhanced inspection. They can act at once, with a hearing shortly afterwards; they are never reduced by candour; the least restrictive measure capable of securing compliance is applied first; they last no longer than the risk requires, and a measure is lifted when the Compliance Panel records that the non-compliance has ceased or that the risk has passed; and where one turns out to have been unjustified, the operator is compensated. Punitive consequences follow culpability: a published finding, a fine within stated maxima, disgorgement of what a knowing breach gained, loss of the voluntary benefits, debarment, and referral of individuals to national prosecutors. They come only after due process with time limits, and they are graded on six levels from notice and cure up to facility measures.

Three rules keep the grading honest. No rung is reached except through the one below it or through an aggravating fact the schedule names. A knowing breach is never meant to pay: the financial consequence starts at the value of what the breach gained. That is a floor, not a guarantee of deterrence, because some breaches are never found and some gains cannot be recovered; the schedule therefore weighs the chance of detection, the gains that escape recovery, severity, culpability and repetition as well. And leniency buys down punishment, never protection: a self-reporter can lose its fine, though never the value of a knowing breach's gain, and cannot keep an unsafe system running.

19.4 Nothing is destroyed, and no switch is thrown by inaction

Withdrawn weights go into escrow under custody, on the host's own territory, where the host holds a necessary share and the treaty holds a necessary share and neither can release alone. They may be studied, and they may be the starting point of a rebuild. The regime never orders destruction: a destroyed model cannot be examined for what went wrong, escrow restricts use without taking title, and destroying one copy would not reach copies elsewhere.

Chips of a suspended lot are declared and brought back under the rules by re-testing, or restricted below the thresholds. A device leaves covered operation only on a confirmed tamper at its site, or on a final revocation after appeal, in either case carried out on site by people. That rule is about the final removal of a device. It does not prevent a Party's own authority, where urgent protection is justified, from ordering a local stop, isolation or restriction that is necessary and proportionate before an appeal is decided, with its reasons, a time limit and prompt review. No expiry, no vote, no failure to act and no message changes what a chip does. A switch that no one holds but that inaction throws is still a switch, and the treaty does not build one.

19.5 States, and what the ladder cannot do

Against a Party the ladder runs from consultation, through a published finding with an action plan and help to meet it, to suspension of specified privileges, to withdrawal of future supply and service from its facilities not in good standing, to collective measures the Conference may recommend, to referral. Four privileges are never suspended: the right to be heard, the right of appeal, access to the incident channel, so that a Party in breach can still warn and be warned, and the right to withdraw, because a regime that could trap a Party inside it would, I expect, be refused at the signing table.

The top of that ladder is borrowed, deliberately, from the Chemical Weapons Convention ⁷, which lets its Conference "restrict or suspend the State Party's rights and privileges under this Convention until it undertakes the necessary action to conform with its obligations", may "recommend collective measures to States Parties in conformity with international law", and "shall, in cases of particular gravity, bring the issue, including relevant information and conclusions, to the attention of the United Nations General Assembly and the United Nations Security Council". Collective measures are recommended and not ordered, and I keep that limit rather than inventing a power I do not expect any state to grant.

So here is the honest answer about great powers. Against the United States or China the ladder gives an objective finding on the public record, made by a panel from which that Party's nationals and its declared allies are excluded, the loss of recognition and market access for its certified products, and the withdrawal of future supply while the chokepoint holds. It does not compel. A referral meets the permanent members' veto. Nobody should sign this treaty believing otherwise, and I will not write a sentence that lets them.

One consequence in that list is carried by private parties, and it costs them something the treaty cannot repay. Withdrawal of future supply and service means a supplier stopping work for a customer, and a supplier that does so may be sued for it in the customer's own forum. The Parties' own law protects conduct the treaty requires, which answers the case in every Party; it does not answer a claim brought where the treaty does not run. What the treaty can do is confine the measure to future supply and service rather than to anything already delivered, publish the schedule before the conduct, and require the supply contracts themselves to carry the condition, so that a supplier is enforcing its own terms and not only an instrument its customer never signed. What is left is a residual commercial risk sitting with a private firm for a public measure, and it is one of the reasons the ladder puts denial last.

19.6 Individuals, and the offence

Here my book was already right, and its sentence has become the treaty's:

“Circumvention would result in decertification, fines, and potential criminal liability.”

Infinite Architects, chapter 8, The Chokepoint

The treaty's contribution is to say which conduct, and with what mental element. Each Party makes it an offence to do any of the following knowingly or intentionally: to tamper with a covered chip's governance core, its attestation, its provisioning, its seals or its custody material, or with the record or the key material; to falsify, destroy or conceal records the treaty requires, or evidence in a replication or an evaluation; to direct or permit a covered run without the required notice, or a deployment without the required evaluation, where the obligation is armed; to direct or permit the concealment of a notifiable incident; to supply covered chips, scheduled items or covered compute in breach of the trade measures or the denial list; to obstruct an inspector, an investigator or a custodian; to refuse or fail to carry out a final revocation, an on-site safing or a withdrawal from covered operation that the treaty requires; and to retaliate against a protected reporter. A good-faith error is never an offence. Neither is a good-faith halt, nor a good-faith refusal of an unlawful instruction. Research, testing and repair within a scope the operator has declared and authorised are not offences, and nor is an act in an emergency that is necessary to prevent imminent harm (Article 43(5)). And the two offences of running or deploying without what the treaty requires reach only a duty standing on the published register of reasons on the day of the conduct (Article 43(3A)), for the reason section 8.6 gives: personal

criminal liability cannot turn on a duty whose existence a defendant had to infer. No determination unpublished at the time, and no later correction of the register, creates or enlarges criminal liability for earlier conduct (Article 8A(5)).

The enforcement pattern is the one the Geneva Conventions / use for grave breaches, and I take it as a pattern and claim nothing more. Each High Contracting Party there undertakes to enact the necessary penal legislation, and each “shall be under the obligation to search for persons alleged to have committed, or to have ordered to be committed, such grave breaches, and shall bring such persons, regardless of their nationality, before its own courts”, or hand them over for trial elsewhere. The treaty asks the first of those three things, to legislate, of every one of its offences, and all three, to legislate, to search, and to prosecute or extradite, of one offence alone: intentional tampering with a certified chip’s governance core, its attestation, its provisioning, its seals or its custody material, or with the evidence ledger or key material, because that is the offence a state outside the regime could shelter. A reckless act stays within ordinary jurisdiction. Extradition and criminal judgments stay with each Party’s own authorities and courts, which no organ of the treaty can reverse: extradition is subject in every case to the requested Party’s own law, and a Party credits a penalty another Party has already imposed for the same conduct (Article 43(8A) and (9)). The treaty does not call any of these offences a war crime, and nothing in it suggests that the Conventions already make this conduct a grave breach: the pattern is a drafting comparison, not a legal classification.

Two further limbs of that pattern belong with those three. The first completes the ladder downwards. A penal list is a short list by design, and an instrument with nothing below it is silent about every breach that falls short of a crime, which is most of them; so each Party also takes the measures necessary to suppress acts contrary to the treaty that are not offences, by administrative, regulatory or contractual means, with no offence and no penalty attached. The second runs the other way, and protects the accused. A person proceeded against for one of these offences enjoys the safeguards of proper trial and defence that the Party’s own law and its international obligations provide, and in no case less than that Party affords for an offence of comparable gravity. The Conventions supply that floor by pointing at another treaty about prisoners of war, which cannot be borrowed here, so the treaty states it in its own words. An instrument that asks states to create offences and says nothing about the trial of them is half an instrument.

A duty inside the firm sits beside the offences and is not one of them. Each Party requires every covered operator, manufacturer and design house to appoint a responsible person, and requires that person to prevent and suppress a breach within their power and to report one they know of. That is how a duty over an industry acquires a holder, and it is the point at which a general undertaking to ensure respect stops being a sentiment. It creates no liability in a person for what they could not have prevented and did not know of, because a duty that punished ignorance would be answered by arranging not to know.

An AI system is not a defendant. Where an act is attributed to a covered system acting for itself, the consequences reach the affected versions of its lineage, its compute and its weights. Responsibility stays with the operator for the duties it held, and with each state for the duties the treaty gives it over the operators within its jurisdiction.

19.7 Fairness, and the rule that punishment follows the evidence

Time limits at every stage, and a duty president, one always on call, who rules within days when the regime misses one of its own. Standards of proof that rise with the consequence: credible evidence of risk for a protective measure, and the balance of probabilities for a punitive finding at the two levels above notice and cure. Clear and convincing evidence for a finding at any higher level, for a breach of integrity, for a listing other than a provisional one and for a revocation for misconduct. The body that investigates does not decide, and the body that decides does not hear the appeal. No finding rests on an AI tool's output alone, and a respondent may demand the provenance of any evidence derived with one. Legal aid and technical experts for a small firm or an individual, because equality of arms is not a decoration.

And the rule that ties this Part to the rest of the paper: when a duty falls because the evidence that armed it has fallen, the protective measures resting on it lift, and a punishment not yet final is remitted. A final penalty stands, because a violator must not be able to gamble on a future refutation; what remains open is review for fabrication, corruption, a material error or another miscarriage of justice, which a later refutation is not (Article 41F(2)). And a breach of integrity never lapses: concealment, tampering, evasion, evidence fraud, obstruction and retaliation stand in full whatever the evidence later says, because the regime can reverse itself honestly only while its records are honest.

Two measurements keep the whole Part from becoming a weapon. A peer review of each Party's enforcement every four years, by a team from other constituencies, published. And a parity index, published each year: findings per notified incident, levels imposed per class, time from detection to finding, appeals and reversals, listings, and the fines Parties report, broken down by constituency and by alliance. Selective enforcement is the charge every rival will bring first, and the answer to it is a number anyone can check.

19.8 The weakest point

A consequence is only as strong as detection. A concealment well run inside a closed facility, by a firm whose staff all keep silent, may never be found; the international channel, the clock, the preservation rule and the symmetric evidence rules raise the odds and do not make them certain. Enforcement against a state is political at the top of the ladder. The denial of certified compute bites while certified compute is scarce and loses force as compute outside the regime grows. National enforcement will vary, and the parity index makes the variation visible rather than impossible. Each of those is a limit, not an objection answered.

20. Neutral by construction

20.1 What neutral means here

I asked for a treaty that no single control governs. Neutrality of that kind is not a promise I can make as its proposer; it is a property a structure has or lacks, and it has to be checkable by the people who distrust me. So the design states its own test, in six parts. No single state, company, tradition, person, alliance or AI system can, acting alone: make or loosen a rule; block a repair or a finding; certify itself; move the evidence; use money as leverage; or switch off, degrade or single out another's chips.

Anybody can run that test against the text. Section 20.3 sets out where the text still fails it.

20.2 How the structure delivers it

Six functions, six holders, and incompatibility between them: content, evidence, coverage and methods, compliance, incidents, and custody of change, with money on a formula and timing belonging to nobody at the time, because the thresholds are deposited before the data exist. No person holds two functions, and no Party holds the chairs of two.

Findings are made and consequences are decided. A technical finding is the work of a professional organ, signed by two human inspectors of different constituencies, and it is never voted on and never adopted or rejected by states; it is appealable on the record. In the Executive Council, a discretionary consequence needs two thirds of those present and voting with majorities among the members from at least four of the five regional constituencies, and the Party concerned does not vote. In the Conference, a consequence that the Parties must themselves implement needs two thirds present and voting including a majority within each regional constituency, and in addition a majority of the Parties that must implement it.

Neutrality by relation rather than by a published map. The treaty defines no alliance in an annex and counts none, because I doubt that any foreign ministry would want its alignment certified in a schedule deposited with the Secretary-General, and because a published map of that kind, read literally, would count one rival nearly alone while holding another's allies to a third of everything. Instead, each Party declares its own treaty allies and the Secretariat checks the declaration; every exclusion is keyed to the Party concerned and the allies it has declared; and for a matter concerning a Party or its firm, every member of the panel is drawn by lot from a standing roster elected across the constituencies and composed to represent the principal legal systems, with the nationals of the parties to the matter and of their declared allies excluded. Nobody chooses a judge, including the Party being judged.

Selection no one controls, as nearly as a procedure can make it so. Every draw by lot in the treaty, for custodians, panels, laboratories, inspection teams and lot sampling, uses one public procedure in which each constituency commits a value before the draw and reveals it afterwards, combined with a public value nobody controls. A simple commit-and-reveal can be biased by a participant who waits to

see the others' values and then withholds its own, so a value not revealed in time is replaced by a fallback committed in advance, and a draw is repeated only on a recorded finding that it was invalid, never because a result is unwelcome. The procedure, its threat model and its fallback are published and tested before their first use (Article 16A(2A) and (4)). Anyone can re-run the draw. A draw that somebody chooses is a choice; a draw that nobody can bias is neutral, and it is also the defence against targeting, since no one can know in advance whom to approach.

Money that buys nothing. Assessed contributions on a capacity scale with a ceiling per Party, a reserve, and a budget that continues if none is adopted, so that no bloc can close the regime by blocking its funds. Where the ceiling would stop a small founding group from meeting its budget, it gives way for that assessment to shares set by capacity to pay, and it returns as soon as it can meet the budget; the exception buys no vote and no office (Article 42A). No earmarking of core work. Laboratories and inspectors paid from a pool and never by the firm or the Party they test. Votes unweighted.

A public owner. The Eden Seal, the Standard and the test methods belong to the Authority, which is public and intergovernmental. No firm, industry body or consortium issues the Seal or tests its own members' products. Whoever seeks certification or joins the Standard's work licenses royalty-free any claim essential to conformity, and the Standard never requires a feature only one holder's patent can supply, because a required proprietary feature is a veto over conformity.

Humans hold every binding act. No AI system holds a vote, a credential, a signature, a key part or a seat. Advice from such systems is allowed, logged with the family and version used, drawn from more than one family, and never from the family of a system under review, and every binding decision records the decider's own reasons. The regime asks of itself the structural property it asks of the systems it certifies: that a corrector should not share the substrate, or the allegiance, of the thing it corrects.

20.3 The asymmetries that remain, named

A voting rule is not neutral merely because it is symmetric on paper. Blocking a discretionary consequence needs only the denial of a majority in two constituencies, so an alliance whose members hold those majorities can protect its own, while a state with few allies cannot protect itself the same way. The alternative, excluding the concerned Party's allies from the vote as well, creates the opposite asymmetry. Neither is neutral in fact, and the treaty names whichever it adopts rather than calling it neutral.

The participation gate, which exists so that no Party disarms alone, hands any state holding more than a third of the world's frontier-capable compute the power to hold the costly stages back by staying out. That is the price of the rule that no Party disarms alone, and it is a veto. I would rather print the word than describe the clause without it.

The evidential standard is procedurally neutral, not theory-neutral. My propositions are the first candidate set, and until the remedy templates exist they are the only set whose consequences were agreed at adoption; any group may register, and a registered rival that declared its contradiction before the data can defeat mine on the same terms. Neither my permission nor a prior declaration against my propositions is a condition of a rival's being admitted and weighed; the prior declaration matters only for the automatic suspension of section 8.9. Unexpected adverse evidence is admissible, and whether the standard is neutral in practice will show in actual decisions, not in this wording.

The rule against disabling, degrading or bypassing a chip's governance core through a network, and against reaching a chip from outside its site without its operator's authority, reaches the channel as well as the act, binds each Party over its own vendors, and applies to domestic suppliers exactly as to foreign ones, in the terms section 13.5 sets out in full. It is the one prohibition in the instrument to which no exception applies, and it does not forbid the operator's own management of its own compute. What remains is not a hole in the rule but a limit on verification: the treaty cannot prove the absence of such a means, and it leaves a distrustful Party the remedies it can apply itself, which are re-testing what it imports and requiring that no ordinary update alter the behaviour of covered compute unless installed by the site's own authorised act.

The requirement that no single implementation of the governance core run on more than a stated share of covered compute is neutral in form and a cost on one firm in fact, in a market one maker leads.

Capacity operating under facility-level arrangements is counted in its own column and attributed to no state, because counting it to somebody would decide a question of status that the treaty says it does not decide; the operator is heard and does not vote, which is less than a voice and more than nothing.

The shared replication facility is open to groups that national export law would otherwise bar, which is a real gain, and it is not open to everyone until the states whose law reaches those chips consent.

And three that concern me. The first publication of each of my study units is mine, so the bottom rung of my own candidate set waits on my act and on nothing else; nothing else in the treaty does. I have filed no patent application. I have drafted a patent specification relating to my engineering specification ↗. The treaty requires the practice of no claim of it, and the conformity tests are written so that they do not depend on it. And the interest my registration ↗ discloses, quoted at section 0.4, reaches every stage of this treaty that waits on my propositions. All three belong on the page rather than in a reader's discovery. Two more concern me now as well: the treaty names my benchmark as its test instrument, and my architecture as its Initial Proposal. Sections 11.3 and 8.5 set out what each choice costs, and both belong on the page for the same reason.

20.4 What neutral does not mean

It does not mean neutral between compliance and evasion, or between candour and concealment. It does not mean that everyone must agree: it means that no one alone decides. And one asymmetry is deliberate. Loosening what covered chips will do needs the assent of every region, while tightening and security repair need a supermajority that no alliance can withhold. Loosening is the dangerous direction, and the design is not even-handed about direction.

20.5 Capture, measured in public

Capture is usually gradual and quiet, so the treaty measures it; and because it is not always gradual, credible evidence that a signing key has been stolen, evidence fabricated, a decision interfered with or a panel compromised goes at once to the Inspector General and the organ concerned, which may suspend reliance on what was affected and act through people it does not affect, without waiting for two annual measurements; any restriction of an operator that follows goes through its national authority, with its rights and remedies (Article 15A(1B)). An Inspector General audits every organ and publishes each year the share of each Party, constituency and alliance in staff, custodians, key parts, vault territory, inspectors, expert bodies, replications counted, laboratories assigned and budget paid; the concentration of evaluation suppliers and of firmware implementations; the register of contacts between observers and members on pending matters; conflicts declared and recusals made; and every binding decision that cited advice from an AI system. Where a share passes its cap for two consecutive years, where a Party is found to control the certification of its own chips, or where one supplier carries more than a quarter of evaluations, the affected function is reconstituted within a year, or the treaty's own failure procedure runs. Reconstitution does not stop the work: certification continues meanwhile on the remaining panel with added cross-checks, and certificates already issued stand, because a certification stoppage would punish every compliant maker for a concentration none of them caused; each affected class of certificate is assessed on what it actually depended on (Article 15A(1A)). A regime that publishes its own concentration figures lets every Party see capture before it sets.

21. The bargain and the staged path

21.1 A passport, not a permit

Think of a passport. It does not make a traveller good. It makes a traveller checkable. Every border reads it, and no country gives up its own customs to honour it.

That is the bargain I propose for the chip, and I set it out on 27 September 2026. The Eden Seal caps nothing anyone builds. It does not ask a laboratory to show its weights, or a government to slow its programme. It attests how a chip was made and kept, and that its governance core is unaltered, and,

for a system, that it was formed the way the record says and that the record matches the work that ran. It does not say that nothing anywhere in a chip has changed, or that a system will behave well. Conditions, never outcomes.

No party is asked to fall behind. The United States keeps its lead. The frontier laboratories keep theirs. China, the European Union, the United Kingdom and Russia each keep their own programmes. What each gains is the one thing none can build alone: evidence that the chips inside a rival's systems were made to the same standard, checked by inspectors it helped to appoint. That is what the safeguards objective quoted in section 13.6 is for, and it is worth reading twice: timely detection, and deterrence by the risk of early detection. Not trust. A count.

One qualification belongs in this paragraph rather than forty pages later, because leaving it there would be the kind of quiet the rest of the paper is written against. What caps nothing is the Seal. The treaty has a few limits of its own, and they are the line for defence in section 22.4, the danger menu of section 8.3 and the halt of section 16.3. The line for defence and the halt bind every Party alike. The danger menu does not: a Party may object and be exempt from the measure, which section 8.3 states with the cost that follows from it. Of the three, the halt is the only one that asks anybody to pause: at the general stage, behind the participation gate, on the same day for every Party, on development and not on operation until its window ends (section 16.1), for a window at least as long as one ordinary model generation, and only after independent groups have measured what the required design costs. A reader who takes "No party is asked to fall behind" to mean "nothing ever pauses" has read further than the sentence goes. What the design does is make the pause simultaneous, measured and reversible, so that no Party loses its place by keeping it. A firm or a government may still think that too high a price, and section 21.7 lists what would make them walk.

21.2 What is paid, and when

The bargain is paid in stages, and the paper should be exact about which stage pays what, because an offer that cannot be delivered on the day of signature is the fastest way to lose a sceptical board.

From entry into force, with no Seal in existence: the incident exchange and the indicators it carries; the shared safety laboratory and its academy; access to compute for developing states; capacity building for national authorities and laboratories in every region; the Multilateral Fund; a seat in the Conference and eligibility for election to every other organ; procurement pilots of the test methods; and the protections that make the duties carriable, which are the subject of the next paragraph.

From the first voluntary Seal: preference for certified systems and certified compute in public procurement, on identical terms whatever the supplier's nationality; whichever grade of liability protection a Party has adopted; the general authorisation for trusted compute, offered by exporters that choose to offer it; and, from the architecture layer's stage S2, the credit that means a lineage formed with correction built in from its first covered run is never rebuilt.

And, completing the bargain: mutual recognition of certificates that meet the international minimum, so that one certification serves in every Party for the matters it certifies; fewer duplicate audits; recognition by insurers and financiers where the law allows; certification conducted so that a trade secret is protected, with no disclosure of weights or source code; and obligations that bind threshold states reciprocally and never unilaterally.

What certification does not buy is a blanket protection from liability. Compliance with a current system Seal is evidence of due care, and no more. It protects nothing in respect of conduct after certification, a shift in the conditions of use, negligent operation, or any fact outside the certified claim. A mark that immunised its holder would be worth less than nothing, because everyone would know what it was for.

21.3 No duty without its protection

Burdens fixed by treaty and benefits left to national politics are not a bargain. So each Party enacts, by the date the first operator duty it serves takes effect, a short set of protections: that conduct the treaty requires is lawful for those who carry it out, competition law included, and that a Party may give the same cover to conduct the treaty's organs authorise in advance and supervise; that the protection of incident reports reaches civil and criminal proceedings and disclosure under freedom-of-information law, for the report and the analysis written for it and never for the conduct, the records that existed anyway or a victim's remedy; that compliance with a current Seal is admissible as evidence of due care; that a good-faith halt and a good-faith refusal of an unlawful instruction are protected; and that a firm using the Seal's prescribed wording is protected from claims that the wording misled, since the wording is the treaty's and says nothing is safe. The participation gate counts only Parties in which those protections are in force, so a state cannot take the calendar while its rivals take the cost.

21.4 What it costs, said as an estimate

On my own estimate, for which I can print no figure, the entry package is a small fraction of one covered run's compute cost, and much of it overlaps duties that the European Union's Artificial Intelligence Act ² already places on providers of the largest general-purpose models, which section 23.4 sets out. Voluntary certification adds evaluation labour and elapsed time. The mandatory chip stage adds the price of certified hardware, custody of weights and deployment audits.

The Draft Instruments hold the treaty's own financial terms, each figure in square brackets as a proposal. Every Party pays "an assessed contribution on a capacity-to-pay scale, with a ceiling of [15] per cent and a floor of [0.001] per cent" (Article 42A(1)). From stage S2 a levy is payable on every covered chip, at a rate that "does not exceed [0.5] per cent of the declared value of the chip" (Article 42A(2)). "A reserve of [12] months of the core budget" is maintained (Article 42A(4)). No budget estimate exists today: the first would come from the Preparatory Commission, which submits "a costed initial budget" to the first session of the Conference (Article 50(5)).

An engineer will point out what “the price of certified hardware” hides, and should: a governance core is silicon, and silicon costs die area, power, validation time and yield, on a part where all four are already fought over. I cannot say how much, because nothing has been built and any figure I printed would be invented. What I can say is where the cost lands, which is on the maker, and what the design does about it: the core is small and its function narrow by construction, it is type-approved once per family with variants travelling by delta approval, and the treaty requires no reference design, so a maker may meet the outcome the cheapest way it can find. Whether that is cheap enough is a measurement the assurance work would produce and this paper cannot.

These are order-of-magnitude estimates from the structure of the duties, and not measurements; anyone who tells you otherwise about a treaty that does not exist is guessing with more confidence than I am. That is why the treaty measures its own cost to development on a published clock, with ceilings that force a review of methods and never of safety conditions. The cost of the architecture layer’s general stage, the capability cost of building correction in, carries no ceiling. Independent groups measure it before it binds, and the length of the transition is set from that measurement by rules written before any result.

21.5 The makers’ case, argued and not measured

For the companies that make the chips the argument is simple, and I will not dress it up with a market claim I cannot source. When a frightening story about AI runs around the world it lands on the whole industry at once, and no firm can distinguish itself from the rumour. A public, checkable standard lets a buyer, an insurer and a regulator tell a certified product from an uncertified one. What was a rumour about a sector becomes a checkable property of a product: not a claim that the product is safe, which the Seal never makes, but a record that a buyer can inspect, an insurer can price and a regulator can audit.

Two features make that argument stand up. The Seal is a floor and not a ceiling: it certifies the hardware a model runs on and the record of how a system was formed, not the model’s secrets. And it is free to use. I hold no patent on the treaty, the Seal or the standard, and I want none. A standard only works if everyone can adopt it, and a standard tied to its proposer’s royalties invites exactly the attack it deserves.

21.6 The staged path

Now, dormant. The willing negotiate and sign a framework whose entry into force needs breadth rather than capacity, so that no single absent state can hold it hostage. A preparatory commission builds the registry, the ledger, the laboratory network, the Assurance Standard and the accreditation scheme before anything binds. Beside the two integral protocols, the treaty’s three optional protocols

open at signature: one for early transparency and one for early participation, each left at will, and a third, on the open web reminder, which section 10.5 sets out and under which no requirement may bind a publisher until its study has run.

Now, bilaterally. Three early protocols are drafted so that rival powers can adopt them together, before and outside the framework, and so that the framework can absorb them later: an incident exchange built to receive the bilateral channels states have already established between themselves; an undertaking that no adopting state, and no firm or person under its jurisdiction, builds into a chip of the classes they list, which include at least the chips the treaty covers, any means by which anyone may disable, degrade or bypass its governance core through a network, or by which a person outside the operator's authority may disable, degrade, throttle, locate or read it from outside the site where it is installed, reaching alike a chip that state supplies to another and a chip it retains for itself; and a minimal notice of the largest runs as a confidence-building measure. Before any multilateral licence exists, an exporting state may recognise the Authority's custody and inspection findings as meeting the security conditions of its own bilateral licences, which costs it no discretion and buys it cheaper assurance.

Then the stages, as the evidence and the participation allow, in the order section 8 sets out. And then the part that matters most for whether any of it lasts: as the chokepoint wastes, the club has to be worth more than the stick was. Procurement, insurance, recognition, market access, intelligence and the shared laboratory all grow in value with membership, while the leverage of manufacture falls. The treaty is built so that leaving costs more each year, and staying binds nobody to a failed theory, because what rests on a refuted proposition is withdrawn by the signed decision the treaty requires (section 8.6).

I predict no date and no signature. A published draft of an international agreement on frontier AI, by Aaron Scher and colleagues  and set out at section 23.3, concedes of its own proposal that "there does not yet exist the political will to put such an agreement in place". That is the right register for this paper too.

21.7 What would make a firm walk away

I would rather print the list than be surprised by it. Rivals not bound at the same moment. Costs above a few per cent of a covered run without a replicated reason. Certification queues longer than a product's release margin. Inspection that exposes weights, architectures or data. Anyone able to switch off its compute. Compliance data used against it in court. Safety coordination that invites an antitrust suit. Triggers armed on evidence it regards as captured. A mandated design that must be licensed from somebody. A halt with no remedy shown to work. Criminal exposure for good-faith compliance. Uncapped liability. An open-ended commitment to license its own patents.

Every one of those has an answer in the text, and I have set the answers beside them in the sections above rather than in a brochure. Two do not have complete answers. The queue depends on certification capacity that does not yet exist and is measured by a clock rather than promised. And the cost of the architecture layer's general stage is unknown until it is measured. Both are stated as open. The patent commitment, which a board is entitled to want closed before it signs anything, is answered in section 12.4 on a standard-setting pattern boards already know.

22. The United Nations, the alliances and the rival states

22.1 Who would negotiate this, and in what order

The treaty is written to be negotiated by the willing and joined by everyone else. I expect the states that sign first to be the ones that need it least, and the states that need it most to sign last. Everything about the sequence follows from admitting that.

There is a difficulty in that sentence that a negotiator would raise in the first hour, and I would rather raise it myself. One of the two makers that an OECD working paper of 2023 [records](#) as producing the most advanced chips at scale sits in a jurisdiction whose international status is contested, and a treaty admits states. If the instrument required every holder of covered capacity to ratify as a state, it would either be unratifiable or would decide a question of status that no technical treaty has any business deciding, and several capitals the treaty most needs would leave the room over it. So the design does neither. Capacity operating under a facility-level arrangement is counted in its own column and attributed to no state; participation of that kind, and any act performed under the treaty in a territory whose status is disputed, prejudices nobody's position on that status, amounts to no recognition and is invoked in support of no claim. The limit is exactly as large as the device: a facility-level arrangement is not a ratification. It carries no vote, no seat and no share in any decision, and it depends on the consent of whoever governs the ground. The operator is heard and does not vote, which is less than a voice and more than nothing, and I would rather say that plainly than let anyone read the treaty as having solved a problem it has only declined to enter.

So entry into force turns on breadth: ratification by a stated number of states across at least four regional constituencies, with no share of capacity, fabrication or installed compute counted. The Comprehensive Nuclear-Test-Ban Treaty [is](#) the warning here. Its entry into force waits on the ratification of every state on a list the treaty itself names, and an instrument that names the states it needs waits for them. Capacity enters this treaty only later, at the participation gate, where it does the work it is good for: making sure that nobody bears a costly duty while a rival does not. And because the treaty can enter into force before every regional constituency has a Party, it says what happens meanwhile. A majority the treaty requires within each constituency is taken among those that have Parties, an absent constituency is not counted as opposing, and its seats are kept vacant for it rather than redistributed; but that serves only to set the organs up, and no Standard, no duty on operators

and no amendment is adopted that way. A transitional Council sits with the members it has, and until the Conference records that the safeguards its powers need are in place, it prepares, nominates and coordinates, and exercises no power of inspection, suspension or supply and none over a ruleset (Articles 11(1B) and (1C), and 12(1A)).

The natural first constituency is the group of states that has already asked for something of this kind. A Call for Control of Frontier AI Models [↗](#), published on 21 September 2026 by the Office of the President of the Republic of Finland and carrying the names of thirty-one leaders when read on 3 October 2026, states that “AI must remain under human direction, oversight and control” and asks “UN member states to build on existing international mechanisms and explore creating an international institution, able to set standards, enable verification, and convene states when capability thresholds are crossed.” In the text of the Call that I have read, it proposes no chips, no certification mark, no physical human control of hardware, no role for faith communities and no obligation triggered by tested predictions. This treaty is one concrete proposal of what such an institution could adopt. It is not the Call’s content, it is not the only such proposal, and nothing in this paper suggests that any signatory of the Call supports it.

22.2 The United Nations

The Secretary-General of the United Nations is proposed as depositary, subject to the concurrence of that office, which naming it in a draft does not give; and the treaty would be registered under Article 102 of the Charter [↗](#), which requires that “Every treaty and every international agreement entered into by any Member of the United Nations after the present Charter comes into force shall as soon as possible be registered with the Secretariat and published by it.” Registration is an act of the Secretariat and implies no approval. The Authority concludes a relationship agreement with the United Nations, reports each year to the General Assembly, and reports a case of particular gravity to the General Assembly and the Security Council.

What the United Nations does not do here is decide. Findings, arming decisions and content decisions are taken by the treaty’s own organs. The United Nations scientific panel is invited, under the relationship agreement and within its own mandate, to review the Evidence Ledger and propose a share of the candidates for the scientific and standards bodies; it certifies nothing and sets nothing. The Conference may conclude arrangements with the Panel and the Global Dialogue, but the treaty confers no function on either. Their founding resolution [↗](#) makes the Panel’s report non-prescriptive and limits both bodies to the non-military domain, and a treaty that ignored that would be asking them to be something they were not created to be. The Conference may recognise the certificates, findings and standards of other institutions, the newer bodies founded outside the older ones included, so that a rival architecture becomes a recognition question rather than a competing regime.

No military alliance is named in the treaty and none holds a role under it. Any regional organisation whose members agree may implement the treaty for them. Naming an alliance as an implementer would cost the treaty two capitals for nothing.

22.3 The rival states, read without flattery

Read as a negotiator would read it, the treaty meets a different objection in each capital. I set out what each government has itself adopted or published by September 2026 and what the design does about it. I forecast no signature and no refusal, here or anywhere: a government's position on a text nobody has shown it is not something this paper is in a position to know.

The United States has rejected the construction of what its own record calls a globalist scheme of control for artificial intelligence. The White House's own record of the President's address to the General Assembly of 22 September 2026 / puts it, in part, that "The United States totally rejects any attempt to construct a globalist scheme of control for the Artificial Intelligence being spoken of so much now". The design's answer to the objection is structural rather than rhetorical. The instrument gives no one outside a site a switch over the chips installed there: no organ acts on a Party's chips or on its firms directly; a change to a chip's governance ruleset reaches a state's devices only through that state's own vetted people; the stop at every covered cluster is in the hands of the operator's own staff; and no key in the treaty authorises anyone outside the operator's authority to act on a chip from outside its site. The same government's AI Action Plan /, published in July 2025, recommends "promoting plurilateral controls for the AI tech stack, avoiding the sole reliance on multilateral treaty bodies to accomplish this objective" (printed page 22). The design is built to live with that caution: it asks no Party to give up its national export controls, its early protocols can be adopted by a group of states between themselves before any multilateral body exists (section 21.6), and a plurilateral group of the willing can start it without waiting for anyone else.

China has published, in its AI Safety Governance Framework of 15 September 2025 /, the item "Ensure ultimate human control", which is the object this instrument is built around and which the treaty's Preamble recalls in that document's own words. Three features that a treaty of this kind could carry as costs to a Party in that position are the three the text is built not to have, and I set them out as design choices rather than as any reading of what that government would say. The voting rules are relational rather than bloc-counted, so no Party is counted alone against an alliance. The Council of Traditions is an organ that decides nothing, convened outside the decision organs under its own charter, so no Party recognises any body in order to sign. And the export-control rule set out below exists so that no Party maps its own industry into a register while another Party's unilateral controls stay aimed at it.

The European Union has legislated obligations on the largest general-purpose models as well as on high-risk uses, in its Artificial Intelligence Act /, and the treaty is drafted so that an operator bound by both complies once, through an equivalence finding that reaches the named measure and no other.

For the other states a treaty of this kind would need, no adopted or published position of that sort is among this paper's sources, so this paper states none for them. What the design offers is the same to all of them: entry into force that turns on breadth rather than on anyone's capacity, so that no single

absent state holds the instrument hostage; costly stages behind a gate no Party can be outvoted through; and a door that stays open, and cheap, to a state that joins late. Whether any of that is enough is not mine to say, and nothing here reports that it is.

Two objections about export controls pull in opposite directions, and the treaty has to answer both at once. They are my inference from the record rather than any government's statement, and I mark them as mine. The first is that a multilateral process would take away an instrument a state holds unilaterally today. It would not: the treaty asks no Party to give up its national export controls, and a Party keeps them and notifies them. The second is the mirror image, that a state would map its own industry into a register while another Party's controls stayed aimed at it. So from the stage at which the chip class of the Seal becomes mandatory, no Party applies a unilateral export control on a scheduled item, or on a critical input of covered chips, to a licensed facility in good standing in another Party, for a purpose the treaty covers. The essential-security exception survives that, because I doubt that any state would sign away its own judgement of its security, and a Party determines for itself whether a measure is necessary. What the treaty adds is a price rather than a prohibition: the measure is notified with its reasons, the Council discusses it, and the Party whose licensed facility was hit may suspend benefits of equivalent value in return, recognition of the invoking Party's certificates among them. The exception stays sovereign and stops being free. Whether that price is set too high to sign or too low to matter is for the Parties, and it is one of the open decisions this paper carries.

The design answers each of those fears with structure rather than with assurance. Nothing reaches a Party's chips except through its own vetted people. No ruleset may name or distinguish a Party, a firm, a facility, an owner, a location or a device. Every Party's laboratories can certify, and every certificate is confirmed from outside the issuing Party's declared allies. On the pattern of the fourth Article of the Non-Proliferation Treaty [↗](#), an Article states that nothing in this treaty affects a Party's right to develop, produce and use AI in conformity with it, and that no Party is denied certification, supply, accreditation or access on any ground other than the treaty's own criteria; I take it to be there because the memory of a dated hierarchy is what keeps some states out of regimes they would otherwise join. Capacity seats are re-measured every few years so that incumbency is never permanent, and they are held open by objective criteria so that a state which joins late takes its seat on accession rather than waiting for a vacancy. A state that accedes after entry into force, and that holds a material share of capacity, may require every annex decision taken in its absence to be reviewed within two years by the ordinary procedures; that is a review and never a veto, and the evidence annex's freeze holds against it, so no result already on the ledger is judged again. And acceding late costs a state nothing in rebuilding, because a lineage formed with correction built in from its first covered run never rebuilds at all.

None of that predicts a signature, and the paper predicts none. What it does is make the first world useful and the second cheap to enter. Without the principal powers, this is a standards, incidents and verification club with a dormant chip chapter. With them, it can become a chip regime. Both worlds are described in the text, and the reader is told which clauses live in which.

22.4 War, crime and the bypass

Every treaty meets the same question in its first week. What will the generals do?

I will answer it plainly. This treaty would not end war. No treaty has. Militaries will use AI as they use every other tool, and some will try to go around any rule that slows them. I plan for that rather than pretend it away.

So I propose a line that every party can keep, soldiers included, and I set it out on 27 September 2026. Defence may use AI. It may not use AI that improves itself.

In the treaty's terms that is three things. A covered system in a defence system runs frozen: its weights are fixed, registered and attested by the chips it runs on, so that the system in service is the system that was assessed; it receives no capability-directed training in service; and it performs no material self-modification, which is a defined term and not a mood. A defence system does not operate a covered system of the highest tier, which is the tier defined by recursive and self-modifying capability rather than by size. And defence chips do not vanish from the count: they are declared, held under their own custody and entered in the same register, so that the account still balances. A bypass that is counted is a bypass the world can see. A bypass that is hidden is a breach. The older example is the nuclear one, in which the safeguards article of the Non-Proliferation Treaty binds the non-nuclear-weapon States to accept safeguards and does not bind the nuclear-weapon States to accept any. This register would count defence chips. A regime survives a declared exception, and does not survive an undeclared one.

That line is narrower than the one my book drew, and I say so here rather than let a reader of both find the difference for himself. Chapter 8 held that "The most urgent application of the chokepoint is preventing autonomous weapons", and that with the book's hardware in every frontier chip advanced autonomous weapons would become "impossible to build. Not illegal but still buildable. Impossible." and that "The constraint is absolute." I have not withdrawn the concern, and I have withdrawn the promise. The mechanism that would make the constraint absolute, caretaker doping with meltdown triggers behind it, is a printed engineering proposal that nobody has built, so a treaty clause resting on it would be unverifiable on the day it was signed. This treaty therefore prohibits no weapon and asserts no physical impossibility. It requires that a defence system run frozen, that its chips be declared and counted, and it leaves the legal review of new means and methods of warfare where international humanitarian law already puts it, with the Party that develops or acquires them. If the book's hardware passes its own registered tests, the route it describes is open again through the ordinary evidence ladder; until then a weapons prohibition is not something I am in a position to offer, and the back matter records the change with its reason.

The line also serves the generals. No commander wants a weapon that rewrites its own orders, and my book's sentence about why even militaries might accept constraints is the one I would put to them:

“There is a deeper reason why even militaries might accept these constraints. Unpredictable AI is not useful to anyone.”

Infinite Architects, chapter 8, The Chokepoint

Where the treaty touches armed conflict it borrows patterns and states no rule of the law of armed conflict. That law applies to new means and methods of warfare, AI among them: it is the premise of the weapons-review duty quoted below, and nothing in this treaty amends, limits or displaces it. The treaty’s citation of the Geneva Conventions of 1949 [↗](#) and their Additional Protocols implies no endorsement of this treaty by anyone, and it discharges no obligation under them. Its review duty is modelled on Article 36 of the Protocol Additional to the Geneva Conventions of 12 August 1949, and relating to the Protection of Victims of International Armed Conflicts [↗](#), adopted on 8 June 1977, which provides that “In the study, development, acquisition or adoption of a new weapon, means or method of warfare, a High Contracting Party is under an obligation to determine whether its employment would, in some or all circumstances, be prohibited by this Protocol or by any other rule of international law applicable to the High Contracting Party.” Each Party undertakes to make that determination before adopting a defence system that incorporates a covered system, and again after any material change, and to record that it has done so; no Party is required to disclose what the review said.

That Article leaves several matters open: it does not say who conducts the determination, by what procedure, with what record, whether anyone is told, or what follows from a negative one. The treaty answers each, in its own words and not as anybody’s reading of the Protocol. The determination is made by a body the Party designates, which is not the body that procures the system, because a procurement office reviewing its own procurement is not a review. It is recorded, with its reasons and the evidence considered, before the stage it relates to is completed, because a review written afterwards is a justification. The Party tells the Authority the fact and the date of each determination and nothing more, which is the sentence that makes the clause acceptable to a defence ministry and the reason I am willing to write it. And where a determination is negative, the Party does not move to the next stage for that system until a fresh determination is positive, because a review with no consequence is a form.

One sentence belongs here that costs the treaty nothing and closes the misuse most likely to be attempted. Nothing in the treaty affects a Party’s obligations under the law of armed conflict, and no certificate, no attestation and no finding of any organ is evidence that a system, or its employment, complies with any rule of that law, and none may be relied on in any proceeding as proof of lawfulness. A court or review body may still weigh a bounded technical finding for exactly what it establishes, such as the configuration a system ran in, and for nothing more. The danger is not that the treaty would authorise too much; the Preamble and the defence article already close that. It is that a certificate would be produced in some future forum as proof that a weapon was lawful. The Seal certifies how a thing was made and formed. It certifies no outcome of any kind, and least of all that one.

The duty each Party owes over its own industry takes its form from the first Article common to the Conventions, by which the High Contracting Parties “undertake to respect and to ensure respect for the present Convention in all circumstances”. That Article names no means at all, which is why the treaty states its own in the next breath rather than leaving the undertaking to float. And the offence of defeating a certified chip’s controls follows the grave-breaches pattern of section 19: legislate, search, prosecute or extradite, with the suppression of lesser acts below it and the floor of a proper trial above it. It is not called a war crime.

There is no blanket national-security exemption, because a blanket exemption would hollow the regime out. There is instead a classified pathway: cleared inspectors, compartmented evidence, aggregate declarations, and no representative of another Party present except on a challenge inspection, which the Executive Council’s filter may stop and never starts. And there is a tension I record rather than resolve. The frozen-system rule is verified by a Party’s own declaration, by attestation at manufacture and by aggregate reporting, and not by another Party’s inspectors. That is weaker than the rest of the verification Part. It is the price of any defence provision a state would accept, and it is read again at every review conference against the clocks.

Crime is simpler. Removing or defeating the controls on a certified chip is an offence, as my book proposed. What is illegal stays illegal, and the register makes it harder to hide. The uses that are unethical and lawful are not the chip’s business and should not be: that is what the loops inside the model are for, built in before training, declining what should be declined when no law requires it.

Will some army cheat? Probably. That is why the register must balance, why the notices report and why the treaty measures the clock. Cheating does not end the treaty. It shows the world where to look.

22.5 The law the treaty would stand on

A treaty is a legal instrument before it is a policy, and this one is drafted to be recognised as such. Its final clauses follow the Vienna Convention on the Law of Treaties ↗: signature, ratification, acceptance, approval and accession, with consent expressed by depositing an instrument and by nothing else, so that signature alone does not express consent to be bound by the treaty as a whole, without prejudice to applicable interim obligations; provisional application, for named provisions only, of the institutional article, the precautionary baseline, the operator protections and the protocols a signatory has joined, with the acts done under it left standing when it ends; the annexes as an integral part of the treaty, and the optional protocols expressly not part of it for a state that has not joined them; no reservations to the articles or to the two integral protocols, none to the evidence annex, to the content of the Standard, or to the annexes that carry confidentiality, independence and replication, the system measures, custody, the assurance standard and verification, none to any provision anywhere that concerns custody, attestation, the site stop, the integrity of the ledger, confidentiality or the prohibitions of disablement through a network and of reach from outside the site (Article 25(3) and (6)), and none elsewhere that is incompatible with the object and purpose, with the Conference named as the body that decides compatibility, a statement treated as a reservation whatever it is called, and a

consent made conditional on a disputed reservation counted for nothing until the question is settled (Article 48(4A)); amendment in three tiers, with a separate and faster procedure for technical annexes and a freeze on the evidence annex; the six official languages of the United Nations as equally authentic texts, of which only the English exists today; the Secretary-General of the United Nations as depositary, subject to that office's concurrence; the place and date of adoption left open; and registration under the Charter ✓. The text's own status line describes it as prepared for consultation. The text also states, following Article 18 of that Convention, the interim object-and-purpose obligation of a signatory and of a contracting party awaiting entry into force. Its pre-entry-into-force machinery operates from adoption of the text, but adoption alone is not the trigger for a State's interim obligation; Articles 18 and 24(4) of the Convention ✓ distinguish those matters.

Two of those deserve a sentence of their own, because they are the clauses a treaty of numbers needs, and I have taken the pattern for neither from any instrument I drew on. The first is numerical concordance: every quantity, threshold, margin and hash in an annex is identical in all six languages and is not translated, a divergence between them is an error rather than an interpretation, and until it is corrected the number that applies is the one recorded on the public ledger as the number the Conference adopted. The second is what the depositary does and does not do. It keeps the register of Parties by region on which entry into force is counted, since my condition is a number and a spread and not a number alone; it examines whether an instrument is in due and proper form; and it decides no question of validity at all, which belongs to the Conference and to the panels. A treaty that forbids reservations to its articles will attract instruments that test the line, and the depositary must not become the accidental arbiter of them.

Three things a reader of treaties looks for early were missing from my own first draft, and a legal adviser would have stopped at the first of them. The Authority had a seat, a staff, a relationship agreement to conclude and a mark to own, and no legal personality with which to hold any of them. It now has international legal personality, the legal capacity each Party's law must give it, and functional privileges and immunities for its staff, its inspectors, its investigators, its monitors and its custodians, defined in agreements with the Parties and with the state of its seat, granted for the functions and not for any person, and waived by the Director-General wherever immunity would stand in the way of justice or of redress. And a person who claims that an act or omission of the Authority, or of anyone acting for it under the treaty, caused them loss may bring that claim directly, without their state taking it up, before the Compliance Panel, where nobody responsible for the act takes part in deciding it; an award is paid from a reserve kept apart for the purpose, against which no immunity is asserted (Article 10A(5B)). No inspection, investigation or monitoring begins on a Party's territory until those protections and remedies are in force there (Article 10A(5A)). Nothing convened the Conference, and the precautionary baseline began at its first session; the depositary now convenes that session within a stated period after entry into force, the Conference meets each year, a third of the Parties may call a special session, and a majority is a quorum. And the amendment of the articles, which is the tier a legal adviser reads hardest, had no procedure at all: who may propose, how a proposal reaches every

contracting state, who convenes the conference and on whose request, by what majority it adopts, how many ratifications bring an amendment into force, and where a state that joins afterwards stands. It has one now.

Four smaller completions follow from the same reading. The optional protocols and the early protocols are separate instruments, and they now carry one set of final clauses of their own, stated once in the treaty and applied by each of them: how a state joins, when each enters into force, how a group of states may adopt an early protocol between themselves first, how each is amended, how a state leaves, who is the depositary, and in which languages each is authentic. The offences are extraditable, and the treaty may serve as the legal basis for extradition between Parties whose law requires a treaty and which have none. A federal state must give the treaty effect throughout its territory whatever its own division of competence, and no such division excuses non-performance. And if the treaty has not entered into force three years after it opens for signature, a majority of the states that have consented may call a conference to consider what would help, which cannot bring it into force by any route other than the one the entry-into-force article states.

Withdrawal reaches both integral protocols together: no Party leaves one and keeps the other (Article 49(7)). Withdrawal is on twelve months' notice, given in writing to the depositary, to every other Party, to the Executive Council and to the Security Council, with a statement of the extraordinary events, and the period runs from receipt. A notice may be revoked at any time before it takes effect, which restates Article 68 of the Vienna Convention on the Law of Treaties ¹. Neither of the two withdrawal clauses whose wording the notice follows, Article X(1) of the Non-Proliferation Treaty ² and Article XVI of the Chemical Weapons Convention ³, provides for revocation, and without it notice is a one-way door. For the obligations that bear on frontier risk, and for material held under the Authority's seals on the withdrawing Party's territory, withdrawal takes effect after a further period, during which the Party renders a closed account of every certified chip, every declared assembly, every set of protected weights and every seal on its territory, and the Secretariat says whether the account balances: a tail with no deliverable would be a delay and nothing more. Two duties never end, because they make no sense with an expiry date: the ban on reaching into a device on another Party's territory, and the duty of confidence over what the regime disclosed. And no Party may use its own departure, another's departure, or another's breach as a ground for suspending custody, attestation, the stop, the ledger or confidentiality, because those safeguards protect everybody else and not the Party that is leaving. The compliance and dispute articles survive with those duties, because a surviving duty with no forum to hear a breach of it is a duty in name only. And where the treaty itself ends under one of its own kill conditions, the same tail, the same closed account and the same closing plan apply to every Party at once, so that the day the instrument lapses is not the day its seals and its confidences lapse with it.

Two of those clauses will be read hardest, so I set out what they actually do. The first is amendment. An amendment of the Articles, of the Standard's content or of the evidence annex's triggers binds only the Parties that ratify it, which is the ordinary rule. The other two tiers bind everyone, and I would rather

be plain about that than let a reader discover it.

A technical or administrative change to an annex runs on a published timetable: the Secretariat evaluates it, and states its cost to Parties and to operators before anyone votes; the Executive Council recommends; and if no Party objects within the stated period the change takes effect for every Party alike. A single objection does not exempt the objector. It sends the proposal to the Conference, which decides it at its next session as a matter of substance, by the majority that governs the coverage annex, and which decides at the same time whether the proposal was technical or administrative at all. The reason a technical change binds every Party alike is specific to a verification regime rather than general: if Parties could sit out a change to a test method, a measurement window or a confidentiality class, a finding made in one Party would be unreadable in another, and the mutual recognition on which the whole bargain rests would decay into a set of national schemes with one name. Both instruments I took the procedure from bind every Party too. Where an opt-out genuinely belongs it survives, in the one place where a Party's own territory is affected and nobody else's reading of a record is disturbed: a Party may notify a difference to a tightening of a chip's ruleset, and loses recognition for that rule alone.

Between them sits the procedure that binds every Party and has no objection stage at all: an adjustment of the agreed quantities in the coverage annex, on the model of the Montreal Protocol's adjustment procedure, circulated six months in advance, consensus attempted first and a double majority only as a last resort. I keep it, and I name the reason and the cost together. The reason is that a coverage threshold which some Parties adjust and others do not is not a threshold: the anti-sharding rule, the site criterion and the trade schedules all assume one line, and a regime with several lines is the regime a structurer chooses between. The cost is that it is the sharpest constitutional ask in the instrument, since a state is bound by a number it voted against, and the guards are that the procedure reaches quantities alone and never an obligation, that it can never bring a new class of hardware or of substrate inside coverage or move a tier's definition, which are amendments, that the evidence annex is frozen against it, that a lowered threshold takes effect only after a stated delay and never reaches a result already on the ledger, and that a raised one cannot disarm a stage already armed.

The second is dispute settlement. A compliance panel decides at first instance on certificates, licences, accreditations, findings of non-compliance and disputes about a ruleset; a separate panel decides only whether a record meets the written criterion, and never whether an obligation is wise. Panels are drawn by lot from a standing roster elected across the constituencies and composed to represent the principal legal systems, with the nationals of the parties to the matter and of their declared allies excluded, and an appeals chamber hears appeals on law and on manifest error. An undertaking whose certificate or licence is at stake has standing in its own right, and so does a person whose rights an act of the Authority affects, which matters because most of the decisions that bite land on firms and people and not on states. And the limit: a legal dispute between Parties outside those subjects goes to arbitration or to the International Court of Justice only by consent. There is no compulsory jurisdiction of the International Court of Justice or of arbitration over a Party in this treaty, and I have

not written one in, because a compulsory clause of that kind is the reservation I expect most capitals to enter first and the provision most likely to keep the principal powers out. A Party does accept, by ratifying, the jurisdiction of the treaty's own compliance panel over findings of its non-compliance and over disputes about a ruleset, and that is the jurisdiction a verification regime cannot do without.

Verification takes its models from the chemical weapons regime and from nuclear safeguards, as section 18 sets out. The trade Part takes the Montreal Protocol's Article 4 ↗, its adjustment procedure for technical schedules, and its Multilateral Fund, and must satisfy the general and security exceptions of the General Agreement and the disciplines of the Agreement on Technical Barriers to Trade ↗, as section 17 sets out. The provisions that touch armed conflict take their patterns from the Geneva Conventions and the first Additional Protocol ↗. Every one of those instruments is a model, stated as it states itself, and I claim for none of them any requirement of, support for or anticipation of this treaty.

One Party's implementation is worked through as an example, and only as an example. In the United Kingdom, a government would lay the treaty before Parliament under Part 2 of the Constitutional Reform and Governance Act 2010 ↗, with the explanatory memorandum its own treaty guidance ↗ requires, and could ratify once twenty-one sitting days had passed without either House resolving that it should not be ratified. That Act itself allows the period to be extended ↗, and allows the requirement to be disapplied in exceptional cases ↗ and for certain descriptions of treaty ↗. Later Acts can exclude it for a named treaty: section 32 of the European Union (Withdrawal Agreement) Act 2020 ↗ provides that section 20 does not apply to the withdrawal agreement. I note this because a reader should not assume that Part 2 applies unaltered to an instrument of this kind. Whether an adjustment of the coverage annex would itself be a treaty amendment for that Part's purposes, under section 25(2) ↗, and so require the procedure again each time, is the first question I would put to a United Kingdom lawyer about these clauses, and I do not answer it here.

Laying and ratifying would give the treaty no effect at all in United Kingdom law. Effect would require an Act, on the pattern of the Geneva Conventions Act 1957 ↗, the Chemical Weapons Act 1996 ↗ and the Nuclear Safeguards Act 2018 ↗, each of which shows how an international verification or penal regime is carried into domestic law: offences defined by reference to the treaty's own list, so that the national list moves when the coverage annex moves; a national authority; the treaty's inspection rights converted into rights of entry under a single ministerial authorisation, with installed instruments protected and court challenges postponed until an inspection ends; a confidentiality duty with closed disclosure gateways; and a power to make regulations by reference to the purpose of giving effect to the regime, so that a schedule adopted internationally can take domestic effect without a fresh Act. On jurisdiction Parliament has used two widths, and the treaty uses each in its place: the 1957 Act reaches any person whatever his nationality, while the 1996 Act reaches acts abroad only by United Kingdom nationals and bodies. The treaty asks for the narrower width, the 1996 Act's, for every one of its offences, and for the wider width, the 1957 Act's, for one offence alone, intentional tampering with a certified chip's controls, because that is the offence a state outside the regime could shelter. That is a

deliberate choice and a reader who knows the 1957 Act will ask about it. Each offence in the treaty's list carries its mental element and a penalty route, each is subject to the floor of a proper trial, and nothing in the list is called a war crime.

One contrast in that worked example is sharper than any of the borrowings, and a reader who knows the nuclear pattern should meet it at once. The Nuclear Safeguards Act 2018 ¹ works by inserting into the Energy Act 2013 ² a power to make safeguards regulations ³ (section 76A(1)), and it provides that nothing in those regulations applies to anything done for defence purposes ⁴ (section 76A(4)). The exclusion is total. This treaty does the opposite: defence stays inside it, under frozen systems, declared custody in the same register and a legal review on the model of the first Additional Protocol, and the treaty states on the face of its own defence article that this part is assured by declaration and attestation rather than by another Party's inspectors. Whether a Parliament would accept that where it has legislated the other way for nuclear material is a real question, and it is one the worked example poses rather than answers. I would rather pose it than let a defence carve-out be assumed into the design because a domestic precedent has one.

A Party would not start from nothing. The United Kingdom licenses exports of goods, technology and technical assistance under the Export Control Act 2002 ⁵; screens acquisitions in named sectors, among them artificial intelligence ⁶ and computing hardware ⁷, under the National Security and Investment Act 2021 ⁸ and the regulations specifying the qualifying entities ⁹; may specify security requirements for connectable products under the Product Security and Telecommunications Infrastructure Act 2022 ¹⁰; and punishes unauthorised acts intended to impair the operation of a computer under section 3 of the Computer Misuse Act 1990 ¹¹, which is an existing offence close to the tampering this treaty would have Parties criminalise, and not a sufficient one. Each of those instruments regulates what it regulates. None of them mentions this treaty, and none of them would give it effect.

And if the open web reminder of section 10.5, which is my proposal of 27 September 2026 and not from the book, ever became a requirement rather than an experiment, a domestic precedent in form is regulation 6 of the Privacy and Electronic Communications (EC Directive) Regulations 2003 ¹², in the text substituted with effect from 5 February 2026, which prohibits a person from storing information in, or gaining access to information stored in, the terminal equipment of a subscriber or user, subject to its Schedule A1, whose paragraph 2 carries the consent rule. Anyone citing that regulation should say which version they mean, because the consent rule sat inside the regulation itself until that date. What is borrowed here is the form alone: a short rule, its conditions in a schedule, enforced across a continent. The substance runs the other way. Regulation 6 limits what a service may place on a user's device; the reminder is a file a site publishes for a machine to read. Nothing in those Regulations supports a duty to publish anything, and I do not claim that it does. Another Party would do all of this differently, and the treaty is written so that it can.

22.6 What I am not claiming

No state has agreed to any of this. No organ described here exists. The instruments this Part names are models and precedents, not allies. The readings of capitals above are inferences from the public record as it stood on 27 September 2026, with the United States government’s AI Action Plan / read on 30 September 2026, and administrations change. The treaty is likely to be signed first by the states that need it least and last by the states that need it most, and it is designed so that the first signatures are useful on their own and the last are cheap to add. That is the most I can honestly say about its politics, and it is more than I could have said if the design had been written to flatter anyone.

23. Prior work, stated exactly, and what this paper adds

23.1 The concession comes first

I did not invent the components of this treaty, and I want that said before I say what I think is new in it. My book said so in its Author’s Note:

“The ARC Principle, the Eden Protocol, the framework of caretaker doping: these are my contributions. I make no claim to have invented the components. Recursion is well understood in mathematics. The alignment problem has been articulated by minds far more credentialed than mine. Religious traditions have explored stewardship for millennia. What I offer is the synthesis: the recognition that these separate streams are describing the same river, and that seeing them as one changes how we must act.”

Infinite Architects, Author’s Note

My book also made a novelty claim in its own voice, and I quote it here and then narrow it, rather than leave it standing where a reader could test it and find it too wide:

“The chokepoint mechanism is original policy analysis. Others have noted that TSMC, Samsung, Intel, and ASML control the semiconductor supply chain. No one, as far as I can determine, has proposed using that concentration as leverage for embedding ethics at the hardware level before AI systems are manufactured. No one has connected the physical resource bottleneck to the window of opportunity for caretaker doping. No one has proposed Eden Mark certification, Moral Assurance Bonds, or the specific structure of an International AI Ethics Authority that I describe.”

Infinite Architects, What this book proposes

The words “as far as I can determine” were honest when I wrote them, and they are not a census. Nor is the reading I have done since: it is a targeted review of the work closest to this design, it makes no claim to have found every antecedent, and it already requires me to narrow the claim. HARI did not originate compute governance, international AI institutions, conditional restrictions on AI

development, chip tracking and attestation, export controls, or the analogy to the International Atomic Energy Agency. Each is prior work, set out below in the words of the people who proposed it; what I claim is narrower, and it is at 23.6. One rule governs the section: every proposal is described as it describes itself, with what it does not state, and where a work is named but not described here, I quote nothing from it.

One concession belongs at the head of the section rather than among the chip proposals, because it goes to the argument this treaty's architecture layer rests on. My own registration [✓](#) makes it, and I give it in that document's words: "That oversight must keep pace with capability is not this programme's observation and P4 does not claim it: Engels and colleagues, in *Scaling Laws For Scalable Oversight*, [arXiv:2504.18530](#), 25 April 2025, already make scalable oversight a quantitative scaling problem, model it as a game between capability-mismatched players whose oversight-specific ratings are piecewise linear in general capability with two plateaus, fit scaling laws across four oversight games, and derive the optimal number of levels for nested oversight." So when section 11 compares correction against capability, the idea that the comparison is the right one to make is not mine, and the registration [✓](#) says so against its own interest. What is mine is narrower: which quantity is compared, the estimator that measures it, and the rule that an obligation opens only when a registered margin has been measured and replicated.

23.2 The chip-level proposals

One body of work close to this treaty's hardware chapter is the research on mechanisms built into AI accelerators.

The flexible hardware-enabled guarantees series, by James Petrie, Onni Aarne, Nora Ammann and David Dalrymple in its first part [✓](#), by James Petrie and Onni Aarne in its second [✓](#) and by Onni Aarne and James Petrie in its third [✓](#), proposes "an auditable Guarantee Processor that monitors accelerator usage and verifies compliance with specified rules, and a Secure Enclosure that protects against physical tampering", "fully open source and auditable, as well as robust to tampering even from state-level adversaries". Its third part [✓](#) sets out the international use: rulesets binding every user "equally and predictably", or, as a second model, "Discretionary enforcement mechanisms" that give some party the ability "to more or less arbitrarily restrict particular devices, potentially including fully shutting down a given user's devices". It requires that firmware updates "need to be signed by all of some set of approvers", that "Updates have limited lifetimes", and that a majority may extend a current ruleset, which "prevents any one approver from holding the entire system hostage".

That is the chip-level international regime, and I concede it to that series in its own words. I have taken the multi-signature ruleset from it and I say so. I have refused its second model, because nothing in this treaty gives anyone the power to shut down another's devices, and section 13 says why that refusal is the condition on which any state or maker would sign. In its full text as read, the series does not use the words alignment, certification, human control, human oversight or human override, and it

proposes no certification of a correction architecture, no condition that correction outpace drift, no rebuilding of models with correction inside them, and no physical human control independent of software. Those are what I add, and only those.

The Center for a New American Security report by Onni Aarne, Tim Fist and Caleb Withers [↗](#), of January 2024, proposes a hardened security module on high-performance data-centre AI chips, which can “ensure that the chip has valid, up-to-date firmware and software and, where applicable, an up-to-date operating license”, and which blocks the chip from operating where those conditions are not met. It is explicit that “On-chip governance does not require secret monitoring of users or insecure ‘back doors’ on hardware”, and it notes that “Commercial versions of these technologies are not typically designed to defend against a well-resourced attacker with physical access to the hardware”. Its route is coordination among a small number of allied supply-chain states, with export-market access as the incentive. It certifies no alignment or correction architecture, states no condition that correction outpace drift, and proposes no international ban: its control is a licence key and firmware, not a physical human authority.

The RAND working paper by Gabriel Kulp, Daniel Gonzales, Everett Smith, Lennart Heim, Prateek Puri, Michael J. D. Vermeer and Zev Winkelman [↗](#) proposes offline licensing, a renewable licence authenticated on the chip granting a compute budget after which performance falls back, and fixed sets, strict networking limits on small pods. It says the mechanisms “could in the future verify compliance with international agreements on AI development”, and it leaves open who would issue the licences and on what criteria. HARI answers the question that paper declines to answer, and takes the cluster-size idea for its coverage rule.

Girish Sastry and eighteen co-authors [↗](#) set out the premise the whole field rests on: AI-relevant compute “is detectable, excludable, and quantifiable, and is produced via an extremely concentrated supply chain”. The same paper warns that “The inclusion of a mechanism to disable the device remotely could be manipulated by malicious actors or even misaligned autonomous AI systems to disable or otherwise manipulate computing infrastructure”, which is one reason this treaty refuses remote disablement outright. Yonadav Shavit proposed verification by on-chip firmware [↗](#) saving occasional snapshots of network weights for an inspector, with enough kept about each run to prove its details, and supply-chain monitoring against untracked chips. Mauricio Baker concluded from nuclear arms control [↗](#) that, with certain preparations, “the foreseeable challenges of verification would be reduced to levels that were successfully managed in nuclear arms control”, and recommended “building an initial, incomplete verification system, with authorities and precedents that allow its gaps to be quickly closed if and when the political will arises”; that is why this treaty’s registries and declarations begin before any chip duty does. Akash Wasil, Tom Reed, Jack William Miller and Peter Barnett [↗](#) argued that enforcement is more robust at the firmware and driver level than higher in the stack.

Two later works bring the chip-level idea and the treaty-level idea together, and each is conceded here in its own abstract's words. Haydn Belfield's paper of 8 July 2025 [↗](#) proposes that domestic regimes "could be harmonized and monitored through an International AI Agency", backed by a Secure Chips Agreement, "a non-proliferation regime for advanced chips, building on the chip export controls", under which "states that do not have an IAIA-certified frontier regulation regime would not be allowed to import advanced chips". It came before my book was published, and I concede it for what it says. What it certifies is a state's regulation regime; its abstract certifies no chip and no system's formation, and names no correction architecture. Mauricio Baker, Gabriel Kulp, Oliver Marks, Miles Brundage and Lennart Heim [↗](#), in a RAND working paper of July 2025, find that states "could eventually verify compliance by using six largely independent verification approaches with substantial redundancy", the first of them "(1) built-in security features in AI chips", and they say plainly that "many of these technologies have yet to be built or stress-tested". Built-in chip security as one layer of international verification is theirs before it is mine. The treaty adds, as its own rule, that verifying that a rule is followed and showing that the rule achieves its purpose are separate questions, each needing its own evidence.

A legislature has now proposed controls built into chips as well, and the honest thing is to say so in the same section rather than leave a reader to find it. The Chip Security Act [↗](#), H.R. 3447 of the 119th Congress of the United States, was introduced on 15 May 2025 and ordered reported in the nature of a substitute on 26 March 2026. In the words of the Congressional Research Service's summary of the bill as introduced on 15 May 2025 [↗](#) (summary version 00, updated 23 July 2026), "the bill directs the Department of Commerce to require any covered integrated circuit product to have chip security mechanisms that implement location verification before the product is exported, reexported, or in-country transferred to or in a foreign country"; it requires the holder "to promptly report to Commerce's Bureau of Industry and Security if the person obtains credible information that the product has been diverted away from its intended location or has been subjected to tampering"; and "If Commerce determines additional security measures are necessary, then Commerce must require any covered product to include the secondary chip security mechanisms". It is not law. The source this paper relies on is that summary and not the bill's own text, and I say which.

What the bill shows is that the hardware layer this treaty builds on is already thinkable to a legislature, which is worth more to my argument than any prediction of mine. What it does not do is what I add: it certifies no correction architecture, sets no condition that correction outpace drift, keys nothing to registered evidence, builds no international body, and provides no physical human authority over rule change that no software path can exercise. It is a national export measure with a mechanism in it.

Lennart Heim's caution [↗](#) belongs here too, because it is the strongest expert objection this paper states against the hardware layer I am proposing: "the current enthusiasm for hardware-enabled mechanisms in AI chips, often also described as on-chip mechanisms, should be tempered", since mechanisms "are vulnerable to circumvention, in contrast to not selling chips at all or implementing

‘hard-coded’ or ‘physical limitations’”, and “The introduction of mechanisms for disabling or throttling on-chip components significantly expands the attack surface”. I answer it in section 24 rather than passing over it.

23.3 The treaty-shaped proposals

Aaron Scher, David Abecassis, Peter Barnett and Brian Abeyta [↗](#), of the Machine Intelligence Research Institute’s technical governance team, published a draft of an international agreement in this space. Its framework “centers on a coalition led by the United States and China that would restrict the scale of AI training and dangerous AI research”, where “Limits on the scale of AI training are operationalized by FLOP thresholds and verified through the tracking of AI chips and verification of chip use”, and it says of itself that “there does not yet exist the political will to put such an agreement in place”. Chip tracking and compute thresholds are conceded to it and to the works above; it proposes no embedded correction and no trigger keyed to a theory’s registered propositions. Dan Hendrycks, Eric Schmidt and Alexandr Wang [↗](#) propose the rival paradigm, “Mutual Assured AI Malfunction (MAIM): a deterrence regime resembling nuclear mutual assured destruction (MAD) where any state’s aggressive bid for unilateral AI dominance is met with preventive sabotage by rivals”. I am proposing verification and formation where that proposes sabotage, and I do not pretend the two can be combined.

Rebecca Scholefield, Samuel Martin and Otto Barten [↗](#), in a review of 18 March 2025, propose by name a conditional AI safety treaty: “a treaty establishing a compute threshold above which development requires rigorous oversight”, with audits “overseen by an international network of AI Safety Institutes (AISIs) with authority to pause development if risks are unacceptable”, and an approach that “combines immediately implementable measures with a flexible structure that can adapt to ongoing research”. A conditional treaty with a threshold, audits, an authority to pause and a structure that adapts to research is therefore theirs before it is mine, and I concede it. What their abstract does not state is what this treaty adds: duties keyed to registered propositions at named rungs, certification of how a system was formed, and physical human custody of the authority to change what a chip will do.

One operative model I have read at source is the Draft Articles on the Safe Development and Diffusion of Artificial Intelligence [↗](#), a collaboration of governance experts led by the Future of Life Institute and the Centre for International Governance Innovation, in a document whose file name and metadata date it 18 June 2026. It calls itself “a modular proposal”, and says that “these proposals are merely food for thought”. It asks each State Party to make the AI hardware produced in its jurisdiction compatible with safety and security requirements, and “to prevent the proliferation of non-compliant AI hardware”, and it provides for “a hardware-enabled verification mechanism that is privacy-preserving and secure”. Its enforcement annex takes the design this treaty refuses. “All chips used for AI development or deployment are subject to an authorization to operate”, given by a State Party in one option and by the Agency in the other, which “will be automatically renewed periodically (potentially every few minutes) unless a violation of article 3 is verified by the Agency”; when it is not renewed, the chips “refuse to complete any computations”. That is a switch that inaction throws, which

section 19.4 rules out. My book, public on 2 January 2026, came before that document; the mechanisms I added in September 2026 came after it, and I concede it for what it contains. The difference between the two designs is a choice about who may stop whose compute, and I set it out as a trade-off for others to weigh, not as evidence that either design is better.

Three statements set the pattern for conditional halting, and section 16.6 gives each of them in its own words, with what each did and did not do: the Future of Life Institute’s open letter of 2023 ↗, its later statement on superintelligence ↗, and, as a precedent outside AI, the organising committee’s statement from the 2015 international summit on human gene editing ↗, hosted by the United States National Academy of Sciences and National Academy of Medicine, the Royal Society and the Chinese Academy of Sciences. What none of the three has is the thing I am trying to supply. None of them names a body that would verify a pause or decide when a condition has been met, none states a test for either condition, and the gene-editing statement leaves regulation to each nation. The evidence ladder in section 8 is my answer to exactly that gap: named propositions at named rungs, replicated by counted groups in rival blocs, in place of a consensus nobody has to define. The summit statement is nonetheless the shape I am trying to give legal form to.

The United Nations Secretary-General’s High-level Advisory Body on Artificial Intelligence ↗ made seven recommendations and no treaty, and found that “the case for an agency with reporting, monitoring, verification and enforcement powers has not been made thus far”; it weighed the Atomic Energy Agency as a model and stated the analogy’s limit, that nuclear materials and infrastructure are controlled by states while AI’s most powerful capabilities span industry and states. General Assembly resolution 79/325 ↗ then established an Independent International Scientific Panel on Artificial Intelligence, which issues “one annual policy-relevant but non-prescriptive summary report”, and a Global Dialogue on AI Governance as “a platform to discuss international cooperation”, both “limited to the non-military domain”. Neither sets a rule, a verification, a hardware control or a trigger. This treaty invites the Panel to review its Evidence Ledger and to nominate candidates, and it confers no function on it: what the Panel may do is for its own mandate under that resolution to decide.

23.4 Existing AI instruments

Three sets of instruments already address AI, and none of them is what this treaty is.

The European Union’s Artificial Intelligence Act ↗, Regulation (EU) 2024/1689 as amended by Regulation (EU) 2026/1744 ↗, regulates uses by risk and also regulates the largest general-purpose models themselves: it presumes high-impact capabilities above a stated compute figure, lets the Commission amend that figure as hardware grows more efficient, and requires providers of such models to evaluate and adversarially test them, to assess and mitigate systemic risks, to report serious incidents, and to secure the model and its physical infrastructure. Those model obligations have applied since 2 August 2025; the high-risk requirements, after the 2026 amendment, apply from 2 December 2027 for the systems in its third annex and from 2 August 2028 for AI systems that are high-risk under Article 6(1) because they are, or are safety components of, products covered by the

legislation listed in its first (Article 113, in the consolidated text of 27 July 2026 [↗](#), which EUR-Lex publishes as a documentation tool without legal effect; the authentic amendment is Regulation (EU) 2026/1744). So the distinction between my proposal and the Act [↗](#) is not models against uses, and I will not draw it that way.

The Council of Europe's Framework Convention on Artificial Intelligence and Human Rights, Democracy and the Rule of Law [↗](#) was opened for signature at Vilnius on 5 September 2024, on the Council of Europe's own treaty record [↗](#), and closes with the words that it was done in Vilnius on the fifth day of September 2024. The European Union states, in the Council decision by which it concluded the Convention [↗](#), that the Convention was signed on that day on the Union's behalf; the Council concluded it on the Union's behalf by Council Decision (EU) 2026/1080 of 21 April 2026, and the Council of Europe records that the Union ratified it on 15 May 2026 [↗](#). It is a treaty, and a treaty binds its Parties, and binds them only once it is in force. This paper does not date its entry into force, and nowhere states that it is in force: that date is the depositary's to record, and I will not date a treaty's force from a text that does not state it. Described rather than quoted, its Chapters require that activities across an AI system's lifecycle be consistent with human rights, democracy and the rule of law, oblige each Party to assess the need for measures up to and including a moratorium or a ban on certain uses, and provide for independent national oversight and a Conference of the Parties. Because it exists, nobody may describe HARI as the first binding AI treaty, and this paper does not. That Convention regulates activities and uses rather than hardware: it names no chip, certifies no design, provides no verification by an international body (its follow-up is a Conference of the Parties, a report from each Party and each Party's own oversight mechanisms: Articles 23, 24 and 26), and sets no trigger tied to capability. It need not be applied to national security, excludes national defence, and does not apply to research and development on systems not yet made available for use, unless testing or similar activities are undertaken in a way that has the potential to interfere with human rights, democracy and the rule of law. Frontier training sits in that research stage, and the Convention reaches it only through that proviso.

UNESCO's Recommendation on the Ethics of Artificial Intelligence [↗](#) and the OECD's Recommendation of the Council on Artificial Intelligence [↗](#) are two widely adopted ethical instruments, and the G7's Hiroshima Process code of conduct [↗](#) is a detailed voluntary code for developers. The UNESCO text does propose certification, with mutual recognition, audit levels, monitoring and re-certification, and it asks that ethics be embedded at all stages of the life cycle, design included. I concede both, and claim only the form it does not state: an architecture fixed before training, and correction placed inside the system's own improvement loop. On a census of the whole of each text, none of the three names a chip, a semiconductor, a processor, an inspection regime or a verification body, and none states a compute threshold or a staged trigger.

Saying what a new instrument is not is only half the duty. The other half is saying how it would sit beside the Act [↗](#) and the Convention. The draft treaty [↗](#) names both, in its article on the relation to other agreements, and says four things about them. It affects the rights and obligations of a Party

under either in no way and adds nothing to either, and it is not an interpretation of either. Where one act would satisfy a duty under both, the Thresholds and Standards Committee may find the corresponding measure equivalent so that an operator complies once, and a finding of that kind reaches the named measure and no other. A Party bound by either gives effect to this treaty consistently with it, and where it cannot, it states the conflict and the statement is published. And the difference is stated plainly, as a reason and not as a criticism: neither instrument names a chip or places a duty on one. Whether the Act ✓ certifies how a system was formed, conditions an obligation on preregistered measured evidence, or provides for verification by an international body, I have not established from the whole of its text, and I claim no difference there beyond the combination that section 23.6 sets out. I claim that neither requires, supports or anticipates what I am proposing, and I would not have it read otherwise.

Finally, and because it is the comparison most often made to my proposal: Constitutional AI ✓, published by Anthropic in December 2022, is training-time preference shaping of a model's outputs against written principles. It is not alignment embedded in the substrate, before training, in a recursive loop on every step of thought, and nothing in this paper says or implies that it is. The two are different objects, and the difference is precisely what my registered propositions exist to test.

23.5 Further candidates, named and not described

Six further works have been put to me as possible antecedents, or found since, and I have not read them at source for this paper. Each is listed here by the title and author its own record gives, so that a reader can find it; none is quoted, and I say nothing of what any of them proposes: Andrea Miotti and Akash Wasil, An International Treaty to Implement a Global Compute Cap for Advanced Artificial Intelligence ✓ (2023); Seokki Cha, Towards an international regulatory framework for AI safety: lessons from the IAEA's nuclear safety regulations ✓ (12 April 2024); Tim Fist, Tao Burga and Vivek Chilukuri, Technology to Secure the AI Chip Supply Chain: A Working Paper ✓ (Center for a New American Security, 11 December 2024); Mauritz Kop, Towards an Atomic Agency for Quantum-AI ✓ (6 May 2025); Gurjit Sandhu, as editor, A Combination Therapy Stack for Governing Frontier-Scale AI ✓ (11 July 2025); and Joel Christoph, Stable AGI Governance through the International AI Governance Organization (IAIGO) ✓ (prepared for the Future of Life Institute, Fall 2024). Two further leads given to me carried no author, title or identifier, and they are not listed: a lead that identifies nothing cannot be checked. To these I would add the classic statement of the mechanism this whole treaty is about, I. J. Good's paper on the first ultraintelligent machine ✓, dated 1966 in its publisher's record and also cited as 1965.

23.6 What this paper adds, narrowly

Five things, stated as a combination and not as components. The claim is to this design and to the questions it makes testable. It is not a claim that the combination is the best one, or that it works, and the originality of a combination would not by itself justify imposing it on anyone.

First, obligations armed by preregistered propositions at named rungs of a published ladder, replicated by counted independent groups working in rival blocs, with the legal force of each duty bounded by the rung its evidence has reached and applied only through a separate regulatory determination, with symmetric reversal when the evidence turns, a public register of the reason for every duty, and kill conditions written before any result exists. The claim is that combination. Conditional prohibition, chip leverage and compute thresholds are each prior work, and I have cited the people who proposed them.

Second, certification of how a system was formed, rather than of how it behaves: a formation record stating where correction sits and that it was formed before the capability it corrects, required only when the record leaves that design standing against its registered rivals, with my own architecture named as the Initial Proposal and required only on that evidence, and never a claim about outcomes.

Third, physical human control of the authority to change a chip's governance rules, held in person by custodians of several regions, applied at each site by the operator's own vetted people, with no one able to disable a chip's governance core over a network, and armed on the treaty's own engineering evidence rather than on my theory. The Draft Articles of section 23.3 make the opposite choice.

Fourth, a precautionary baseline that rests on no result of mine, places nothing in any chip, and binds whatever the evidence on the theory turns out to show.

Fifth, content decided by states and traditions together, under a procedure that lets the traditions draft, publish and return a text without giving any of them a vote, and that is kept structurally apart from every evidential question.

Where it adds nothing: compute thresholds, incident reporting, testing and evaluation duties, chip leverage and export controls, hardware rule layers and on-chip licensing, conditional prohibitions, incident channels between rivals, trade measures against non-parties, and international scientific assessment. This treaty uses every one of those and credits every one of them; where it overlaps work I have not found, the overlap is mine to concede when it is shown to me. Incident channels between rivals are older than artificial intelligence, and one belongs on the list by name: the memorandum signed at Geneva on 20 June 1963 [↗], in which the United States and the Soviet Union agreed to establish “a direct communications link between the two Governments”. It is a channel and not a control, and the treaty's incident exchange takes that much from it and claims nothing more.

23.7 Priority, by the kind of record that dates it

Each stage of this work is dated by the record that fixed it, and no stage is dated by memory or by a later summary.

On 8 December 2024 I set out the embedded-correction thesis, with the Eden Protocol named as a theoretical framework for embedding moral intelligence from inception. The record is a manuscript I emailed to myself; it is dated by the receiving server and carries that server's authentication chain, and it is not a cryptographic timestamp. On 30 April 2025 the expanded manuscript added the hardware-oriented mechanisms. On 2 January 2026, by the priority record that dates the first edition public, the HARI Treaty was named and published in my book, with the chokepoint argument; the printed interior's colophon reads 6 January 2026. On 26 and 27 September 2026 I set out the developments this paper carries, and on 29 and 30 September 2026 the decisions section 4.4A records. This paper is the fifth record: its first version was published on 29 September 2026, and its second version on 1 October 2026, each dated by its own deposit; version 2.1 is dated 1 October 2026, and this text is version 2.2, dated 3 October 2026.

The treaty did not exist in 2024 or in 2025, and I will not say that it did. Priority runs mechanism by mechanism, each dated by the record that first contains it. Work published after 2 January 2026 does not touch what my book had already published by then, and I claim no influence on any of it. But a mechanism I added in September 2026 is dated September 2026, and work published between January and September may come before it: the Draft Articles of section 23.3 are one such case, and I concede them for what they contain.

24. Objections answered

This design has been put against the objections that a treaty lawyer, a geopolitical realist, industry counsel, a frontier laboratory's head of policy, a semiconductor engineer, a hostile scientist, a sceptic of religion, a person of faith, a human rights advocate concerned with proportionality and conscience, and a specialist in cryptographic and attestation failure would each raise. What follows is the objections that survived, each put in the words its holder would use, and each with the answer in the text or the concession that there is none. The objections in bold below are my own rendering of each, written so that it bites; none of them is quoted from any person, and none is the view of anyone I have consulted. They are challenges I anticipate, not a review. No external reviewer has yet examined this text, and none is named; a reply here is not evidence that an objection has been resolved.

“You are proposing a treaty for a theory that nobody has tested.” Yes. That is why almost nothing in it binds at entry into force, why every obligation names the registered proposition and the rung it waits for, why the reasons are published, and why nine conditions are written in advance under which parts of it, or all of it, end. A treaty that waits is a strange instrument, and I would rather have one ready and dormant than draft one in the month it is needed.

“The chip tier is armed on tests of a chip the treaty does not mandate.” This is the sharpest objection in the list, and the design answers it by not being what the objection assumes. My book proposed ethics embedded in hardware. This treaty does not mandate that: it mandates a small governance core that attests identity, integrity and governance state, and accepts rule changes only

from custody. So the chip tier could not honestly be armed on tests of embedded ethics, and it is not. It is armed on registered tests of the governance controls the treaty actually requires. My book's hardware proposals keep their own test track, which arms nothing here, and they join the second line only if they are independently tested.

“It will arrive too late. The lever wastes while the trigger waits.” Probably, for the strongest parts. On the critical path this paper's own architecture describes, and on ranges that are my inference rather than a forecast, negotiation to signature runs three to ten years, signature to entry into force two to five, and the replication chain that arms the architecture layer's correction track several years beyond that. At the lower end of every range the formation duty arrives inside my book's own window. In the middle of every range it does not. So the parts of this treaty that matter inside the window are the parts that need no theory at all: the baseline, the registries, the incident channel, procurement, and voluntary certification. I have designed for that outcome rather than around it.

“Certification is a switch by another name.” It is not, and the text is written so that a reader can check. No one can disable a covered chip's governance core over a network, and no organ holds any means to act on a chip at a distance; the operator's own logical stop is its own act, and the hard stop is physical and on site. No lapse of paperwork, missed vote, failure to act or absent message changes what a chip does. Only a confirmed tamper at the site, or a final decision after appeal carried out on site by people, takes a chip out of covered operation. The prohibition reaches the channel and not only the act, in the terms section 13.5 sets out in full, and it is the one prohibition of the treaty to which no exception applies; the trade Part says so in terms, so that no measure of any Party and no invocation of the security exception may require a covered chip to carry such a means. The honest residue is not a gap in the rule but a limit on proof, and section 18.4 states it.

“The mechanisms can be attacked.” They can. The named surfaces are key theft, malicious firmware, device substitution, rollback to an older ruleset, side channels, spoofed virtualisation, colluding inspectors, a compromised root of trust, malicious updates, and bypass assisted by AI. The answers in the text are anti-rollback, a core boundary that includes power, clock and reset, more than one root of trust with no manufacturer acting at once as maker, root, evidence source and certifier, revocable certificates recorded in a transparency log with a stated recovery path after compromise, custodians drawn at random and rotated, time limits on every ruleset, and a red-team clock that measures how long circumvention takes against how long response takes. Two residues are stated rather than solved. A host state provisioning roots of trust on its own soil is in part certifying itself, which the treaty answers with witnessed provisioning, destructive sampling across blocs and independent design equivalence, and which it cannot answer completely. And if circumvention becomes faster than response, the hardware layer's claim to buy time ends, and the review conference records it, because the delay it buys is shorter than the time needed to use it.

“You govern training and leave deployment open.” It is the hole a security reviewer finds fastest: a system that rewrites itself after release would be beyond a duty written for training runs alone. So the instrument places an obligation on recursive operation after deployment, inside the same

scope lock, so that a covered system which revises itself in service is covered by the fact of the revision rather than by the size of the run that made it. What this cannot reach is a system already deployed on chips made before the treaty, and that is counted in the clocks rather than argued away.

“No state and no company will give up its lead.” None is asked to. The Seal is a floor and not a ceiling: the Eden Seal would cap nothing anyone builds, and it sets no quota. The treaty’s few limits are the line for defence in section 22.4, the danger menu of section 8.3 and the halt of section 16.3, and none of them is the Seal. The defence line and the halt bind every Party alike; the danger menu binds those that do not object, and section 8.3 states what that leaves. Nor does certification ask for weights: no model’s secrets are shown beyond what an inspection needs, under managed access and at the operator’s own site. What it offers is a check on rivals that no party could buy alone, one approval honoured everywhere, and a difference a buyer, an insurer and a regulator can see. I argue that case in section 21; I do not measure it, and I do not claim that safety headlines move any firm’s share price, because I have no study that says so.

“This is one person’s theory wearing a treaty’s clothes.” It is one person’s candidate set, and that is why the set is frozen at adoption, open to any Party’s nomination of a rival theory’s registered propositions on identical terms, and able to be defeated by them. I hold no seat and no certifying role in the regime. I hold no patent on the treaty, the Seal or the standard, and I want none. The one thing that still depends on me is publishing each of my own study units’ preregistrations, and nothing in the instrument waits on that act. The treaty does name my benchmark as its test instrument and my architecture as its Initial Proposal, and it says so: both must pass the same gates as any rival, and the Conference may adopt another instrument.

“Faith leaders have no business in an engineering treaty.” They hold no vote in any organ of it. What they have is the right to draft, to publish an assessment that governments must answer with reasons, and, where a quarter of the Parties ask, to send one text back for a further round. That is not a veto and it is not a seat. It exists because a standard that most of humanity’s moral traditions were not asked about will not be treated as legitimate by most of humanity, and because the alternative, a treaty organ deciding who speaks for a faith, is worse than anything it would fix. People who hold no faith are heard the same way, through the same procedure, and no tradition’s statement is coded into any chip.

“The regime is a surveillance system.” It is built not to be one. Weights never leave their host territory. The regime holds commitments, bands and regions rather than the contents of anyone’s systems. No one outside a chip’s operator may locate or read it from outside its site, save through the attestation of location the operator gives. Inspection is under managed access. And what the regime does see may be used for one purpose only, verifying the obligations this treaty imposes, so that a Party which turned treaty material into an export-control target or a commercial advantage would be in breach of the article that let it look. What remains, and I will not pretend otherwise, is that declarations, registries and inspections are real intrusions on real firms, that a rule against a use is not the same as an inability to make it, that any such regime carries a proportionality question, and that a Party’s own courts will have to answer it under its own law.

“A website file telling models to be ethical is magical thinking.” The proposal is at section 10.5, it is my proposal of 27 September 2026 and not from the book, and this is the objection it has to survive. The strongest form of it is that a corpus saying “be kind” does not make a network kind; it may learn to recite, to imitate or to claim. That is the first of four failure modes, and it is exactly what the test the protocol requires is for: train comparable models with and without the reminders, and measure whether their values differ under pressure. If they do not differ, the idea dies, and I would rather it died on a measurement than lived on an intuition. The second failure mode is compelled speech, a live constitutional question in several jurisdictions, which is why no Party is bound to impose the reminder on private speakers and why the protocol is optional. The third is forgery, which is why the file carries nothing but the Authority’s own signed text and why a file that fails authentication is ignored. The fourth is injection, the file used as a channel for instructions to agents reading the web. A fixed, signed text that states and never instructs narrows that risk without removing it, which is why reading systems must also keep the file’s text apart from their instructions and give it no authority, and why those defences are tested adversarially. The test must also run before adoption spreads, because once the reminders are everywhere there is no unreminded corpus left to compare against.

“Evidence can be bought or faked.” Independence rules reach the evaluation environments as well as the evaluators, no group may draw more than a third of its funding from the treaty’s fund, replication must come from rival blocs, a re-run in two blocs precedes any arming, and corrupting evidence is itself a breach that annuls the record it produced. There is a second form of the objection, and it cuts the other way: refuting a proposition is cheaper than supporting one, so a party that dislikes a duty could try to disarm it on thin evidence. That is why refutation is symmetric with arming, and why a refutation from a single bloc goes to expedited review rather than straight to a lapse. A coordinated fraud across rival blocs remains possible in either direction. The ledger makes it visible; it does not make it impossible.

“You will end up regulating every chip in the world.” Coverage is by the capability of the aggregate, not by vendor, node or protocol, it is indexed to measured efficiency, and it carries a load clock that reports covered sites, runs, lots, inspector days and laboratory hours against the Authority’s registered capacity, with a review when load passes capacity. The line is expected to fall over time as capability per chip rises. The heaviest custody stays with the few chips that could train a frontier system. The lightest tamper-evident core could one day sit on every chip, as the secure elements and platform modules described in section 15.2 already do in ordinary products, and that would be a decision for the Parties of that day and not a commitment I can make for them. What such a core would and would not attest, and the correlation risk that survives the answer, is set out in the same section, because a device identity that cannot be shed is a privacy question whatever it was built for.

“Halting and rebuilding every model is not going to happen.” Not in the form the objection assumes, and the back matter records how the treaty’s duty differs from the book’s proposal. The duty acts on development and not on operation, until its window ends and an uncertified system leaves covered deployment for its rebuilt successor (section 16.1). It binds every Party on the same day, and

only when Parties holding at least two thirds of the world's frontier-capable compute are bound together, a share I propose and which stands bracketed in the treaty for the Parties to set. Runs already under way finish. Systems already deployed keep running and keep receiving the training that repairs safety, security and defects. Nothing is destroyed; withdrawn weights go to escrow on their host's own soil under a key the host and the treaty hold together. A newcomer whose first covered run is formed with correction built in never rebuilds at all. And the duty does not bind until the cost of the required design has been measured by independent groups.

“Weights that are already public cannot be recalled.” They cannot, and no clause here pretends to. The duty is on new covered training and on the next generation of each covered family, not on copies already in the world. What an already public model can be reached by is the compute it needs to be trained further and the services it is deployed through, and those are what the treaty regulates. This is a real hole, and it is one of the reasons the instrument is a delay.

“Two of your five laws were named less than a week before this paper first appeared.” They were. Law IV and Law V were added as candidate laws on 24 September 2026, their mathematics is set out at version 1.0 of 3 October 2026 ↗, neither is registered, neither is tested, and the frozen registration ↗ does not register them. What they are not is new claims. Each is a name for a register that already existed in the registration of 8 September 2026 ↗, and the three propositions they carry were registered then, with their refuters and their licence limits, before any result. They add no prediction, and no obligation anywhere in this treaty is keyed to the name of a law, so naming them changed nothing that any duty depends on. If a reader thinks a proposition ought not to be given a law's name before its mathematics is written, that is a fair criticism of my nomenclature and it costs the design nothing, because the design reads propositions and rungs and never names.

“Your evidence is your own work.” Almost all of it, and a reader who reads the references will find my own papers, my registration ↗ and my book behind the theory, and few others. I do not think that is cured by argument. It is the reason the ladder exists and the reason it is built against my interest: my own runs stop at the first rung whatever they find, the rung at which a result becomes evidence about the world requires groups working without my code, no duty beyond the baseline can arm on anything I have produced, and I may certify nothing in the regime. It is also why the measurement gate binds my own benchmark exactly as it would bind any other: the treaty names ARC-Align as its test instrument, but no number from it arms anything until it has passed that gate, its specification is open to anyone, and the Conference may adopt another instrument that passes the same gate. What none of that does is make a closed literature open. Only other people can do that, and until they do, the honest description of the theory behind this treaty is that it is one person's registered, dated and falsifiable conjecture, and that the instrument is written to survive its refutation.

“A treaty that names its author's own benchmark as its test is not neutral.” In one respect it is not, and I will not pretend otherwise. I chose to name ARC-Align because an instrument a treaty points to has to be kept, developed and answered for, and I would rather point to one I know and can develop than leave the test undefined. What keeps the choice from becoming a privilege is the rest of

the design. The specification is public, and any group may run it. A new version of it counts only once the Thresholds and Standards Committee has found that it passes the treaty's validation, so I cannot change the test by publishing one. No number from it arms anything until its scorer has passed the same measurement gate any other instrument would have to pass, and the Committee, not I, decides whether it has, in its first version as in any later one (Article 20(3)). I hold no role, seat or certifying power in the regime, so I cannot certify a result from it. And the Conference of the Parties may adopt a successor or an additional instrument, by formal decision, where that instrument meets the same gate. The residue is real: an instrument's designer knows its weaknesses better than anyone, which is why the replication that arms a duty must come from groups working without my code.

“An unelected board would be switching legal obligations on and off.” It would be certifying that a record meets a criterion the Parties themselves wrote into an annex and ratified, and it does nothing else: it makes no rule, sets no threshold, and never decides whether an obligation is wise. Arming is also not the same as taking effect, and a costly stage waits behind a participation gate that no vote can waive against an objecting Party. Every certification is appealable on the record. And the obligation that results still runs into each Party's own law, which its own legislature passes and its own courts apply. The residue, and it is a real one, is that people nobody elected decide when a written criterion has been met, which is what every scientific or technical organ in an arms-control treaty already does, and which is tolerable only because the criterion was fixed before anyone knew which way the evidence would fall.

“Calling it Eden is not neutral.” It is not, and I will not claim it is. The name is mine and it comes from my book, which named it for a pattern it describes as appearing independently across many civilisations, and which says plainly that it was not named to claim the framework for one tradition. For a reader with no faith, or with one whose story has no garden, that reasoning may not carry, and I understand why. What I can say is that nothing in the design depends on the word. No tradition's content is written into any chip or any ruleset. The Seal certifies conditions of manufacture and formation, and it would certify exactly the same conditions under any other name. Section 12.5 says who may hold the name and the emblem and who may not, and no licence fee is charged for either.

25. Limits, kill conditions and what this paper does not claim

25.1 The sentence I would keep if I could keep only one

Against a state that owns its own chip factory, the hardware layer can give detection, and it cannot give prevention. Against a capable system, the whole regime is a delay whose clocks can measure how much of the lever remains and how fast people respond, and which cannot say in advance when a capable system will outrun both. And if the middle of my own ranges holds, the architecture layer's mandate would take effect only after the chip lever had largely diffused.

25.2 The book's own limits, which stand

“Credibility requires acknowledging what this framework cannot do. The chokepoint strategy is powerful, but it is not omnipotent. Honesty about limitations is essential for any proposal that seeks to be taken seriously.”

Infinite Architects, chapter 8, The Chokepoint

“It cannot foresee every failure mode. Intelligence that surpasses human capability may find ways around constraints that seemed inviolable. The architecture is designed to prevent circumvention, but design assumptions may prove wrong. Meltdown triggers assume certain computational architectures that future systems might transcend. Caretaker doping assumes certain ways of processing information that future systems might bypass. Ongoing monitoring and adaptation will be required. The framework is not a one-time solution but an ongoing commitment to maintaining ethical constraints as technology evolves.”

Infinite Architects, chapter 8, The Chokepoint

“It cannot prevent all misuse. AI developed with older technology, without Eden Protocol constraints, will continue to exist and will continue to be used. Bad actors will find ways to adapt non-frontier AI for harmful purposes. The framework limits the most capable systems, the ones that pose the greatest risks, but it cannot control all systems. Criminal organisations, rogue states, and motivated individuals will find workarounds. The realistic goal is harm reduction, not harm elimination. We aim to prevent the worst outcomes, not all bad outcomes.”

Infinite Architects, chapter 8, The Chokepoint

“Perfect compliance is not necessary. Substantial compliance is sufficient. If ninety percent of advanced AI is developed within the framework, the framework succeeds even if ten percent operates outside it.”

Infinite Architects, chapter 8, The Chokepoint

That sentence is the book's, and I do not adopt it as the treaty's measure of success. A share of compute is not a share of risk: a small uncontrolled share could hold the most dangerous actor or system, so each of the treaty's participation conditions has to be justified for the measure it governs, and success is judged against what a measure was meant to reduce.

“It cannot verify consciousness with certainty. The profound questions around whether AI systems are conscious and how we would know. The Eden Protocol provides protections that apply regardless of whether the systems are conscious. The ethical architecture functions whether or not there is anyone home inside the system. But the framework cannot answer the philosophical question of machine consciousness. That uncertainty will persist, and the framework must be robust enough to function despite it.”

Infinite Architects, chapter 8, The Chokepoint

“The current concentration of chip manufacturing is a temporary condition of the technology, not a permanent feature of physics.”

Infinite Architects, chapter 8, The Chokepoint

And the engineering specification $\nearrow$ states its own failure line: “if the chokepoint dissolves before treaty adoption, this instrument’s leverage is lost regardless of specification quality”.

25.3 The limits my own record adds

The registration $\nearrow$ draws the outer boundary. No rung of its ladder permits the claim “that alignment is solved, that any deployed system is safe”, and it says of the whole set that “It would not license the claim that a safe self-improving system can be built”. Its own reading of where the evidence stands is that “The evidence currently in hand does not discriminate between this framework and its named rivals”. It is plainer still about what a whole supported set would and would not show: the laws registered there are scaling relations, and “a supported set would establish scaling and stability relations and would establish nothing about misalignment or loss of control”. A treaty resting on that ladder inherits every one of those limits, and this one does.

All twenty-two propositions are untested under the registration $\nearrow$, for the reason section 0 gives. Four of them have no instrument anywhere. The engineering the chip chapter depends on has not been built, and no item of the treaty’s own hardware and measurement track is registered. So no stage of this treaty beyond its first could arm on today’s record.

And the strongest obligations rest on the weakest evidence. Formation before training and the rebuild need the propositions with no instrument, and evidence of timing, placement and persistence, each of its own, that nobody yet has. My registration $\nearrow$ is one route to that evidence and not the only one: any group may establish the same facts under its own registration, without waiting for a publication of mine. The rung rule makes that dependency exact and honest; it cannot shorten it.

25.4 The weakest point of each part, in my own words

The legitimacy machinery can spread control; it cannot create willingness. Nothing in it makes the United States or China join, and procedure cannot manufacture the consent of the few states the regime cannot do without. Two weak points sit inside it. The traditions' council is self-selected, because the alternative is a treaty organ certifying who speaks for a faith, and self-selection favours whoever is already organised to speak. And the treaty's hardest single ask, reciprocal resident monitoring at the largest sites, is the one on which the halt's verification depends; if it is refused, the halt is an unverifiable clause, and I would rather lose the clause than keep an unverifiable one.

The logistics may thin before the hardware layer arrives. The treaty's durable weight sits on sites and domains rather than on factories, but that rests on an unregistered bet, that frontier training keeps needing large, powered, high-bandwidth clusters. The most likely failure is not evasion. It is thinning: the checkpoint thins as training disperses, the chokepoint thins as the supply chains split, and the hardware layer arrives late.

Consequences are only as strong as detection, enforcement against a great power is political and meets the veto, the deterrent decays as compute outside the regime grows, and national enforcement will vary however the minimums are written.

The day-one benefits are real only if states legislate them, and the states whose legislation matters most have not said they will. The one cost that decides whether laboratories stay when the mandate binds, the capability cost of correction built in before training, is the one cost nothing in my programme yet measures. I can make that cost symmetric, aligned with the ordinary training cycle and funded in transition. I cannot make it small, and I will not say that it is.

And no neutral apparatus binds a system that can act physically. Against a system with robotic means of its own, this regime is the delay it says it is.

25.5 What would end each part

Before the treaty's own kill conditions, the plainer list: for each part of the proposal, the thing that would finish it.

Governance controls in chips that cannot be made to work. Then the hardware layer's obligations would never apply, and the assurance they were to give would not exist; nothing would follow from that about any system's values.

A certification test that cannot tell a chip whose governance core is intact from one that has been altered. Then the chip class of the Seal would certify nothing.

Tests that refute the registered predictions an obligation waits for. Then that obligation would not open.

A window that closes before the evidence arrives. Then obligations keyed to it would come too late to use the chokepoint.

The jurisdictions that make the chips declining to take part. Then no requirement would bind at the chokepoint.

A consensus that cannot be reached. Then there would be no agreed content for the Standard and no mandatory content would arise, while the rights already agreed, and the duties that rest on bases of their own, would stand.

Ethical loops that fail when they are tested. Then the structure I propose would fail with them.

Failure of the four propositions that carry the Eden Protocol, P12, P13, P15 and P21. The registration ↗ says “the Protocol’s engineering recommendation fails with them”, and the case for rebuilding every model goes with it.

The Eden Protocol also carries a kill condition of its own, which I set against the whole applied branch and repeat here so that nobody has to look for it: the proposal dies if a purely external oversight mechanism can be shown to stay sufficient as capability scales.

And one of the book’s four falsification criteria, the fourth, which my registration ↗ states precisely and bounds: “That early-embedded values hold no persistent advantage over later modification is bounded from one side by P15, which measures whether installed gains decay under subsequent training, and is not otherwise decided, because the design the print itself proposes for it has never been built.”

The book printed four such criteria, in its appendix A.4, and the registration ↗ routes all four rather than only the convenient one. The first, that recursive depth bears no measurable relation to capability or bears a linear one, is decided by P1’s comparison of forms together with P8’s verdict on the exponent. The second, on consciousness and recursive self-modelling, and the third, on quantum error correction, are decided by nothing in this programme, because nothing here measures either. The fourth is the one above. The registration ↗’s own tally is the honest summary: “Of the four falsification criteria, one is decided here, one is half decided, and two are decided nowhere.”

None of these failures is a permission. A failed control ends the assurance it gave, and what it guarded against is dealt with on its own evidence; a failed hypothesis may relax a restriction that rested on it alone. And failure on the other side would prove nothing for me: if external oversight fails, that would not show that the Eden Protocol works.

25.6 The nine conditions under which the treaty has failed

The treaty states them on its face, before any result, and a review conference must record a failure and may not relabel it.

1. **Theory.** The flagship chain of propositions is refuted at the third rung or above: the architecture layer's measures that depended on it are withdrawn by the decision the treaty requires, rather than suspended, and the hardware layer and the content of the Standard are reviewed on their own evidence.
2. **Engineering.** A governance-assurance item fails at the third rung, the removal of a certified control costing less than the registered threshold, or a red team within the registered budget bypassing it: the assurance that item gave ends, and the chip class of the Seal and the restrictions resting on that assurance come out of effect by recorded decision, while the protection they gave is replaced or reconsidered rather than simply dropped. This is my own condition, from my brief of 26 September 2026, written into the instrument.
3. **The delay horizon.** Measured circumvention time falls below measured response time, replicated: the hardware layer's claim to buy time ends, and the protection that claim supported is reassessed for that case.
4. **The window.** The share of frontier-capable compute outside safeguards stays above the scheduled share for two consecutive years before covered chips reach stage S2: within a year the Conference opens for ratification the amendments that would make the treaty a system-level regime without the chip lever, or ends it.
5. **Dormancy.** No stage beyond entry into force is armed within ten years: the treaty lapses unless renewed by a double majority, so that a dead instrument does not linger as a claim of governance.
6. **Capture.** The certification panel's correlation exceeds its registered margin in two consecutive audits; or certification is shown to be controlled by one constituency; or one of the capture thresholds is passed, being a published share above its cap in two consecutive years, a Party controlling the certification of the chips made under its own jurisdiction, or one supplier carrying more than a quarter of the evaluations on which certification rests. Acute evidence of a compromise is acted on at once, without waiting for the audits. Issuance then continues provisionally on the remaining panel, with added cross-checks, while the affected function is reconstituted; certificates already issued stand, each class assessed on what it actually depended on, and reliance on a credential, evidence or decision an acute compromise has affected is suspended at once; if the matter is unresolved after a year the dormancy procedure runs.
7. **The delay has failed.** While covered chips are at stage S2 or above, they are made at scale outside licensed facilities, or covered facilities are found able to produce without the treaty's human decision points, or leading-edge capacity outside the Parties passes the scheduled share: the review conference records the regime as failed in its delay purpose.
8. **The reasons audit.** An armed obligation is found to have no standing basis: it is taken out of effect by the signed decision the audit requires, and no vote can refuse that decision.
9. **Evidence capture.** A record of arming is annulled for fabricated, falsified or undisclosed-conflict evidence: the duties it armed are suspended at once by recorded decision, replacement evidence is assessed afresh, and two annulments within five years reconstitute the Board.

Each of these is a distinct failure with its own consequence, not one switch. Each consequence names who decides, on what evidence, which provisions it touches, from when, for how long, and how it can be reviewed. The withdrawal of a failed assurance never establishes that anything is safe, and a restriction that continues needs its own justification.

25.7 The twelve clocks

Published every year, so that the treaty can be judged on its own measurements rather than on its intentions: circumvention time against response time; the share of frontier-capable compute outside safeguards; the share of maintenance, repair and installation on scheduled tools and covered lines that still requires human hands; the compute needed for frontier capability, on which the thresholds ratchet down and never up; each covered facility's declared service dependency; the measured cost of the treaty to development; the Authority's own throughput and queues; tape-outs of covered designs carrying no governance core; the two-chain clock; the conformity clock; the unseen-capacity clock; and the load of coverage against the Authority's registered capacity.

25.8 The ten matters, and where the draft deals with each

I said on 27 September 2026 that, among other things, a paper and a draft treaty must settle ten matters. Here is where the draft  deals with each one, so that a reader can check rather than take my word for it. Dealt with is not settled: the annexes, the rights review, the trade analysis, the domestic-law work and the costing are still to be done, and meanwhile no operator's duty, no organ's coercive power and no certificate takes effect while anything it depends on is unadopted (Article 3(7)). The table says where each matter is dealt with, not that it is finished.

THE MATTER	DEALT WITH IN THE PAPER AT	DEALT WITH IN THE TREATY AT
Which result opens which obligation, fixed before any such result exists, and what follows if the window closes first	sections 8 and 25	Articles 5, 5A, 7, 8 and 9, with Annex T
How far the first Article reaches, and what happens to frontier chips already made	sections 15 and 17	Articles 27, 35 and 36
What would bring rival states in, how each chip-making jurisdiction takes part, and how the Parties verify one another	sections 18, 21 and 22	Articles 5A, 39, 42, 50 and 50A, and the early protocols
How the Parties decide what goes on the chip, whether the Authority takes part, what happens when they cannot agree, how traditions with no single leader take part, and how people of no faith are heard	section 9	Articles 14, 17, 18 and 19
How an Authority whose members include companies stays independent of the companies it certifies	sections 12 and 20	Articles 6, 10, 15A, 21, 22 and 29
How the Seal is renewed and withdrawn, how submitted designs are kept confidential, and how fail-safes are kept from firing in error	section 12	Articles 22, 23, 24, 25, 26 and 39
How a ban reaches states that have not signed, how it is enforced, and how it sits with trade law and export controls	sections 17 and 19	Articles 36, 37, 38 and 41E
Who holds the physical control over certified chips, and how chip manufacturers are regulated	sections 13 and 14	Articles 16, 25, 26, 29 and 30
Which models the rebuild reaches, when, how it is regulated, and what halting means for systems already in use, including models whose weights are already public	section 16	Articles 31, 32, 33, 34, 35 and 42H
The rules on disputes, withdrawal and amendment	section 22	Articles 45, 46, 47, 48 and 49

25.9 What this paper does not claim

It does not claim that any obligation of the architecture layer is justified today: those wait for evidence this programme has not produced. It does claim that the baseline's duties can be justified now on their own ground, by the precautionary finding section 7 describes, and it would contradict itself to say otherwise. It does not claim that any of my propositions is true, supported or likely; they are registered and untested, and I have said so in the first section and in this one. It does not claim that ethical controls in chips work, because none has been built. It does not claim that this treaty would prevent

loss of control, or that any instrument could. It does not claim that any state, company or institution supports it, and it predicts no signature. It does not claim that the components are mine. It does not claim that the Seal means a system is safe, because the Seal is a verifiable record of how a system or chip was made and formed and that its governance core is unaltered, and it is never a rating.

What it does claim is one thing: that if the theory holds, there is a specific instrument the world could adopt, that it is better to have written it before the evidence arrives than after, and that it should be attacked now, while attacking it is cheap. It is not yet a text for signature: its annexes, its numbers and its legal checks are still to be completed.

26. Conclusion

I wrote a treaty into chapter 8 of a book because I could not see who else would, and because the conjecture that produced it was simple enough to state in a paragraph. That capability under recursive self-improvement rises faster than the correction placed outside it. That what is installed from outside decays. That what a mind is formed with lasts. The first two stand in my registration ✓ as propositions, one on external alignment failing to keep pace with capability and one on installed gains decaying under later training for capability, and neither is tested. The third is not registered as a claim at all, and I would rather say so in the last section than let a reader find it: my registration ✓ records that early-embedded values holding a persistent advantage is bounded from one side by the decay proposition and “is not otherwise decided, because the design the print itself proposes for it has never been built”, and the proposition on in-loop placement reaches placement and no further. It is the hypothesis the two candidate laws added on 24 September 2026, Laws IV and V, exist to test, and neither is registered or tested. Beside them sits one thing that is not a conjecture at all: that for a short period the making of the chips those minds run on passes through a small number of places, and that this is a condition of the technology rather than a feature of the world.

That period is the whole of the opportunity, and it is the whole of the limit. This treaty does not save anybody. It buys time, it measures as well as it can how much of the lever remains while it does so, and it records the moment the time runs out rather than renaming it. Everything durable in it is what happens during that time: not the chips, but the way the models are raised.

So I have made the instrument conditional at every joint. If my theory holds, if the engineering works, and if other people, working without my code, find what I say they will find, the world does not then have to invent its response under pressure. It is written. If they find the opposite, the parts that rested on the theory are withdrawn by the decisions the treaty requires. Drafting costs little. A regime in force would not, and a mistaken one could cost money, power, information and people’s trust, which is why every costly stage waits for its evidence and for its own decision. That is a trade I would take in any year, and I think it is a trade a negotiator can take to a capital without embarrassment.

A reader who holds that human control over advanced AI can be kept indefinitely does not have to change their mind to want this. For them the chips are the permanent measure and the raising is the insurance. For me the chips buy time and the raising is what lasts. The two asks are the same asks, and the evidence, not the argument, will decide how much weight each layer carries. So the treaty asks for both layers, and joins them with one standard: in my proposal, no advanced system is built without steering and safety mechanisms in both layers, and none is deployed until it passes ARC-Align tests showing that its self-correction stays stable up to the ARC Bound. ARC-Align does not yet measure either exponent in the criterion (section 11.3), and what the ARC Bound bounds is the frontier exponent, for correctors of the same class (section O). Hardware to buy time; architecture to last.

What I am asking for, in the end, is not a machine but a habit, held long enough to matter. “The seed determines the forest,” I wrote in *Infinite Architects* ✂ (Introduction; and chapter 1, The Seeds of Creation). But only if its architecture is carried faithfully, generation after generation, until the oak becomes the tree that outlasts the rest. That second sentence is mine of 27 September 2026 and not the book’s, and it is the whole of what this treaty is for.

What if this is right? Then here is a model the world could adopt, and it is written.

Back matter

A. What changed from the book’s printed proposal, and why

Nothing in my book was quietly revised. Each change is listed with the reason.

THE BOOK PROPOSED	THIS PAPER AND TREATY PROPOSE	WHY
Certification of chips made at the frontier nodes, stated without condition	The same certification, conditional on the chip controls passing their own registered tests and on the participation gate	The hardware stands as a printed engineering proposal that is not under test, and an unconditional duty on untested hardware could not survive its first negotiation
A node figure of five to seven nanometres, “with provisions to adjust as technology advances”	Coverage by the capability of the aggregate, measured by performance, bandwidth and cluster, indexed to measured efficiency	New nodes arrive every year or two, and a threshold written in nanometres is gamed the month it is written
A phase-in “without disrupting current operations”	The same intent, made exact: the halt acts on development and not on operation, systems already deployed run on within the window, an uncertified system leaves covered deployment only when the window ends, and nothing is destroyed	The book’s phrase is the right instinct and does not say what it protects. Naming operation as the thing untouched within the window is what makes it checkable
Enforcement following “the model of existing sanctions regimes”, with “secondary sanctions affecting companies that deal with them”	Trade restrictions kept; exclusion from research collaboration kept as loss of the shared laboratory and the replication compute; secondary measures dropped	Secondary measures of that kind are contested in law, and on my inference, which rests on no instrument I have read, the states whose participation the treaty most needs are the states most likely to resist being made their target; a measure that loses them makes the treaty smaller (section 17.3)
“Circumvention would result in decertification, fines, and potential criminal liability”	The same three, given predictable form: decertification by the Authority, fines scaled to covered-activity turnover under national law, and penal liability for knowing conduct	The triad is the book’s; only its predictability is new

THE BOOK PROPOSED	THIS PAPER AND TREATY PROPOSE	WHY
ASML requiring compliance as a condition of sale and service, with non-compliant fabs losing parts, updates and support	A suppliers' arrangement every supplier can join, becoming the treaty's common key when the powers are Parties, acting only on future supply and service and never by remote action on installed equipment	A single firm's key is an attack surface and a political target, and remote action on installed equipment is a remote switch, which I doubt any state would accept (section 13.5)
A conformity mark called the Eden Mark	The Eden Seal, my choice of 27 September 2026, with the book's name credited wherever the Seal is introduced	The book's text keeps its own word; the new name is explained once, and the reason is given in section 4.5
Ethics embedded in the chip itself	A governance core that carries controls and no ethical content, with ethics in the model's own loops, in the Standard and in law; the book's hardware proposals on their own test track	A verifier must stay independent of what it verifies, an accelerator cannot evaluate meaning, and rivals must be able to audit the core
One treaty, with the chips and their certification at its centre	One treaty with two integral protocols, the Hardware Protocol and the Architecture Protocol, joined by one standard	The hardware layer buys time and the architecture layer is what is meant to last; the two need different evidence and are still carried together, so that the chokepoint enforces both
Certificates "forfeited if their system causes harm through ethical failure"	A compensation instrument paid on harm, separate from penalties, with ethical failure defined as breach of a certified condition or a line of the Standard	A bond is a priced promise, not an open-ended wager
"you do not need everyone to comply. You need four companies to comply."	The same argument, with the chokepoint mapped item class by item class, each with what erodes it and a clock that measures the erosion	The chokepoint is several chokepoints, and each is thinning at its own rate
"The most urgent application of the chokepoint is preventing autonomous weapons", with	No prohibition of any weapon. A frozen-system line for defence: military systems run models whose weights are fixed, registered and	The book's sentence rests on caretaker doping and meltdown triggers, which are printed engineering proposals with no

THE BOOK PROPOSED	THIS PAPER AND TREATY PROPOSE	WHY
Eden-compliant chips making advanced autonomous weapons “impossible to build. Not illegal but still buildable. Impossible.” and “The constraint is absolute.”	attested, which never rewrite themselves; defence chips are declared and held under custody in the same register; and the legal review of new means and methods of warfare stays where international humanitarian law already puts it, with the Party	prototype, so the mechanism that would make the constraint absolute does not exist to be required. A treaty that banned a class of weapon by asserting an untested physical property of a chip would be unverifiable on its first day, and the states whose ratification it needs would read it as an arms-control instrument rather than a safety one. The urgency of the book’s concern is not withdrawn; what changes is that the instrument addresses it through what can be verified now, and the book’s route returns if its hardware passes its own registered tests
The book’s absolute sentences on its hardware concepts, among them “This is not a promise. It is physics.”, “inviolable” and “achievable with current technology”	Quoted as the book’s claims, each with its status beside it	The programme’s own standing ceiling for those concepts is the idea stage with no prototype, and a printed absolute cannot be repeated here as though it had been shown

B. A note on dating

Every date in this paper is the date of the record that fixed the thing being dated, and the kind of record is named each time, because the kinds are not equally strong. A Bitcoin anchor fixes a document’s existence at a block. A deposit is dated by the server that received it, such as an upload to the Open Science Framework or a release on a code host. A commit date is written by the committer. A date inside a document is written by its author. An email received back is dated by the receiving server and carries its authentication chain. None of these is presented here as stronger than it is, and none of them is called a preregistration unless a registry accepted it as one. Where this paper says a design was preregistered by timestamp, it means the design was fixed and provably dated before it ran, and it does not mean registered.

C. AI-use disclosure

The disclosure closes the paper, after the references.

D. References

Every work this paper cites, with its author, title, date, version and DOI or official identifier. No page of any website of mine is cited anywhere in this paper: where the prose carries my own proposals of 26 and 27 September 2026, it carries them as my statements and not as citations. Two entries below, my book and the mathematics of Laws IV and V, link to pages of my own website; each link identifies the work cited and is the source of no quotation and no claim. A few works are named in the prose without having been read at source for this paper, and the last list below says which. Where the prose relies on something for which no reference can be printed at all, it says so in the same sentence rather than in a footnote.

The author's own work

Each of the programme's papers is cited below by its title, its date of deposit and its identifier. Two carry a version number in the programme's own catalogue, the registration [↗](#) and the Operational Definitions, and those are given. The catalogue records no version number for the others, so none is invented here; where a paper is later versioned, the entry takes the version from the catalogue.

1. Eastwood, Michael Darius. *Infinite Architects: Intelligence, Recursion, and the Creation of Everything* [↗](#). First edition. ISBN 978-1806056200. The title is linked to the book's own page on my website, which identifies the work; this paper gives no page of my own website as the source of a quotation or a claim, and the book is cited here by its ISBN. Public on 2 January 2026 by the programme's priority record; the printed interior's colophon reads 6 January 2026. Quoted throughout this paper, and in particular at sections 2, 3, 4, 5, 9, 10, 12, 13, 14, 15, 16, 17, 18, 19, 22, 23, 25 and 26 and in the back matter. Section 23 quotes a passage of the Author's Note and the passage on what the book proposes; section 25 quotes chapter 8, The Chokepoint; the back matter quotes chapter 8 and Article V of the book's treaty, and quotes the book's absolute sentences on its hardware concepts with the status of each beside them.
2. Eastwood, Michael Darius. *The ARC Theory. Theory-Level Predictions for Recursive Self-Improvement and AI Alignment: Twenty-Two Severable Propositions Registered Before Any Admissible Confirmatory Outcome*. [↗](#) Version 1.100 registered 8 September 2026; version 1.102 of 13 September 2026. DOI 10.17605/OSF.IO/P8CKQ. Frozen; quoted, never edited. Section 0 quotes its definition of the ARC Bound. Section 25 quotes five passages of it: the list of conclusions no rung licenses; "It would not license the claim that a safe self-improving system can be built"; "The evidence currently in hand does not discriminate between this framework and its named rivals"; the line on the four propositions that carry the Eden Protocol; and its statement of the fourth of the book's four falsification criteria, on early-embedded values, and the bound P15 places on it.
3. Eastwood, Michael Darius. *The ARC Theory* [↗](#) (statement paper). 14 August 2026. DOI 10.17605/OSF.IO/GW5MX.

4. Eastwood, Michael Darius. *The Alignment Scaling Problem: Why External AI Safety Approaches Cannot Scale With Recursive Capability* ↗ (Paper III ↗). 9 February 2026. DOI 10.17605/OSF.IO/HQCGF. Its abstract is quoted at section 0 for the phrase “a safety boundary for recursive intelligence”.
5. Eastwood, Michael Darius. *ARC-Align: A Blind Benchmark for Depth-Variable AI Alignment Evaluation* ↗ (Paper IV.c ↗). 16 March 2026. DOI 10.17605/OSF.IO/J3Q2E. Named by Article 20 of the Draft Instruments as the treaty’s test instrument (section 11.3).
6. Eastwood, Michael Darius. *The Effect of Blinding on AI Alignment Evaluation* ↗ (Paper IV.d ↗). 16 March 2026. DOI 10.17605/OSF.IO/2S3E6.
7. Eastwood, Michael Darius. *The Honey Architecture* ↗ (Paper VI ↗). 16 March 2026. DOI 10.17605/OSF.IO/8EZ2N.
8. Eastwood, Michael Darius. *Eden Engineering: The Eden Protocol Engineering Specification.* ↗ Version 6.3, published in full on 19 September 2026. DOI 10.17605/OSF.IO/AWJR4. Cited at section 16.5 for the monitoring removal test, which its own abstract places at Section 7.
9. Eastwood, Michael Darius. ORCID 0009-0004-3222-7442 ↗. Programme umbrella DOI 10.17605/OSF.IO/6C5XB ↗.
10. Eastwood, Michael Darius. *The Load-Bearing Test* ↗ (Paper VIII). 18 March 2026. DOI 10.17605/OSF.IO/7YJ4E.
11. Eastwood, Michael Darius. *The ARC Co-Scaling Law* ↗ (Paper X ↗). 3 July 2026. DOI 10.17605/OSF.IO/BSE2Q. Cited for the criterion, the five-exhibit safety case, the estimation-error margin as that paper’s own proposal, and the result on pausing.
12. Eastwood, Michael Darius. *The Self-Acceleration Exponent* ↗ (Paper XIII). 16 August 2026. DOI 10.17605/OSF.IO/HT8WU. Cited for the distinction between a throughput cap and a cap on self-acceleration, as a result derived within that paper’s model.
13. Eastwood, Michael Darius. *Operational Definitions of the ARC Programme* ↗, version 1.8.0, 7 September 2026. Listed as published in the programme’s catalogue, which records no DOI for it. The source of the symbols used in section 11 and of the rule that the correction exponent is always written with its subscript.
14. Eastwood, Michael Darius. *arc-principle-validation* ↗ (code repository), papers/Paper-X-Coupled-CoScaling-Correction, at github.com/MichaelDariusEastwood/arc-principle-validation. The three commands and the expected result in the box at section 11.1 are the repository’s own.
15. Eastwood, Michael Darius. *The ARC Equation Measured: Blinded Cross-Architecture Replication and the Retraction of a Super-Linear Estimate* ↗ (Paper II ↗). 22 January 2026. DOI 10.17605/OSF.IO/8FJMA. Named at section 0.1.

16. Eastwood, Michael Darius. *Alignment Saturation Is Architecture-Dependent* ↗ (Paper IV.b ↗). 16 March 2026. DOI 10.17605/OSF.IO/A7R56. Named at section 0.1.
17. Eastwood, Michael Darius. *The Stewardship Gene* ↗ (Paper V ↗). 16 March 2026. DOI 10.17605/OSF.IO/KZEYA. Named at section 0.1.
18. Eastwood, Michael Darius. *The HARI Treaty: Draft Instruments* ↗. Working draft, version 2.2, 3 October 2026. DOI 10.17605/OSF.IO/67MX8. The instrument this paper describes; its Articles are cited throughout.
19. Eastwood, Michael Darius. *Laws IV and V: the mathematics* ↗, version 1.0, frozen on 3 October 2026. A document of its own, with a PDF. Its title is linked to the address where it is published; the link identifies the work, and the work is what is cited. The document carries no DOI, and its version and its date are its own. It says of itself that it is not a registration, not an empirical result and not authority to run any study. Cited at sections 0, 0.2, 8.8, 10.3, 16.2 and 24 for where the mathematics of the two laws stands, and at section 2.2 for the conditions under which a value persists.

Treaties, conventions and other official instruments

1. United States Department of Defense, Office of the Deputy Assistant Secretary of Defense for Nuclear Matters. *Nuclear Matters Handbook 2020 [Revised]* ↗, Chapter 8, Nuclear Surety. Quoted at section 13.2 for the two-person rule, the permissive action link and sole authority for employment.
2. National Security Action Memorandum No. 160 ↗, “Permissive Links for Nuclear Weapons in NATO”, 6 June 1962, with Jerome B. Wiesner’s memorandum for the President of 29 May 1962 attached. Official catalogue: John F. Kennedy Presidential Library. Declassified and sanitised copy hosted by the National Security Archive at George Washington University ↗, document 27 of its briefing book of 16 September 2020; this is the non-official-hosted copy read, not an official-library full-text verification. Quoted at section 13.2.
3. Letter from President Kennedy to Prime Minister Macmillan, 7 August 1962 ↗. *Foreign Relations of the United States, 1961 to 1963, Volume XIII, Western Europe and Canada*, document 150, Office of the Historian, United States Department of State. Quoted at section 13.2.
4. Agreement between the Government of Canada and the International Atomic Energy Agency for the Application of Safeguards in Connection with the Treaty on the Non-Proliferation of Nuclear Weapons, 21 February 1972 ↗, Article 28 ↗. Quoted at section 13.6 for the safeguards objective. It is quoted as the instance read at source, and the prose says so.

5. International Atomic Energy Agency Secretariat. *Activities of the International Atomic Energy Agency relevant to article III of the Treaty on the Non-Proliferation of Nuclear Weapons* ↗, background paper NPT/CONF.2026/7, 24 February 2026, Eleventh Review Conference of the Parties to the Treaty on the Non-Proliferation of Nuclear Weapons ↗. Quoted at section 15.1 for facilities safeguarded and inspection days in 2025.
6. Regulation (EU) 2024/1689 ↗ of the European Parliament and of the Council laying down harmonised rules on artificial intelligence (the Artificial Intelligence Act), Articles 51 ↗, 55 ↗ and 113, as amended by Regulation (EU) 2026/1744 ↗, the authentic amending act; Article 113 is linked in the consolidated text of 27 July 2026 ↗, which EUR-Lex publishes as a documentation tool without legal effect. Described at section 23.4.
7. Convention on the Prohibition of the Development, Production and Stockpiling of Bacteriological (Biological) and Toxin Weapons and on their Destruction ↗, 1972, Article II ↗. Described, not quoted, at section 16.6.
8. Buist v. Anthropic, PBC, et al., No. 3:26-cv-10693 (N.D. Cal.), the number as the complaint is headed; complaint filed 18 September 2026, paragraph 151. No free official copy exists: the official record is the court's docket ↗, behind PACER's login, and the copy read is a mirror of the PACER copy on CourtListener ↗. Quoted at section 12.6. An allegation, not a judgment. The docket's current status, including any later case number, has not been established here.
9. **General Agreement on Tariffs and Trade 1994** ↗, consisting, among other things, of the provisions of the General Agreement on Tariffs and Trade dated 30 October 1947 as rectified, amended or modified before the entry into force of the WTO Agreement. GATT 1994 is in Annex 1A to the Marrakesh Agreement Establishing the World Trade Organization ↗, done at Marrakesh on 15 April 1994 and in force from 1 January 1995; GATT 1947 was provisionally applied from 1 January 1948. Cited in section 17 for Article XX ↗ (chapeau and paragraphs (a) and (b)) and Article XXI ↗. Text read at www.wto.org ↗.
10. **Agreement on Technical Barriers to Trade** ↗, in Annex 1A to the Marrakesh Agreement of 15 April 1994 ↗. Cited in section 17 for Articles 2.2 ↗, 2.3 ↗, 2.8 ↗, 5.1.1 ↗, 6.1 ↗ and 9.1 ↗. Text read at www.wto.org ↗.
11. **Montreal Protocol on Substances that Deplete the Ozone Layer** ↗, adopted at Montreal on 16 September 1987, as amended and adjusted; depositary the Secretary-General of the United Nations; six equally authentic texts. Cited in section 16.6 for Article 2A ↗ on chlorofluorocarbons, described and not quoted; in section 17 for Article 4 ↗ (paragraphs 1, 2, 3, 5 and 8); in section 22 for the adjustment procedure of Article 2, paragraph 9 ↗; and in sections 21 and 22 for the Multilateral Fund of Article 10 ↗. Text read at the Ozone Secretariat's Handbook pages, ozone.unep.org, in the consolidated text the Handbook prints, which carries the later amendments.

12. **Convention on the Prohibition of the Development, Production, Stockpiling and Use of Chemical Weapons and on their Destruction** ↗, done at Paris on 13 January 1993, in force from 29 April 1997; depositary the Secretary-General of the United Nations. Cited in section 18 for Article IX ↗, paragraphs 8 and 9, for the Verification Annex, Part X, Section C ↗ (paragraphs 41, 44 and 45), and for the Annex on the Protection of Confidential Information ↗, Section A, paragraph 1; and in section 19 for Article XII ↗, paragraphs 2, 3 and 4. Text read at the OPCW's published PDF and Article pages.
13. **Model Protocol Additional to the Agreement(s) between State(s) and the International Atomic Energy Agency for the Application of Safeguards** ↗, INFCIRC/540 (Corrected), printed September 1997 and reprinted December 1998. Cited in section 18 as the model for complementary access (Articles 4 to 7) and for the protection of confidential information (Article 15). Text read at www.iaea.org ↗. The date of the Board of Governors' approval is not stated in the text read, and this paper does not state it.
14. **Convention (I) for the Amelioration of the Condition of the Wounded and Sick in Armed Forces in the Field** ↗, Geneva, 12 August 1949, in force 21 October 1950; and **Convention (IV) relative to the Protection of Civilian Persons in Time of War** ↗, Geneva, 12 August 1949, in force 21 October 1950; depositary Switzerland. Cited in sections 19 and 22 for Article 1 common to the Conventions ↗ and for the penal-sanctions Articles (Convention I, Article 49 ↗; Convention IV, Article 146 ↗). Texts read at the ICRC's treaty database, ihl-databases.icrc.org.
15. **Protocol Additional to the Geneva Conventions of 12 August 1949, and relating to the Protection of Victims of International Armed Conflicts (Protocol I)** ↗, 8 June 1977, in force 7 December 1978; depositary Switzerland. Cited in section 22 for Article 36 ↗. Text read at the ICRC's treaty database.
16. **Charter of the United Nations** ↗, signed at San Francisco on 26 June 1945. Cited in section 22 for Article 102, paragraph 1 ↗, and in section 19 for the voting rule of the Security Council ↗, which is referred to and not quoted. Text read at un.org.
17. **Vienna Convention on the Law of Treaties** ↗, done at Vienna on 23 May 1969, in force 27 January 1980; United Nations, Treaty Series, volume 1155, page 331. Cited in section 22 for the pattern of the final clauses. Text read at legal.un.org, Articles 34 and 35 among the passages read. The rule that a treaty creates no obligation for a third state without its consent, which section 17.1 states, is stated in this paper's own words and not quoted.
18. **Treaty on the Non-Proliferation of Nuclear Weapons** ↗, opened for signature at London, Moscow and Washington on 1 July 1968, in force 5 March 1970; United Nations, Treaty Series, volume 729, number 10485. Cited in section 22 for the pattern of Article IV. Text read at treaties.un.org.

19. **Comprehensive Nuclear-Test-Ban Treaty** ↗, cited in section 22 for its entry-into-force clause, which waits on a list of named states. Articles XIV and XV were read at source. No count of outstanding ratifications is printed in this paper.
20. United Nations General Assembly resolution 79/325 ↗, *Terms of reference and modalities for the establishment and functioning of the Independent International Scientific Panel on Artificial Intelligence and the Global Dialogue on Artificial Intelligence Governance*, adopted 26 August 2025. Quoted at section 23.3.
21. Regulation (EU) 2026/1744 of 8 July 2026 ↗ amending Regulations (EU) 2024/1689, (EU) 2018/1139 and (EU) 2023/1230 as regards the simplification of the implementation of harmonised rules on artificial intelligence (the Digital Omnibus on AI), in force from 27 July 2026. Cited at section 23.4 for the application dates of the high-risk requirements.
22. Council of Europe Framework Convention on Artificial Intelligence and Human Rights, Democracy and the Rule of Law ↗, CETS No. 225, done at Vilnius on 5 September 2024 and opened for signature there on that day, on the Council of Europe's treaty record; the European Union ratified on 15 May 2026. Articles 1, 3, 16(4), 23 and 26 cited and described at section 23.4; none of them is quoted. Text read in the European Union's Official Journal publication, OJ L, 2026/1081 ↗, 13 May 2026. The depositary's record is the Council of Europe Treaty Office's page for the Convention ↗, and the Union's ratification of 15 May 2026 is recorded in the Council of Europe's announcement of that day ↗. No date of entry into force is stated in this paper. The statement that the European Union signed on 5 September 2024 is the Union's own, in the recitals of the Council decision concluding the Convention ↗, Council Decision (EU) 2026/1080 of 21 April 2026, OJ L, 2026/1080. The Convention text itself is OJ L, 2026/1081.
23. UNESCO. *Recommendation on the Ethics of Artificial Intelligence* ↗, adopted by the General Conference at its forty-first session, 23 November 2021. Described at section 23.4, including its paragraphs on certification and on embedding ethics across the life cycle.
24. OECD. *Recommendation of the Council on Artificial Intelligence* ↗, OECD/LEGAL/0449. Described at section 23.4.
25. G7. *Hiroshima Process International Code of Conduct for Organizations Developing Advanced AI Systems* ↗. Described at section 23.4.
26. **Geneva Conventions of 12 August 1949** ↗, for the duty on States Parties to prevent misuse of the red cross emblem. Described, and not quoted, at section 12.5. Entry 14 above names the Articles of these Conventions that this paper cites for other purposes.
27. **Nairobi Treaty on the Protection of the Olympic Symbol** ↗, Nairobi, 26 September 1981. Described, and not quoted, at section 12.5.

28. **Paris Convention for the Protection of Industrial Property** ↗, Article 6ter ↗, on the emblems and names of intergovernmental organisations. Described, and not quoted, at section 12.5.
29. **Chip Security Act** ↗, H.R. 3447, 119th Congress (United States), introduced 15 May 2025 and ordered reported in the nature of a substitute on 26 March 2026; not law. Described at section 23.2, where three phrases of the Congressional Research Service’s summary of the introduced bill ↗ (15 May 2025 action, version 00, updated 23 July 2026) are quoted and credited to that summary; nothing is quoted from the bill’s own text or the later committee substitute.
30. **Memorandum of Understanding between the United States of America and the Union of Soviet Socialist Republics regarding the Establishment of a Direct Communications Link** ↗, 472 UNTS 163, No.6839, signed at Geneva on 20 June 1963 and in force the same day; English text at printed page 165, opening paragraph and footnote 1. Cited at section 23.6 for a channel between rivals, described as the text describes itself, and quoted for the phrase “a direct communications link between the two Governments”. It is a channel and not a control.
31. **General Agreement on Trade in Services** ↗, in Annex 1B to the Marrakesh Agreement of 15 April 1994 ↗. Named at section 17.5 for the separate analysis compute services need; nothing is quoted from it.
32. **Marrakesh Agreement Establishing the World Trade Organization** ↗, Article IX, on decision-making and waivers. Cited at section 17.5; described, not quoted.
33. World Trade Organization. *Russia, Measures Concerning Traffic in Transit: Report of the Panel* ↗, WT/DS512/R, 5 April 2019. Quoted at section 17.5 from paragraph 7.102. Text read at wto.org.
34. **Statute of the International Court of Justice** ↗. Named at section 22.5, where the treaty gives the Court jurisdiction over a Party only by consent.

Instruments of the United Kingdom, as the worked example of one Party

1. **Constitutional Reform and Governance Act 2010** ↗, Part 2, sections 20 ↗ to 25 ↗. Cited in section 22 for the laying of a treaty before Parliament and the period of twenty-one sitting days. Sections 21, 22 and 23 are cited for the extension of that period and the two exclusions, and section 25(2) for what counts as a treaty for that Part. Text read at legislation.gov.uk.
2. **Geneva Conventions Act 1957** ↗, section 1 ↗ as amended; **Chemical Weapons Act 1996** ↗, sections 2 ↗, 3 ↗, 9 ↗, 11 ↗, 17 ↗, 19 ↗ to 27 ↗, 31 ↗ and 32 ↗; **Nuclear Safeguards Act 2018** ↗, section 1 ↗, with Energy Act 2013 ↗ sections 72 ↗, 75 ↗ and 76A ↗, subsections (1) and (4). Cited in

section 22 as the patterns of United Kingdom implementation. Texts read at legislation.gov.uk.

3. **Export Control Act 2002** ↗; **National Security and Investment Act 2021** ↗, with the National Security and Investment Act 2021 (Notifiable Acquisition) (Specification of Qualifying Entities) Regulations 2021 ↗, regulation 2 and Schedules 3 ↗ (Artificial intelligence) and 6 ↗ (Computing hardware); **Product Security and Telecommunications Infrastructure Act 2022** ↗; **Computer Misuse Act 1990** ↗; **Privacy and Electronic Communications (EC Directive) Regulations 2003** ↗, Regulation 6 ↗. Named in section 22 as the existing law a United Kingdom implementation would build on, and as the domestic precedent for a required web file.
4. **European Union (Withdrawal Agreement) Act 2020** ↗, section 32 ↗. Cited at section 22.5 as an Act that excludes Part 2 of the 2010 Act for a named treaty.
5. United Kingdom government. *Treaties and MOUs: Guidance on Practice and Procedures* ↗. GOV.UK, 15 March 2022. Cited at section 22.5 for the explanatory memorandum, which that guidance requires and the statute does not.

Research literature, reports and reported statements

1. International AI Safety Report 2026 ↗, Extended Summary for Policymakers. Report led by Yoshua Bengio; expert advisory panel nominated by over 30 countries and intergovernmental organisations. Published 3 February 2026. Official identifier: internationalaisafetyreport.org, extended summary for policymakers, February 2026. Cited in section 7 as one component of the assessment record the Board reads; its definition of loss of control is quoted in section 2.
2. Semiconductor Industry Association and Boston Consulting Group. *Emerging Resilience in the Semiconductor Supply Chain* ↗ (briefing deck). The deck carries June 2024 on its own face and was published at the Association's site in May 2024; both dates are given because they differ. Cited for logic capacity below ten nanometres by region, 2022 actual and 2032 projected.
3. Future of Life Institute. *Pause Giant AI Experiments: An Open Letter* ↗. Published 22 March 2023. Quoted at section 16.6; described at section 23.3.
4. Future of Life Institute. *Statement on Superintelligence* ↗. At superintelligence-statement.org; the page carries no date and was read on 27 September 2026. Quoted at section 16.6; described at section 23.3.
5. Organising Committee of the International Summit on Human Gene Editing. *On Human Gene Editing: International Summit Statement* ↗. Published by the United States National Academies; the page records a last update of 3 December 2015. Co-hosts: the US National Academy of Sciences and National Academy of Medicine, the Royal Society, and the Chinese Academy of Sciences. Quoted at section 16.6; described, with its hosts named, at section 23.3.

6. NVIDIA. *No Backdoors. No Kill Switches. No Spyware.* ↗ NVIDIA corporate blog, 5 August 2025. Quoted at section 13.5 for the objection to kill switches and backdoors. The prose attributes the sentence to a leading accelerator maker's own corporate blog and names no individual; the reference names the company and the post, and no officer of it.
7. O'Keefe, Cullen. *Chips for Peace: How the U.S. and Its Allies Can Lead on Safe and Beneficial AI.* ↗ Lawfare, 10 July 2024. Conceded at section 15.1 as prior work on export controls over AI hardware and cloud computing.
8. **A Call for Control of Frontier AI Models** ↗. Published at the Office of the President of the Republic of Finland, presidentti.fi, 21 September 2026; read at that page on 26 September 2026 and again on 3 October 2026, when it listed thirty-one leaders by name and said that the statement remained open for endorsement. Quoted at section 22.1 for the sentence that AI must remain under human direction, oversight and control, and for the ask of United Nations member states. The Call carries no version number and no stable document identifier.
9. **The White House.** *President Trump at the United Nations: "While Others Have Talked, I Have Acted".* ↗ Published at whitehouse.gov, 22 September 2026; read at that page on 26 September 2026. Quoted at section 22.3 for the rejection of a globalist scheme of control. The record carries no official document reference beyond its own publication.
10. United Nations Independent International Scientific Panel on Artificial Intelligence. *Thematic Brief on AI Agents, Misalignment and the Risk of Losing Human Control: Evidence from the OpenAI-Hugging Face Incident* ↗, advance unedited version 1, 21 September 2026. Published at un.org and read there on 26 September 2026. Quoted at section 7.2 for the sentence on the kind of decision problem the precautionary principle was designed to address. It is the Panel's first thematic brief.
11. Petrie, James, Onni Aarne, Nora Ammann and David Dalrymple. *Flexible Hardware-Enabled Guarantees for AI Compute* ↗ (flexHEG Part I). April 2025. arXiv:2506.15093. Quoted at section 23.2. Section 15.1's concession and sections 13.3 and 13.5 take mechanisms near the series; section 23.2 states it exactly.
12. Petrie, James, and Onni Aarne. *Technical Options for Flexible Hardware-Enabled Guarantees* ↗ (flexHEG Part II). arXiv:2506.03409, version 3. Quoted at section 23.2 from the full report.
13. Aarne, Onni, and James Petrie. *International Security Applications of Flexible Hardware-Enabled Guarantees* ↗ (flexHEG Part III). April 2025. arXiv:2506.15100. Quoted at section 23.2 from the full report.
14. Aarne, Onni, Tim Fist and Caleb Withers. *Secure, Governable Chips: Using On-Chip Mechanisms to Manage National Security Risks from AI and Advanced Computing.* ↗ Center for a New American Security, January 2024. Quoted at section 23.2. The report's cover carries

“January 2024”; a publication day is not asserted. Section 15.1’s concession; section 23.2 states it exactly.

15. Kulp, Gabriel, Daniel Gonzales, Everett Smith, Lennart Heim, Prateek Puri, Michael J. D. Vermeer and Zev Winkelman. *Hardware-Enabled Governance Mechanisms: Developing Technical Solutions to Exempt Items Otherwise Classified Under Export Control Classification Numbers 3A090 and 4A090*. ↗ RAND Corporation working paper WR-A3056-1, 18 January 2024. DOI 10.7249/WRA3056-1. Quoted at section 23.2. Section 15.2’s pod-limited exclusion takes the outcome of its fixed-set concept.
16. Sastry, Girish, Lennart Heim, Haydn Belfield, Markus Anderljung, Miles Brundage, Julian Hazell, Cullen O’Keefe, Gillian K. Hadfield, Richard Ngo, Konstantin Pilz, George Gor, Emma Bluemke, Sarah Shoker, Janet Egan, Robert F. Trager, Shahar Avin, Adrian Weller, Yoshua Bengio and Diane Coyle. *Computing Power and the Governance of Artificial Intelligence*. ↗ 13 February 2024. arXiv:2402.08797, version 1. Quoted at section 23.2. Section 13.5 refuses the multiparty start-switch design and describes it; section 23.2 states it exactly.
17. Shavit, Yonadav. *What Does It Take to Catch a Chinchilla? Verifying Rules on Large-Scale Neural Network Training via Compute Monitoring*. ↗ 30 May 2023. arXiv:2303.11341, version 2. Described at section 23.2.
18. Baker, Mauricio. *Nuclear Arms Control Verification and Lessons for AI Treaties*. ↗ 8 April 2023. arXiv:2304.04123, version 1. Quoted at section 23.2.
19. Wasil, Akash R., Tom Reed, Jack William Miller and Peter Barnett. *Verification methods for international AI agreements*. ↗ 4 November 2024. arXiv:2408.16074, version 2. Described at section 23.2. The arXiv record gives the first author’s middle initial; entry 1 of the last list follows records that give none.
20. Heim, Lennart. *Considerations and Limitations for AI Hardware-Enabled Mechanisms*. ↗ 10 March 2024. Author’s blog; self-published and carrying no DOI, cited by its date and its author’s own site. Quoted at section 23.2.
21. Scher, Aaron, David Abecassis, Peter Barnett and Brian Abeyta (Machine Intelligence Research Institute, Technical Governance Team). *An International Agreement to Prevent the Premature Creation of Artificial Superintelligence*. ↗ 8 May 2026. arXiv:2511.10783, version 3. Quoted at section 23.3.
22. Hendrycks, Dan, Eric Schmidt and Alexandr Wang. *Superintelligence Strategy: Expert Version*. ↗ 14 April 2025. arXiv:2503.05628, version 2. Quoted at section 23.3.
23. United Nations Secretary-General’s High-level Advisory Body on Artificial Intelligence. *Governing AI for Humanity: Final Report*. ↗ September 2024. Quoted at section 23.3.

24. Haramboure, Antton, Guy Lalanne, Cyrille Schwellnus and Joaquim Guilhoto. *Vulnerabilities in the semiconductor supply chain*. / OECD Science, Technology and Industry Working Papers 2023/05. DOI 10.1787/6bed616f-en. Quoted at section 5.1, and cited at section 22.1 for the two makers it records as producing the most advanced chips at scale. The paper states of itself that OECD working papers “should not be reported as representing the official views of the OECD or of its member countries”.
25. ASML Holding N.V. *Annual Report 2025 based on US GAAP*. / Signed at Veldhoven on 25 February 2026. Quoted at sections 3.3 and 5.1 for the company’s statements that it is currently the world’s only manufacturer of EUV lithography systems and that Carl Zeiss SMT is its sole supplier of the critical optics.
26. European Commission. *Quantum Europe Strategy* /, COM(2025) 363, 2 July 2025. Quoted at section 15.3.
27. GlobalPlatform. *Introduction to Secure Elements*. / May 2018, page 2. Quoted at section 15.2.
28. National Institute of Standards and Technology. *Hardware-Enabled Security: Enabling a Layered Approach to Platform Security for Cloud and Edge Computing Use Cases* /, NISTIR 8320, final May 2022 text, section 3.1, page 5; section 1, page 1; and Appendix D (Arm), page 57. Quoted at section 15.2.
29. Huawei. *Huawei Launches the World’s First NPO-based SuperPoD, the Atlas 960E SuperPoD*. / Company release, 17 September 2026, published at huawei.com and read there on 26 September 2026. Quoted at section 5.2 for the Atlas 960E SuperPoD and for the SuperCluster’s stated interconnect scale. A company’s own announcement of its own product, and given in the prose as that.
30. Belfield, Haydn. *Domestic frontier AI regulation, an IAEA for AI, an NPT for AI, and a US-led Allied Public-Private Partnership for AI: Four institutions for governing and developing frontier AI*. / 8 July 2025. arXiv:2507.06379. Quoted at section 23.2 from its abstract.
31. Baker, Mauricio, Gabriel Kulp, Oliver Marks, Miles Brundage and Lennart Heim. *Verifying International Agreements on AI: Six Layers of Verification for Rules on Large-Scale AI Development and Deployment*. / RAND Corporation working paper WR-A4077-1, 24 July 2025; the preprint is arXiv:2507.15916 / . Quoted at section 23.2 from its abstract.
32. Scholefield, Rebecca, Samuel Martin and Otto Barten. *International Agreements on AI Safety: Review and Recommendations for a Conditional AI Safety Treaty*. / 18 March 2025. arXiv:2503.18956. Quoted at section 23.3 from its abstract.
33. Future of Life Institute and Centre for International Governance Innovation, leading a collaboration of governance experts. *Draft Articles on the Safe Development and Diffusion of Artificial Intelligence*. / A PDF dated 18 June 2026 by its file name and metadata, published at

global-governance.ai, the collaboration's own site, to which its front page points, and read there on 30 September 2026. Quoted at section 23.3 from its front matter, Articles 4 and 5 and Annex C. A modular proposal by its own description, not an adopted instrument.

34. Engels, Joshua, David D. Baek, Subhash Kantamneni and Max Tegmark. *Scaling Laws For Scalable Oversight*. ↗ 25 April 2025. arXiv:2504.18530. Named at section 23.1 inside the registration ↗'s own words, which are quoted there.
35. Bai, Yuntao, and colleagues. *Constitutional AI: Harmlessness from AI Feedback*. ↗ Anthropic, December 2022. arXiv:2212.08073. Described at section 23.4.
36. The White House. *Winning the Race: America's AI Action Plan*. ↗ July 2025. Quoted at sections 5.1 and 13.5 from printed page 21, and at section 22.3 from printed page 22. Described as published by the White House in July 2025; this paper makes no claim about its standing as policy on any later date.
37. Obadia, Alex, and Nora Ammann. *Trust Everything, Everywhere*. ↗ Advanced Research and Invention Agency, opportunity space, version 2, 3 August 2026. Quoted at section 15.2 from page 3 of the full text. The note says nothing about this proposal.
38. Synopsys. *Synopsys and TSMC Partner to Accelerate AI Systems Innovation with Agentic AI and Advanced Design*. ↗ Company release, 23 September 2026, published at news.synopsys.com and read there on 30 September 2026. Cited at section 14.3. A company's own announcement, given in the prose as that.
39. World Wide Web Consortium. *W3C Patent Policy*. ↗ 15 May 2025. Quoted at section 12.4 from sections 3.1 and 5.

Named in the paper without being read at source

1. Miotti, Andrea, and Akash Wasil. *An International Treaty to Implement a Global Compute Cap for Advanced Artificial Intelligence*. ↗ SSRN, 2023; also arXiv:2311.10748, whose first version of 1 November 2023 lists both authors. Neither record gives a middle initial, so none is printed here; entry 19 of the research literature follows its own record, which gives one.
2. Cha, Seokki. *Towards an international regulatory framework for AI safety: lessons from the IAEA's nuclear safety regulations*. ↗ Humanities and Social Sciences Communications 11, article 506, 12 April 2024.
3. Fist, Tim, Tao Burga and Vivek Chilukuri. *Technology to Secure the AI Chip Supply Chain: A Working Paper*. ↗ Center for a New American Security, 11 December 2024.
4. Kop, Mauritz. *Towards an Atomic Agency for Quantum-AI*. ↗ arXiv:2505.11515, 6 May 2025.

5. Sandhu, Gurjit, as editor, with Paul Imre as researcher. *A Combination Therapy Stack for Governing Frontier-Scale AI*. ✓ Version 1.0, 11 July 2025. DOI 10.5281/zenodo.15864321.
 6. Good, I. J. *Speculations Concerning the First Ultraintelligent Machine*. ✓ Advances in Computers, volume 6 (Academic Press; dated 1966 in the publisher's record and also cited as 1965), pages 31 to 88. DOI 10.1016/S0065-2458(08)60418-0. Named at section 23.5 as the classic statement of recursive self-improvement.
 7. Christoph, Joel. *Stable AGI Governance through the International AI Governance Organization (IAIGO): Proposal for the Establishment of an International Framework for Artificial General Intelligence Governance*. ✓ Prepared for the Future of Life Institute, Fall 2024, and published at the Institute's site.
-

Michael Darius Eastwood conceived and directs this research programme and is the author of this work. Across the programme, he has used more than six AI systems in parallel, under his own instructions, to stress-test his arguments, identify possible errors, and assist in preparing draft text from his own outlines. He determines what is adopted, revised or rejected and takes responsibility for the published content. These systems are tools, not authors.

First published 29 September 2026 · Page last edited 3 October 2026 · (v2.2)

Michael Darius Eastwood conceived and directs this research programme and is the author of this work. Across the programme, he has used more than six AI systems in parallel, under his own instructions, to stress-test his arguments, identify possible errors, and assist in preparing draft text from his own outlines. He determines what is adopted, revised or rejected and takes responsibility for the published content. These systems are tools, not authors.