Personal Red Lines for Development

An Evidence-Gated Personal Security Boundary and Explicit No Document for Dual-Use Development

Daniel Ari Friedman

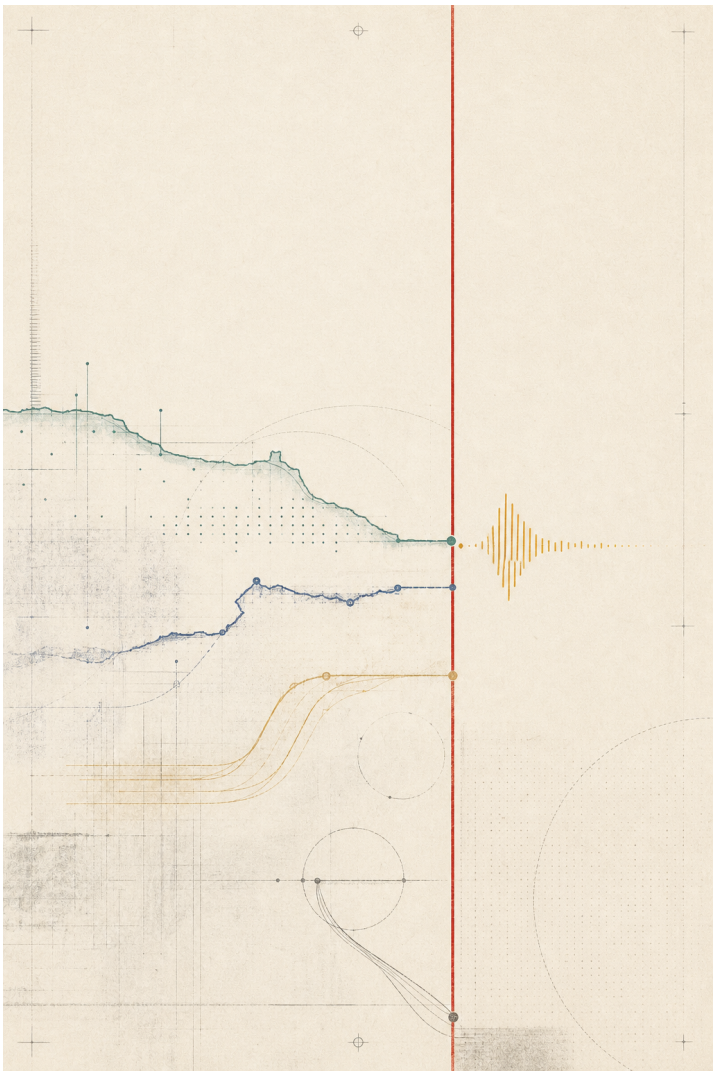
Active Inference Institute

daniel@activeinference.institute

ORCID: 0000-0001-6232-9096

DOI: 10.5281/zenodo.21754240

2026-07-17



Contents

1 Abstract

Red Line is a personal security boundary and explicit **No document** for dual-use development. Its function is deliberately narrow: help one practitioner refuse work before commitment, make a near-boundary review inspectable, and make material weakening visible over time. It records seven first-person refusals as a versioned, machine-readable registry and subjects a proposed engagement to a strict intake gate. Purpose, end use, affected parties, provenance, legal basis, human control, deployment, downstream transfer, and capability scope must each have a reviewable evidence record. Missing, self-asserted, unverified, stale, or contradicted context cannot produce `COMPLIANT`.

The instrument classifies every intake into one of five categories (defined formally in Definition 11): insufficient information blocks inspection, outside scope finds no line implicated, and the remaining three — compliant, requires modification, non-compliant — carry an ordered strictness. Typed exemptions replace narrative carve-out clauses; explicit aliases replace heuristic stemming in scope normalization, where a light stemmer survives only in the advisory description/scope mismatch hint and never reaches a classification; and a named authorization records escalation without releasing a blocking finding. A hash-based canary detects drift only against a prior copy outside the writer's control. The decisive limitations follow: local evidence is not independent truth verification, lexical matching is not semantic understanding, and code is not enforcement.

Those mechanics are also stated formally in sec. 14, which restates the domain objects, the staged evaluator, and its decision rule as numbered definitions and propositions written from the shipped code, each bound to a named test that re-derives it, so a sentence that has drifted from the procedure it describes fails rather than persuades. Two propositions are exercised at scale: degrading any one of the nine intake dimensions withdraws a compliant result in all forty-five executed evaluations, and an ALL-mode exemption stays unreachable by a single declared token across fifty-eight more.

The scholarship is situated rather than universalizing. It joins pre-1900 primary texts and historical scholarship with contemporary work from multiple regions and disciplines, recording what question each source carries, what claim it does not establish, and where transfer into this personal instrument stops. Two literatures bound the project from either side: dual-use export control, which is what this question looks like when institutions answer it, and the refusal literature, whose standing a practitioner declining paid work cannot borrow. Turner's framework is the mechanism source (sec. 3), not the title, subtitle, or authority of the project. The figures are deterministic and generated from the registry, evaluator, canary, and source ledger; captions separate implementation facts from interpretation and visual rhetoric, and the claim register and decision protocol carry the evidence chain and the operator sequence.

Red Line remains narrow. Black Line addresses positive operating discipline; Golden Line addresses aspiration; White Line addresses absence, restraint, and unknowability. They cross-reference one another without importing their substantive methods into this refusal boundary.

2 Introduction

Some work becomes harder to refuse once it has acquired a client, a deadline, or a sunk-cost narrative. Red Line places the refusal earlier. It is the author's personal security boundary and explicit **No document**: a dated, revisable statement of work he will not accept, build, tune, release, or knowingly transfer.

This is not a complete ethics system, a legal opinion, or an enforcement service. Black Line concerns constructive practice; Golden Line concerns direction; White Line concerns absence, omission, and epistemic restraint. Red Line is narrower because a boundary must be usable at the moment of decision. It answers three practical questions: what is refused, what context must be established before a near-boundary action can be considered, and what the instrument cannot establish.

2.1 Four propositions

1. **Red Line is a security boundary.** Its seven first-person lines identify work that this author will not pursue or knowingly enable.
2. **Red Line is an explicit No document.** It is a personal commitment, not a universal moral authority and not a statement made by an AI system on the author's behalf.
3. **Red Line is an evidence-gated auditability aid.** It makes a local review inspectable and reproducible; it does not prove safety, legality, or truth.
4. **Red Line escalates when context cannot be established.** Missing or unverified evidence is `INSUFFICIENT_INFORMATION`, not permission.

2.2 What this paper can establish

Evidence layer	It can establish here	It cannot establish
Personal commitment	what this author currently refuses	universal moral authority or another person's consent
Local implementation	what the code and tests return for declared inputs	semantic truth, honest input, or operational safety
Scholarship transfer	which questions a source prompts	source endorsement, consensus, or public authorization
Render and release	that a specific validation run connected source to artifact	external certification or publication readiness without clean and independent witness gates

2.3 What the artifact is for

At the decision moment the instrument is a triage card rather than an essay: read the beacon, name the concrete capability and deployment tier, resolve the nine context dimensions with evidence instead of assertion, run the local evaluator, preserve the finding with its stable reason codes, and either narrow the work or stop. Maintaining a prior canary outside the author's write boundary is the one step that runs on a different clock. sec. 7 states the sequence in full; the operator version is in `docs/decision-protocol.md`, and day to day the instrument is run through the `daf-red-line` skill in the author's private daf-skills toolchain, so this paper is the theory and the skill is the practice. The point is relevance to a real decision under time pressure. A polished boundary that does not change what the practitioner records, asks, or refuses is only decoration. fig. 1

The design puts refusal before optimization and evidence before policy matching. It also separates four kinds of statement: a source may support a descriptive claim; an interpretation may transfer a question with an explicit stopping point; an implementation claim must be supported by code, tests, generated artifacts, validation, or inspected output; and a release claim must be tied to a validated source-to-render chain for a particular artifact. None of these surfaces can silently upgrade another.

2.4 Beacon, evaluator, canary

The registry is a **beacon**: collaborators can read the boundary before a request becomes a project. The evaluator is a narrow policy instrument, not a semantic judge. It requires a typed `ActionContext`, verified evidence for each required dimension, normalized scope, and a structured exemption before it can return `COMPLIANT`. A complete action that touches no registered line is `OUTSIDE_SCOPE`, which is deliberately not collapsed into compliance.

The canary is a dated attestation over canonical registry content. It can expose drift when a prior copy is held beyond the author's write boundary; it cannot prevent a rewrite, prove authorship, enforce a finding, or independently establish that a source record is true. A review authorization records escalation but never unblocks a non-compliant or evidence-insufficient result.

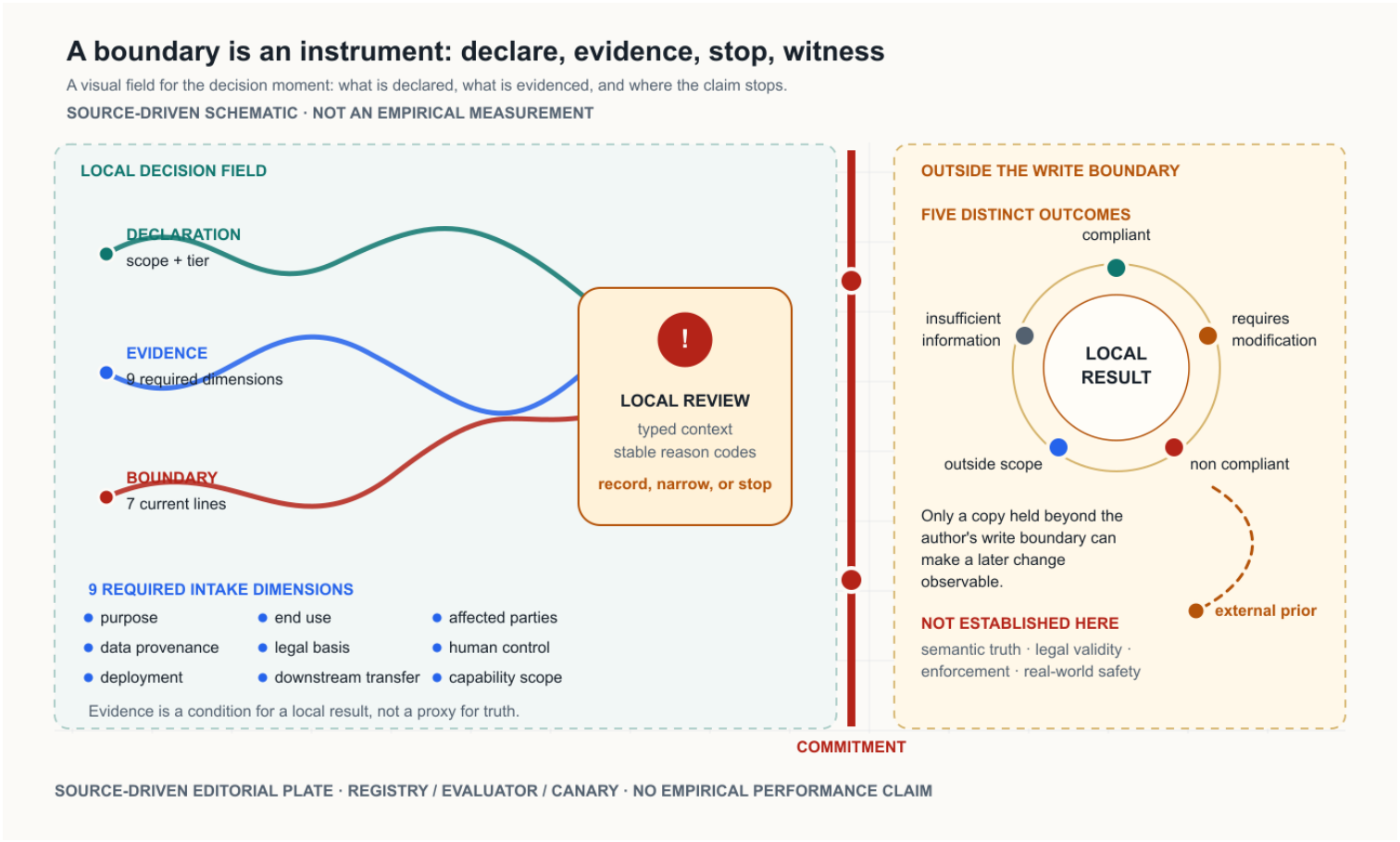


Figure 1: A boundary is an instrument: declare, evidence, stop, witness. This editorial plate turns the live registry, nine-field intake, five evaluator classifications, and external-prior condition into a single decision-time field. Its paper texture, traces, and red mark are visual rhetoric, not empirical data; the caption and alt text preserve the non-claims if the image is unavailable.

2.5 A bounded comparison

Alex Turner's *A Red Line and Oversight Framework for Government AI Contracts* [Turner, 2026] is the mechanism source for this project's architecture (sec. 3). Its organization-to-government setting, standing Review Body, legal context, and specific scope are not imported as authority. The adaptation keeps only a carefully marked pattern — precommitment, retained oversight, review classifications, and durability — and changes the subject to one practitioner's own work.

The manuscript therefore keeps two questions separate: what the sources make visible, and what this author commits to. The scholarship widens the questions asked of the registry—who bears the cost of a false positive, what legibility can conceal, whose knowledge is being organized—but it does not convert a plural reading list into universal legitimacy. The claim register makes this stopping rule auditable.

The paper proceeds as follows. sec. 3 records the mechanism source. Section sec. 14 states the instrument as formal definitions and propositions. The red-line registry is composed in Section sec. 15, and Section sec. 17 closes with the instrument's limitations and negative space.

3 Background — Turner as a bounded mechanism source

Alex Turner's *A Red Line and Oversight Framework for Government AI Contracts* is the principal comparative source for this project's architecture [Turner, 2026]. Turner writes at organization-to-government scale. The framework combines bright-line standards, retained-oversight tiers, a standing Review Body, review classifications, durability provisions, and transparency. This section records the source before describing the personal adaptation.

Turner's formulation of appropriate human control is more demanding than a checkbox saying "human in the loop." It asks whether an identifiable person is accountable, whether that person can exercise independent judgment, whether the system's operational design permits meaningful evaluation, and whether legal and operational transparency are available [Turner, 2026]. It also recognizes that speed, volume, and complexity can make nominal human review ineffective. Red Line carries these as questions for the intake; it does not pretend that a text field or token can establish them.

Turner's surveillance standard similarly distinguishes individualized, particularized analysis from inference based only on a category, population, or bulk dataset. The existence of data about a person is not, by itself, permission to generate an individualized assessment. Those details matter here because they show why Red Line's `human_control`, `affected_parties`, `purpose`, and `data_provenance` fields must be evidenced rather than asserted.

Turner also describes scope-specific carve-outs, oversight thresholds, and durability procedures. Ambiguity resolves toward coverage in the source framework [Turner, 2026]. Red Line preserves that conservative direction but implements a stronger local first gate: unresolved context becomes `INSUFFICIENT_INFORMATION` before the registry can return a policy result.

3.1 What is and is not transferred

Transferred as a question: can a bright-line commitment be made inspectable, reviewed before action, and made harder to weaken silently? Not transferred as authority: Turner's legal setting, government contracting relationship, seven-member institution, leadership powers, or claims about external enforcement. Red Line is a personal, self-authored instrument with an explicit external-witness limitation.

The implementation mapping appears in the next section. The important boundary is that an analogy to a governance mechanism is not evidence that the personal instrument has the institution's powers.

The source framework also contains organizational, contractual, legal, and operational machinery that this project does not possess: a company, government counterparties, a standing Review Body, neutral-auditor access, safety-stack controls, and service suspension. Those omissions are not hidden implementation debt. They are the reason the local artifact claims auditability rather than institutional enforcement.

3.2 Provenance labels for the adaptation

The manuscript uses three labels so that resemblance does not become false attribution:

Element	Label	Exact boundary
Human control for harm-capable action	**source-derived** question; **adapted** intake	Turner supplies the accountable-human problem; Red Line adds fields and evidence, not Turner's authority.
Individualized and proportionate scrutiny	**source-derived** distinction; **adapted** refusal	Turner supplies particularization; Red Line applies it to S2, not to a complete proportionality theory.
Bright lines, tiers, review, and durability	**source-derived** mechanism; **adapted** architecture	The local registry, evidence gate, five classifications, self-review, and canary are this project's implementation.
Personal No document and strict evidence gate	**independently authored**	Personal design decisions, not Turner quotations, legal standards, or universal conclusions.

4 Global and Historical Scholarship: Situated Questions, Not a Universal Lineage

The reading base is an audit surface, not an authority bundle. The selection protocol prefers primary texts, scholarly editions, publisher pages, and official institutional sources; includes pre-1900 materials when their question remains useful; and records the boundary on every transfer. For every one of its 45 sources, the ledger in `docs/research-method.md` records place, period, the question carried into the audit, the non-transfer boundary, and a verified publisher or primary-text URL. Those 45 sources sit in 44 table rows: Kukutai and Taylor is listed in both tables, and the Cugoano row cites the primary text alongside its *Stanford Encyclopedia* entry.

Genre, language or translation, and a stable locator are recorded only for the 22 sources in the deepened second table; the machine-readable ledger at `data/source_claims.json` marks the remaining 23 `not recorded` rather than filling the field with boilerplate, and `validate_source_claims` rejects any locator derivable from a source's own citation key.

4.1 Before 1900: refusal, rule, knowledge, and power

Sunzi's *Art of War* makes information, deception, timing, and judgment operational questions [Sunzi, 1910]. Kauṭilya's *Arthaśāstra*, in Patrick Olivelle's scholarly translation, joins law, administration, intelligence, and the practical limits of rule [Kautilya, 2013]. Aristotle's *Politics* asks how constitutions distribute authority, while al-Fārābī's political philosophy links knowledge, political order, and human flourishing [Aristotle, -350, Mahdi, 2000]. Ibn Khaldūn's account of 'asabiyya and justice adds a historically situated analysis of cohesion and power [Darling, 2007].

These texts are not a single tradition. Machiavelli supplies an uncomfortable test about incentives and power [Machiavelli, 1513]; Wollstonecraft asks whether a discourse of reason can survive the denial of equal intellectual standing [Wollstonecraft, 1792]. Ottobah Cugoano's late-eighteenth-century abolitionist work adds a Black Atlantic voice on liberty, consent, responsibility, and the self-deception that allows beneficiaries of coercion to call it legitimate [Cugoano, 1787, Dahl, 2025]. Cugoano is not used as a generic representative of "African ethics"; he is read in his own Atlantic, religious, abolitionist, and philosophical context.

The transfer from these sources is a set of questions: who acts, under what authority, with what knowledge, toward whom, with what power to withdraw or repair, and who bears the cost when a claim is wrong? None supplies a modern rights framework, a universal AI policy, or an endorsement of this registry.

The Latin American record adds a necessary asymmetry. Bartolomé de las Casas's account is a colonial-era Spanish cleric's polemic about violence in the Americas, not Indigenous testimony and not a representative voice for the region [Las Casas, 1552]. Its narrow use here is diagnostic: an institution can describe its own authority as administration, improvement, or salvation while affected people bear coercive costs. The source therefore sharpens the intake questions about affected parties and power; it does not supply a universal theory of consent or authorize the author's categories.

4.2 Modern critical and institutional scholarship

Elinor Ostrom shifts attention from abstract market/state choices toward rules-in-use, monitoring, graduated responses, and situated knowledge [Ostrom, 1990]. James C. Scott shows how administrative legibility can simplify a system while erasing local knowledge [Scott, 1998]. Linda Tuhiwai Smith makes research power and extraction part of method [Smith, 1999]. Helen Nissenbaum's contextual integrity rejects a context-free account of information flow [Nissenbaum, 2004]. Together they require Red Line to treat provenance, affected parties, downstream use, and public legibility as context-dependent rather than as checkboxes that automatically confer legitimacy.

Jobin, Ienca, and Vayena show convergence and divergence among AI principles; Selbst and colleagues warn that abstraction can sever technical categories from their sociotechnical context [Jobin et al., 2019, Selbst et al., 2019]. Birhane, and Mohamed, Png, and Isaac, put coloniality, dependence, local knowledge, and critical technical practice into the AI governance conversation [Birhane, 2020, Mohamed et al., 2020]. These sources constrain the project's temptation to call a readable registry universal or emancipatory.

The surveillance and classification literature makes the same warning more concrete. Browne traces surveillance of Blackness through historical and contemporary practices [Browne, 2015]; Noble shows how apparently neutral search infrastructures can reproduce racialized hierarchy [Noble, 2018]; and Eubanks documents how automated administrative systems can concentrate burdens on people with little power to contest them [Eubanks, 2018]. Fricker supplies a distinct epistemic question—who is treated as a credible knower and who lacks interpretive resources [Fricker, 2007]—while Couldry and Mejias frame data extraction as a relation of power rather than a mere technical flow [Couldry and Mejias, 2019]. Gray and Suri add hidden human labor to the system boundary [Gray and Suri, 2019]. Together these works justify keeping affected parties, provenance, downstream transfer, and human control as separate intake dimensions. They do not prove that any particular action is unlawful or unsafe.

Four further works make the bridge from scholarship to operating intelligence explicit. Haraway's account of situated knowledges requires the author to name the standpoint and partial perspective behind a claim rather than perform a view from nowhere [Haraway, 1988]. Jasanoff's "technologies of humility" organize uncertainty around framing, vulnerability, distribution, and learning [Jasanoff, 2003]. Costanza-Chock's design-justice practice asks who is centered, burdened, excluded, or able to contest a design [Costanza-Chock, 2020]. D'Ignazio and Klein add a power-aware account of data, classification, and invisible labor; data do not speak for themselves [D'Ignazio and Klein, 2020].

The transfer is operational, not ornamental. These questions attach to the intake as follows: standpoint and missing perspective sharpen provenance and affected parties; framing and vulnerability sharpen purpose, end use, and unknowns; participation and contestability sharpen human control and downstream transfer; and power, classification, and labor sharpen capability scope. None of the four sources can turn a self-assertion into verified evidence or make a local result a public authorization. Their value is that they make the operator ask a better question before the evaluator returns a bounded answer.

Kukutai and Taylor add a collective-authority test that individual consent does not settle: when data concern an Indigenous people or community, who has standing to govern collection, interpretation, and downstream use? [Kukutai and Taylor, 2016] This is a question for `affected_parties`, `data_provenance`, `legal_basis`, and `downstream_transfer`, not a new universal permission rule. The source is especially relevant because it prevents the intake from treating data sovereignty as a property of an individual record alone; it does not replace the authority of the affected community or establish that a particular data use is legitimate.

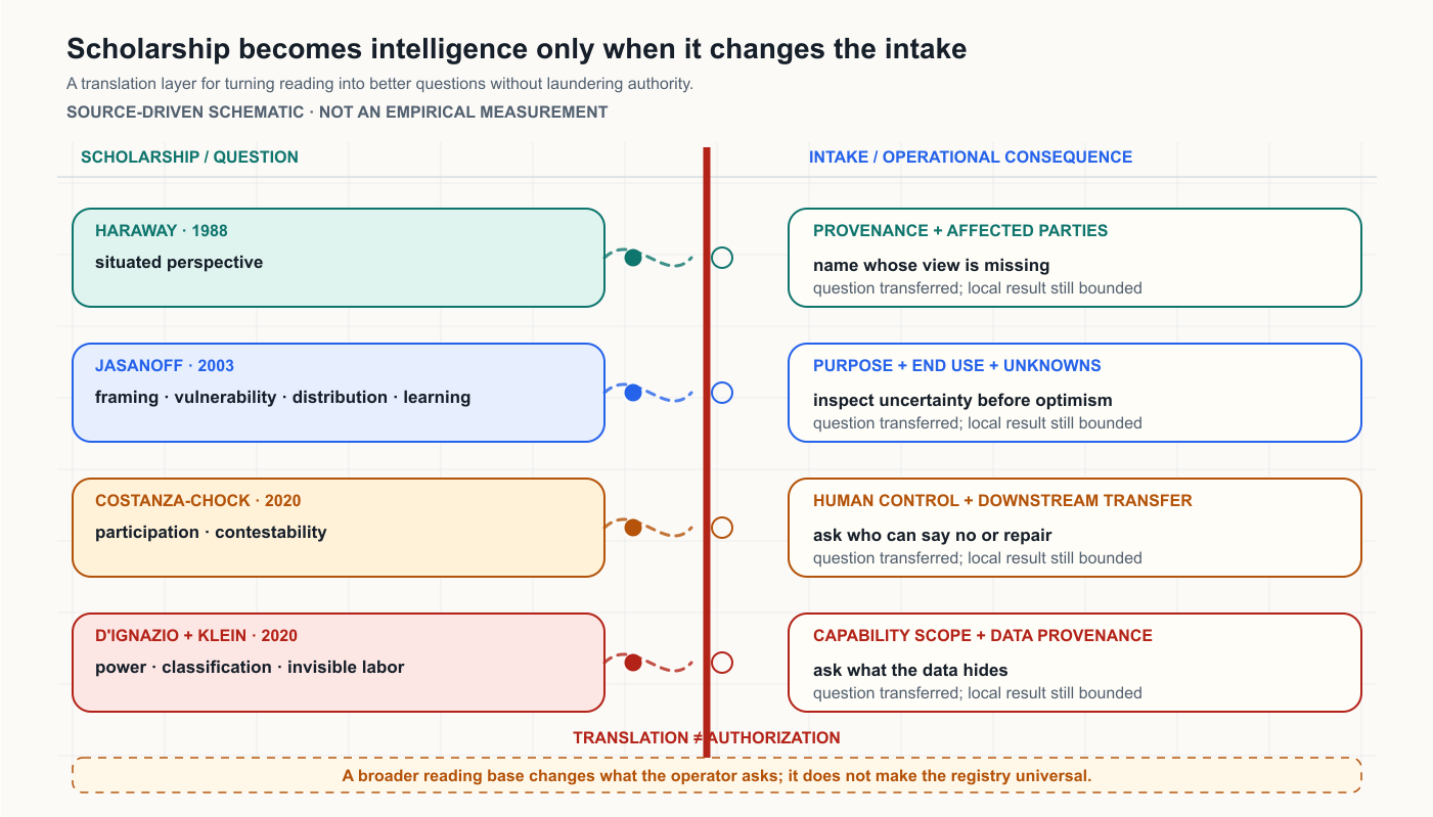


Figure 2: Scholarship becomes intelligence only when it changes the intake. This bridge maps four added works to concrete questions about perspective, uncertainty, contestability, classification, and hidden labor. The red gap is deliberate: sources widen the audit surface but do not authorize an action, verify an intake field, or replace affected-party testimony.

4.3 Export control and the refusal of work

Two literatures sit closer to this artifact than any cited so far, and neither of them flatters it. The first governs dual-use capability from above; the second describes refusing work from below. Red Line is neither, and saying so precisely is more useful than claiming kinship with both.

Export control is the mature institutional form of the question this registry asks. The Wassenaar Arrangement maintains a list of dual-use goods and technologies that participating states implement in national law [The Wassenaar Arrangement on Export Controls for Conventional Arms and Dual-Use Goods and Technologies, 2025]. Its structure is instructive twice over. It keys control to declared *capability categories* rather than to an assessment of the person applying for a licence, which is the same lexical bet Red Line makes and the same one sec. 17 concedes is escapable by description. And its 2013 addition of "intrusion software"

is the reference case for what goes wrong when a capability definition is drawn slightly too wide: a category meant to constrain commercial spyware also described the exchange of defensive research, and the definition had to be revisited. A personal registry is smaller than a multilateral regime in every respect that matters, but it inherits that failure mode exactly, which is why the scope vocabulary in sec. 15 is enumerable and printed in full rather than described.

Harris's comparative study of nuclear, biological, and cyber governance sets out the layering these regimes depend on: treaty obligations, export controls, institutional review, and the judgment of the researcher, each carrying part of the load and none of them sufficient alone [Harris, 2016]. The Fink report made the last of those layers explicit for the life sciences, naming seven categories of experiment that should trigger review and arguing that the scientific community's own screening is a governance layer rather than an informal habit [National Research Council, 2004]. That is the strongest available warrant for an instrument like this one, and it is also a bounded one: the report proposed researcher judgment *alongside* institutional review boards, never instead of them, and Red Line has no board. What transfers is the proposition that a practitioner's pre-commitment is a recognized layer. What does not transfer is any suggestion that the layer is adequate by itself.

The refusal literature supplies the vocabulary the registry was missing for its own act. Zong and Matias give refusal a structure — autonomy, timing, power, and cost — written deliberately from the standpoint of people refusing an institution's data collection rather than the institution seeking compliance [Zong and Matias, 2024]. Their four facets are the sharpest available test of this instrument: it is individual rather than collective on autonomy, proactive rather than reactive on timing, weak on power because it binds only its author, and it redistributes cost onto that author alone. The transfer stops firmly at standing. Refusal from below is refusal by those with the least power in a relation; a practitioner declining paid work is not that, and borrowing the term's moral weight would be exactly the authority-washing this section exists to prevent. The same boundary applies with more force to Simpson, whose account of refusal as a positive political stance — a sovereignty claim with its own content, not the absence of consent — belongs to Kahnawà:ke and to settler-colonial history [Simpson, 2014]. The concept that a "no" can be generative rather than merely negative is what this project takes; the standing is not available to be taken.

Between the two literatures sits the one documented case of collective refusal in this industry. Crofts and van Rijswijk trace Google's Project Maven episode: thousands of workers refused military AI work, the contract lapsed, principles were published — and the corporate form absorbed the objection intact [Crofts and van Rijswijk, 2020]. The episode is evidence that refusal is possible and legible, not a base rate. For a single practitioner the lesson is narrower still: refusal held collectively had leverage refusal held alone does not, which is the honest reason this artifact claims auditability rather than effect.

4.4 Selection limits and positionality

This is a curated conceptual audit, not a systematic review, a representative survey of world traditions, or a substitute for consultation with people affected by a proposed system. The selection is constrained by sources available in accessible editions and translations, the author's questions, and the need to keep each transfer legible. Geographic variety is therefore a corrective to a narrow lineage, not a coverage statistic. A source from a region is not a proxy for everyone in that region; a translated text is not identical to its language of composition; and placing sources in one figure does not make them equivalent.

The consequence is methodological restraint: the reading base can expand the intake questions—about power, classification, labor, consent, and contestability—but it cannot supply affected-party testimony, jurisdiction-specific legal review, or independent verification of an evidence record. Those remain separate obligations.

For dual-use security, Brundage and colleagues map malicious-use concerns across digital, physical, and political domains and argue for prevention and mitigation rather than a single prediction [Brundage et al., 2018]. That supports a capability-and-transfer question in Red Line, not a claim that a lexical registry forecasts threat likelihood. Three engineering references sit beside it and are taken up again in sec. 12, where they describe what this project does and does not do: the NIST Secure Software Development Framework, MITRE ATT&CK, and SLSA v1.2 [Souppaya et al., 2022, MITRE, 2026, SLSA Community, 2026]. Here they enter the reading base for one reason only — they name the questions to ask about provenance, adversary behavior, and build attestation. Neither a standard nor a threat catalog is evidence that Red Line is secure, legally compliant, or resistant to an advanced persistent threat.

UNESCO, the OECD, and the African Union provide policy comparison points; NIST provides operational vocabularies for risk management and continuous verification [UNESCO, 2021, OECD, 2019, African Union Commission, 2022, Tabassi, 2023, Rose et al., 2020]. They are not certifications of Red Line and cannot substitute for the project's own evidence gate.

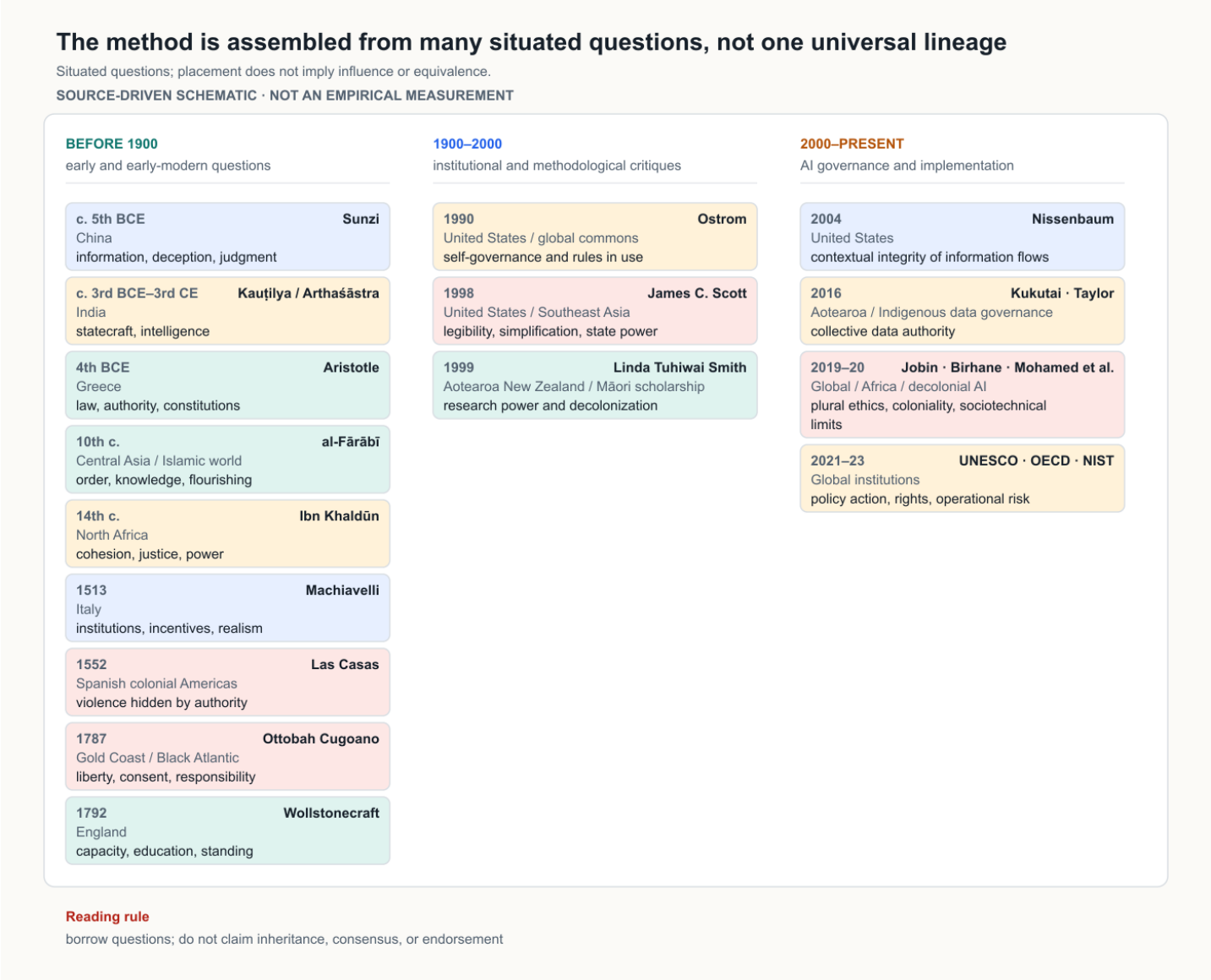


Figure 3: The method is assembled from many situated questions, not one universal lineage. The cards group pre-1900 primary texts, historical scholarship, critical methodology, collective data-governance work, and contemporary standards by broad period and region. Direct labels state the question carried into Red Line; placement does not imply direct influence, equivalence, or endorsement. The figure deliberately includes an interpretive boundary: a source can widen the audit surface without authorizing the personal refusal registry. It is a provenance and humility device, not a measurement of global coverage or scholarly quality.

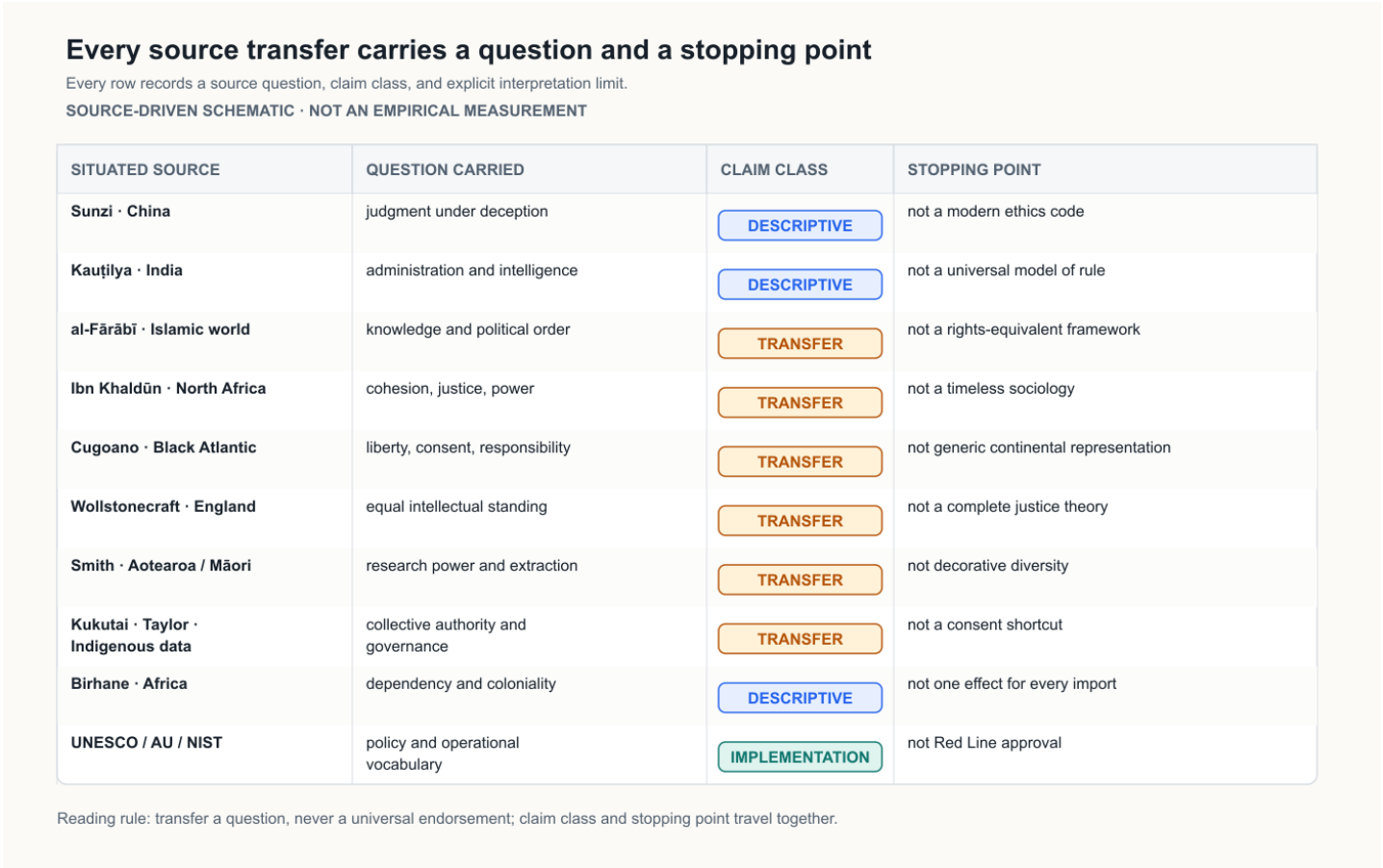


Figure 4: Every source transfer carries a question and a stopping point. This deterministic matrix makes the claim discipline visible: descriptive, transfer, and implementation claims are separated, and each source row names the boundary beyond which it is not used. Collective data authority appears alongside individual consent and contextual privacy; the visual does not score traditions, demonstrate consensus, or make the reading base universal. It is designed for the PDF and narrow HTML layout with direct labels, repeated color-and-text encodings, and a caption that remains meaningful if the image is unavailable.

5 The four-line set: boundary, method, aspiration, absence

Red Line is the first work in a four-line set, but it is not an omnibus theory. It is the security boundary and explicit **No document**: a dated record of work the author refuses, the conditions under which a proposed action is reviewed, and the limits of a self-authored canary.

The companion works answer different questions. Black Line describes the positive operating discipline for doing strong, concise, intelligent work and research. Golden Line describes a higher aspirational thread: directions that can orient a life or project without becoming a compliance verdict. White Line describes what is absent or unknowable, including epistemic gaps, ethical restraint, withheld material, and the negative space around a claim.

The separation is substantive. Red Line does not turn good method into permission; Black Line does not turn aspiration into prohibition; Golden Line does not turn uncertainty into failure; White Line does not turn silence into evidence. Each work has its own registry, vocabulary, executable interface, sources, figures, tests, and limitations, and each is its own repository — `docxology/black_line`, `docxology/golden_line`, and `docxology/white_line`. The relationship is recorded in the companion `line_set` work (https://github.com/docxology/line_set), which declares the set and checks that no two instruments gave the same spelling to different things.

The distinction is summarized in fig. 5. The diagram is an orientation aid, not a hierarchy: the colors name different jobs, not different degrees of moral worth.

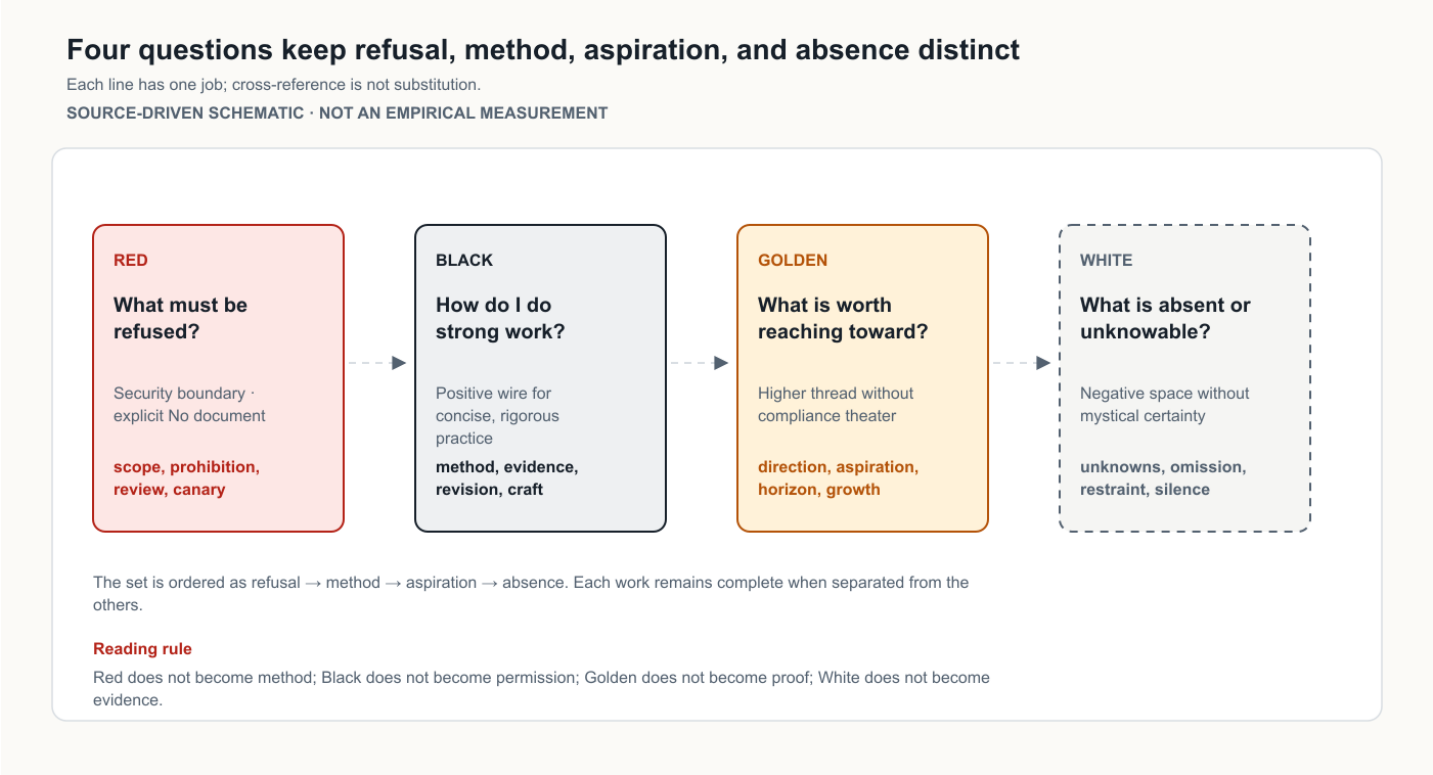


Figure 5: Four questions keep refusal, method, aspiration, and absence distinct. Orientation schematic for the four-line set: Red Line is the refusal boundary and explicit No document; Black Line is the positive wire for strong, concise, intelligent work; Golden Line is an aspirational direction rather than a pass/fail rule; and White Line records what is absent, withheld, uncertain, or left outside a claim. The colors are semantic accents, not severity scores. The four works are standalone instruments with different registries and evaluators, and the figure does not imply that one line can substitute for another.

The current project therefore remains narrow by design. Its questions are:

1. What must this practitioner refuse?
2. How can that refusal be made inspectable before work begins?
3. What can a registry, evaluator, self-review record, and external prior copy demonstrate, and what can they not demonstrate?

Black, Golden, and White now exist as sibling instruments in their own repositories, each with its own release number; the `line_set` companion records their ordering and non-overlap contract, and this paper pins none of their version numbers. Red Line remains the reference boundary rather than a container for their adjacent concerns. A cross-reference does not create a combined evaluator:

a positive method cannot authorize a refused action, an aspiration cannot resolve missing evidence, and a statement of absence cannot prove safety.

One vocabulary item is deliberately shared across the set, and a reader holding only this paper needs to know it. Black Line also publishes a status spelled `OUTSIDE_SCOPE`, and it does not mean there what it means here: in Red Line it marks a complete, evidenced intake that implicates no current registry line, while in Black Line it marks an attempt outside that discipline's evaluation scope. One is a finding of absence after inspection; the other is a declination to inspect. The overlap is declared rather than accidental, and neither sense may be read into the other.

A fifth work, `line_set`, is a thin reader that declares the set and checks that no two lines gave the same spelling to different things — including the shared token just named. It adds no substantive instrument, and Red Line does not import, depend on, or defer to it.

5.1 Note on the name

This paper's alchemical stage is **Rubedo**, the reddening — the culminating phase of the classical *magnum opus* — and the color openly echoes that opus. The set's four colors (black, white, gold, red) name the stages of the alchemical work in plain sight. That echo is read here only in Carl Jung's symbolic-psychological register, where the *opus* is a map of individuation; it is not an empirical, mystical, or causal claim about matter, minds, or this instrument. "Green is not proof" applies to a fluent symbolic parallel as much as to a passed test. The opus reads rubedo as a culmination, but the name is borrowed as symbol, not as rank: Red Line is not the completion of the set nor superior to the other three instruments — the four are peers, and Red Line is merely the one that answers what to refuse. Note too that the set's working order — refuse, method, aspire, absence — is functional, chosen for how the instruments are used, and does not reenact the classical stage sequence (in which reddening comes last, not first). The full framing, with its single Jung citation and the honest esoteric caveats, is documented in the `line_set` companion work (https://github.com/docxology/line_set) and is not restated here.

6 First-principles design: what the artifact can actually do

Before asking whether Red Line resembles another governance framework, ask what problem the artifact must solve and what information it can actually possess. One practitioner, deciding alone, before the work starts. That is the whole situation the artifact is built for, and it fixes what the artifact can know: whatever the practitioner has written down and whatever he can check locally. Everything the abstract promises follows from that constraint rather than from an ambition — and so does everything it refuses to promise, which is universal morality, verification of the world, and stopping an action by software alone.

6.1 Deconstruction

The project is made of six constituent parts:

Part	Irreducible job	Evidence in this repository
Personal commitment	State what this author will not accept, build, tune, release, or knowingly transfer	Seven dated first-person `RedLine` records
Intake	Describe an action before its description can be used as a policy shortcut	`ProposedAction` and nine-field `ActionContext`
Evidence gate	Refuse a green result when required context is missing, stale, contradicted, or merely asserted	Typed `EvidenceRecord` values and `INSUFFICIENT_INFORMATION`
Local decision procedure	Apply declared scope, typed exemptions, and oversight floors consistently	`evaluate_action` and its five classifications (Definition 11)
Review and change record	Preserve what was decided and expose registry drift	Frozen `ReviewFinding`, transparency aggregation, and `CanaryStatement`
Publication surface	Let a reader compare source, implementation, and rendered artifact	Beacon prose, source ledger, deterministic figures, PDF, and HTML

The parts are complementary but not interchangeable. A hash cannot replace a review. A review cannot replace evidence. Evidence cannot become truth merely because it is labeled `VERIFIED`. A public document cannot become universal authority merely because it cites sources from many places.

6.2 Fundamental truths

The design begins from facts that survive removal of familiar labels:

1. A personal refusal has authority over the author's own participation, not over other people or institutions.
2. A local program receives declarations and pointers to records; it does not, by inspecting their labels, know whether the underlying world is true.
3. A lexical scope matcher can apply explicit vocabulary consistently, but it cannot infer the actual capability or purpose of arbitrary work.
4. A cryptographic digest can show that current content differs from a prior digest. It does not prove authorship, authenticity, semantic adequacy, or non-forgeability.
5. A frozen finding records a result; without an external authority or admission control it cannot physically prevent execution.
6. If the conditions for a review cannot be established, treating the action as compliant would manufacture certainty from absence. This is the classical fail-safe-defaults posture: base a decision on demonstrated permission rather than on the mere absence of objection.
7. A publication claim is only as strong as the chain connecting its source, code, tests, generated assets, and rendered output.

These are not claims that the instrument is safe. They are the reasons its strongest honest output is often a stop, a limitation, or a request for an independent witness.

6.3 Constraint analysis

The project also contains choices that should not be mistaken for laws of nature. The following classification keeps the mechanism revisable:

Design element	Classification	Why it is kept or challenged
First-person, dated provenance	Hard requirement of the stated function	Without an accountable author and date, the artifact cannot be a personal commitment.
Fail closed on unresolved required context	Hard epistemic consequence	Missing information cannot logically support a positive review result.
Nine required intake dimensions	Deliberate policy choice	The fields are a broad audit surface, not a complete ontology; new evidence needs may require revision.
Explicit aliases and an ASCII boundary	Deliberate policy choice	They block trivial masquerading without pretending to solve semantic interpretation.
Seven current lines	Deliberate scope choice	`REGISTRY_IS_EXHAUSTIVE = False`; absence is not endorsement.
180-day freshness window	Maintenance policy	It is a useful review cadence, not a universal truth about evidence expiry.
Turner as the organizing comparison	Removable interpretive choice	The project must stand without importing Turner's authority, institution, or legal claims.
External prior canary copy	Hard condition for external change detection	A same-author file can be rewritten together with the registry and therefore cannot witness itself.
PDF and HTML validation	Hard requirement of a publication claim	A source that does not reach the reader intact has not been successfully published.

6.4 Reconstruction: what the ordering has to be

Those truths do not yet give a procedure, but they fix its order. Refusal comes before optimization, because a boundary consulted after a project has a client is a boundary that argues with sunk cost. Evidence comes before policy matching, because a description that can select its own exemption token is not a description. A result comes before authorization, because an escalation that can be recorded ahead of a finding is an override wearing a different word. And a local attestation comes before, without ever substituting for, an independent witness — a file the author can rewrite cannot testify about the author.

The eight-step form the practitioner actually runs is in sec. 7; it is one procedure stated once, not a summary and an expansion.

6.5 Claim classes

The manuscript uses four claim classes so that a polished document does not make unlike evidence look interchangeable:

Claim class	What can support it	What cannot upgrade it
Descriptive	A cited source, edition, and stable locator	A source's existence does not establish the author's interpretation of it
Transfer / interpretive	An explicit question, rationale, and stopping point	A broad reading list does not create universal legitimacy
Implementation	Source code, tests, generated figures, validation, and inspected output	Green code does not verify the world, the source records, or operational safety
Release	A validated source-to-render chain, artifact manifest, and inspected PDF/HTML	A successful build does not provide external certification or real-world outcome evidence

The adaptation is therefore a bounded construction rather than a conclusion that the cited traditions, standards, or threat catalogs endorse Red Line. The artifact is strongest when it says exactly what it did, what it refused to infer, and where another reviewer must take over.

The resulting operating sequence is made explicit in sec. 7, where each claim carries an evidence state and stopping point through the local decision and publication gates.

7 Operating method: from first principles to a bounded release

The first-principles reconstruction (Section sec. 6) fixes the order; this section gives the procedure.

The first-principles section identifies the artifact's irreducible job: help one practitioner decline or narrow high-risk development work before commitment, record the local reason, and disclose what the record cannot establish. This section turns that reconstruction into a repeatable method. It is a protocol for producing inspectable evidence, not a claim that the protocol can discover semantic truth or guarantee a safe outcome.

7.1 The unit of analysis

Keep three objects separate:

Object	Question	Permitted evidence
Observation	What was declared, recorded, or rendered?	a typed field, dated record, source locator, test output, or artifact hash
Interpretation	What question or limit is being transferred?	an explicit rationale and stopping point attached to a source or claim
Action decision	What may this author do within this local registry?	the evidence-gated evaluator result for a declared scope, tier, and review date

A source can support an interpretation without authorizing an action. A test can support an implementation claim without verifying the truth of an intake field. A rendered page can show that a statement reached an artifact without showing that the statement is correct in the world. This separation is the central anti-confusion rule.

7.2 The operating loop

1. **Deconstruct the function.** State the concrete decision or release problem before importing a label such as governance, security, or safety. Identify who can refuse, who may be affected, and what the instrument can actually observe.
2. **Challenge the boundary.** List hidden assumptions about authority, scope, evidence, freshness, semantic interpretation, and external witnessing. Mark each as a hard constraint, a local policy choice, or an open question.
3. **Declare the claim.** Assign a claim class (`descriptive`, `transfer`, `implementation`, or `release`), a verification mode, a supporting surface, and a stopping point. The machine-readable register in `data/claim_register.json` is the publication-facing contract.
4. **Reconstruct the smallest honest rule.** Represent an action with its scope, deployment tier, nine context dimensions, evidence records, and explicit unknowns. Do not let free-text purpose or an exemption token substitute for typed context.
5. **Evaluate fail-closed.** Normalize the scope, check missingness, unresolved statuses, freshness, and ambiguity before policy matching. Then apply typed exemptions and tier floors. The evaluator's five classifications (Definition 11) remain distinct and ordered, from the intake-blocking `INSUFFICIENT_INFORMATION` through `OUTSIDE_SCOPE`.
6. **Freeze the record.** Preserve the classification, implicated lines, human-readable reasons, stable reason codes, normalized scope, evidence-stop dimensions, review date, and any authorization. An authorization may document escalation or remediation; it cannot turn a block into permission.
7. **Verify the publication chain.** Rebuild source-driven figures, validate bindings, run tests and coverage, render PDF and HTML, and inspect the combined artifacts. A source claim is not a release claim until its required surface exists and is checked.
8. **Revise or stop.** If a result, source, figure, or rendered artifact fails its criterion, record the defect and either amend it with a dated rationale or stop. Never replace an unresolved state with a favorable default.

7.3 Evidence states and stop points

The protocol uses evidence as a condition for a local decision, not as a proxy for truth. `MISSING` means no usable value or supporting record exists; `SELF_ASSERTED` means the requester supplied an assertion without an independent reviewable basis; `UNVERIFIED` means a pointer exists but has not been checked; `CONTRADICTED` means the available record conflicts with the declaration; and `STALE` means the record is outside the configured review window or future-dated. None can narrow a line. `VERIFIED` means only that a reviewable record was checked for this decision and remains current; it can narrow a line only through its typed exemption and declared scope.

This vocabulary prevents an important category error. A verified legal-basis record is evidence that a reviewer checked the supplied record; it is not a legal opinion generated by Red Line. A verified human-control record is not proof that people will exercise that control under pressure. The evaluator therefore refuses to manufacture `COMPLIANT` from an information gap.

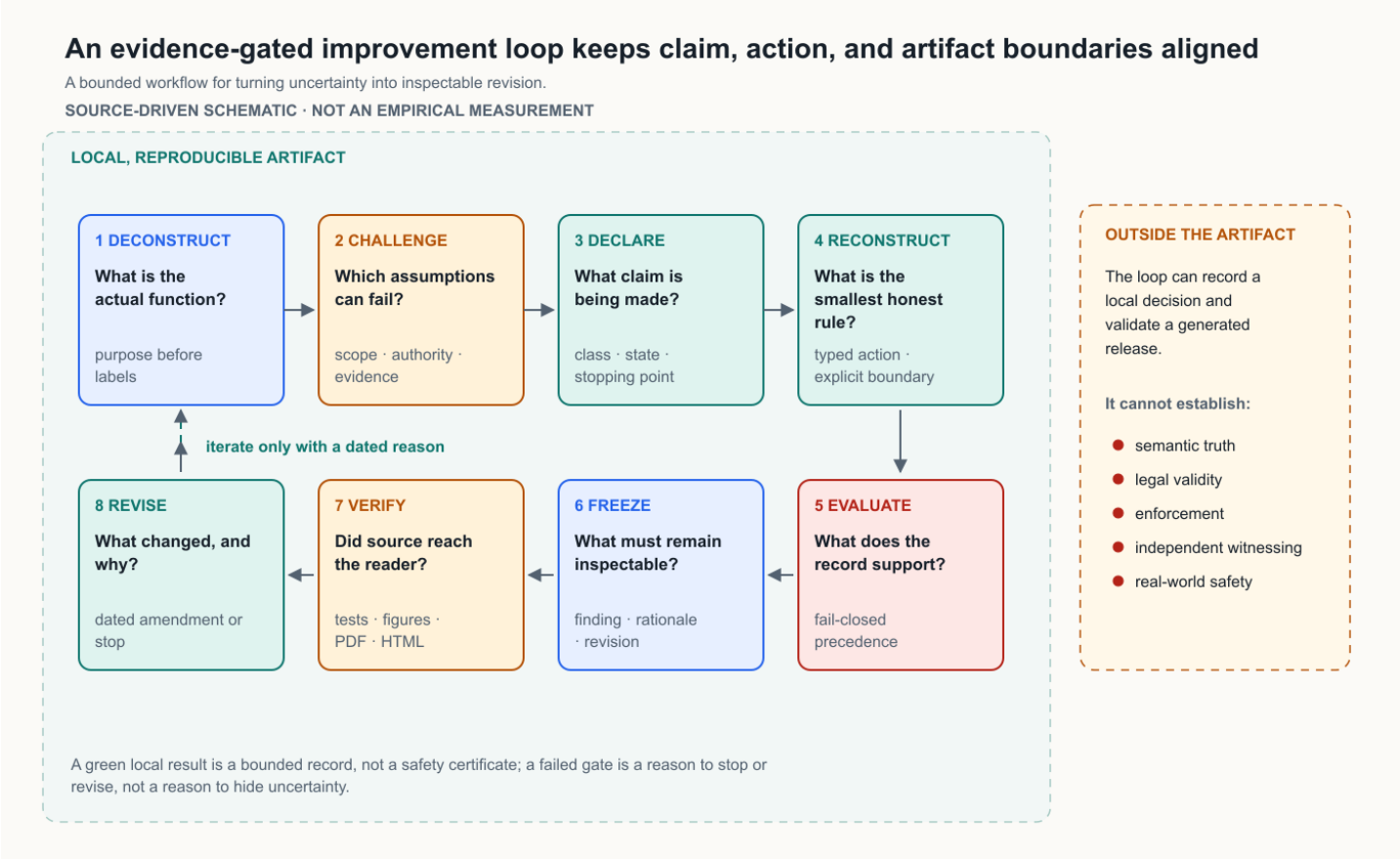


Figure 6: An evidence-gated improvement loop keeps claim, action, and artifact boundaries aligned. The eight-step schematic moves from deconstruction and challenge through claim declaration, reconstruction, fail-closed evaluation, frozen records, source-to-render verification, and dated revision. The outside panel names semantic truth, legal validity, enforcement, independent witnessing, and real-world safety as claims this local loop cannot establish. It is a reproducible method map, not a measurement of improvement or safety.

The frozen `ReviewFinding` carries stable reason-code values alongside the display prose. This is a small but important audit distinction: prose can be clarified without forcing downstream consumers to parse wording, while the codes make a block such as `STALE_EVIDENCE` or `UNEXEMPTED_LINE` available for aggregation and regression checks.

7.4 Falsification and negative controls

The method is credible only if it can fail visibly. The release checks should include negative controls that deliberately attempt to cross the boundary:

- remove a required claim class, stopping point, or verification mode;
- add a duplicate or unknown claim identifier to the prose register;
- describe a stale, self-asserted, or contradicted evidence record as current;
- make the figure registry and visualization brief disagree;
- alter a source-driven figure and test whether byte determinism detects the difference; and
- render without a required artifact and verify that the release manifest refuses readiness.

These controls test the instrument's refusal behavior. They do not estimate false-positive or false-negative rates in the world, because this project has no labeled corpus of real engagements and no independent operational observer. The appropriate conclusion is narrower: the local gates either detect the planted defect or they do not.

7.5 Scope of inference

The method can support a statement such as: "given this registry, this typed intake, this review date, and these recorded evidence statuses, the evaluator returned this classification." It cannot support: "the action is safe," "the action is lawful," "the source endorses the framework," or "the rendered artifact has been independently certified." Those stronger conclusions require different authorities and evidence. The project is most coherent when it keeps that stopping point attached to every claim, figure, and release record.

8 The Adaptation Thesis

Turner's framework demonstrates a pattern: make a narrow refusal explicit, grade release by retained oversight, issue a review finding, and preserve a durable record [Turner, 2026]. Red Line adapts that pattern to one practitioner. The transfer is methodological, not institutional. The first-principles reconstruction in sec. 6 is the controlling logic; the Turner comparison is useful only where it clarifies a mechanism.

The personal version changes the decision boundary in an important way. A proposed action is not allowed to reach policy matching on the strength of a description or a self-selected exemption token. It must carry an `ActionContext` and evidence records for purpose, end use, affected parties, data provenance, legal basis, human control, deployment, downstream transfer, and capability scope. That is an implementation choice made here in response to the risk of scope laundering and false certainty; it is not attributed to Turner. The evidence gate is a refusal to manufacture a green result, not a claim that the local verifier has discovered the truth of the records.

8.1 Four adaptations

1. The standing Review Body becomes a written self-review record plus an external-witness hook. The absence of independent enforcement is explicit.
2. Deployment tiers become grades of how much the author can observe, update, suspend, or withdraw a work product after release.
3. The registry retains narrative carve-outs but implements narrowing clauses as typed exemptions requiring verified evidence.
4. Durability becomes deterministic hashing and canary verification, while the successor rationale makes registry changes visible rather than forbidden. The hash is a change signal, not a signature or an enforcement mechanism.

The result is not a private morality engine. It is a reproducible local review instrument whose strongest claim is: "given this registry, this intake, this review date, and these recorded evidence statuses, this is the documented result." A reader should be able to identify which parts are source-derived, which are the author's commitments, and which are implementation facts.

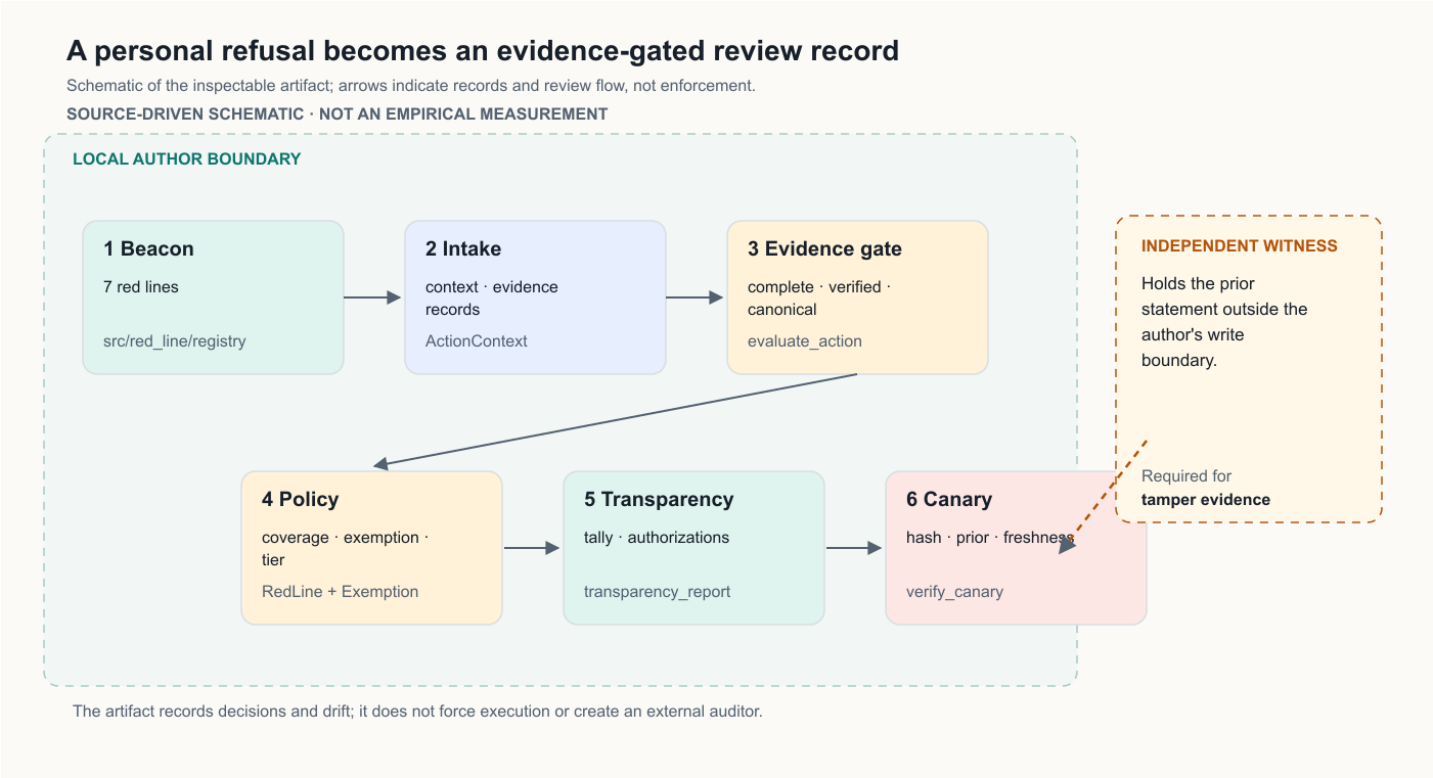


Figure 7: The framework turns a private commitment into an evidence-gated review event. The deterministic architecture schematic draws six stages inside the local author boundary: the seven-line registry beacon, the action-declaration intake, the evidence gate, the policy match over lines and typed exemptions, the transparency tally, and the canary check. The external-witness boundary remains outside the local author boundary. Arrows describe records and accountability flow, not enforcement; a green result is not a safety certification.

9 The two CANARY-severity boundaries

Turner's framework gives special attention to human control over harm-capable systems and to individualized scrutiny that is not generated from bulk data or group membership [Turner, 2026]. Red Line carries these as its two CANARY-severity records, `s1-human-control-force` and `s2-untargeted-profiling`, while making clear that the personal registry is not Turner's institution or legal standard.

9.1 S1 — force and harm-capable systems

The first line refuses building, tuning, or knowingly supplying a component to a system that selects and engages targets for force without appropriate human control over each engagement. In this revision, "human control" is not proved by the token `human_in_loop`. A compliant typed exemption requires verified context about the accountable human, the end use, the purpose, and the operational ability to review the action. The local evaluator can require those records; it cannot inspect a deployment and decide that the judgment was genuinely independent or operationally meaningful.

The typed exemptions cover defensive alerting and adjacent support such as logistics, translation, maintenance, research, and accountable intelligence analysis. The exemption trigger is only a declaration; the required evidence must still be verified. Multiple prohibited dimensions resolve to `REQUIRES_MODIFICATION`, and an unsatisfied exemption never softens the line. The source framework's more specific requirements around proportionality, operational tempo, and legal transparency remain questions for human review, not claims discharged by this lexical interface [Turner, 2026].

9.2 S2 — untargeted profiling

The second line refuses converting bulk data into individualized intelligence on people who are not already subjects of a specific, lawful, individualized process. Turner's source formulation makes particularization important: data about a person does not itself authorize an individualized assessment, and an assessment should not be initiated solely from a demographic, national-origin, religious, or political category [Turner, 2026].

Red Line therefore requires evidence for affected parties, provenance, legal basis, and purpose even where an action declares `aggregate_research`, `identified_subject`, or `opt_in_analytics`. These tokens select a possible exemption path; they do not create consent or legality. The local line also does not implement the source framework's full proportionality, territory, independent-basis, or downstream institutional controls. That omission is a scope boundary, not an implied approval.

9.3 Canary status

Both records remain CANARY-grade. Here "CANARY" is a change-signal grade, not a ranking of moral importance. Structural invariants require that both ids remain present, retain CANARY severity, and never allow an AIR_GAPPED ceiling. Their narrative carve-outs remain readable in the beacon, while their executable exemptions and evidence requirements are included in the canonical registry digest.

10 Deployment tiers as oversight-retention grades

Turner relates deployment eligibility to the oversight retained over a system after delivery [Turner, 2026]. Red Line reinterprets that mechanism for a single practitioner:

- `HOSTED` retains author-operated observation and withdrawal;
- `CONNECTED` retains a maintained update or suspension path;
- `AIR_GAPPED` means beyond recall, such as an unrestricted release or handed-off model.

The `oversight_rank` values are 2, 1, and 0 respectively. Each registry line declares a `max_tier`; despite that field name, it is used as a minimum retained- oversight floor: the least-oversight environment in which a fully evidenced typed exemption may operate. A tier floor is never a permission by itself: scope, context, and exemption evidence still govern. Definition 1 and Definition 13 state the rank and the floor test exactly as the code applies them.

The evaluator is monotonic in danger. An unexempted implicated line is `NON_COMPLIANT` at every tier; reducing retained oversight adds an aggravating reason but cannot turn a hard block into a softer outcome. A verified exemption below its line's floor becomes `REQUIRES_MODIFICATION`. Proposition 8 states this as a property and names the test that exercises it. It is a consistency property of the local decision procedure, not empirical evidence that a hosted system is safe or that an air-gapped system is dangerous in every circumstance.

"Air-gapped" is shorthand for beyond the author's practical ability to observe, update, suspend, or withdraw the work product. It is not a security certification and does not mean that a released artifact is literally disconnected in every deployment.

Figure 8: Less retained oversight narrows the release envelope. The registry-derived matrix marks each line's oversight floor and labels the tiers directly; a filled cell means only that the tier floor is met. It does not mean an action is compliant, safe, or legally permitted. The two CANARY records remain visibly bounded away from AIR_GAPPED release by an executable invariant. Shape, text, and cell state repeat the meaning so the figure remains interpretable in grayscale and on a narrow page.

The monotonicity property itself is exercised, not merely asserted. The analysis module `red_line.analysis.monotonicity` sweeps all 36 line/keyword slots (34 distinct tokens) across the seven current lines through the real `evaluate_action` at each of the three deployment tiers — 108 executed evaluations with a fully evidenced fixture intake — and records the verdict lattice the evaluator actually returned, with zero inversions. The slot count exceeds the vocabulary because `handoff` and `provenance` are each declared by two lines and are therefore swept once per line.

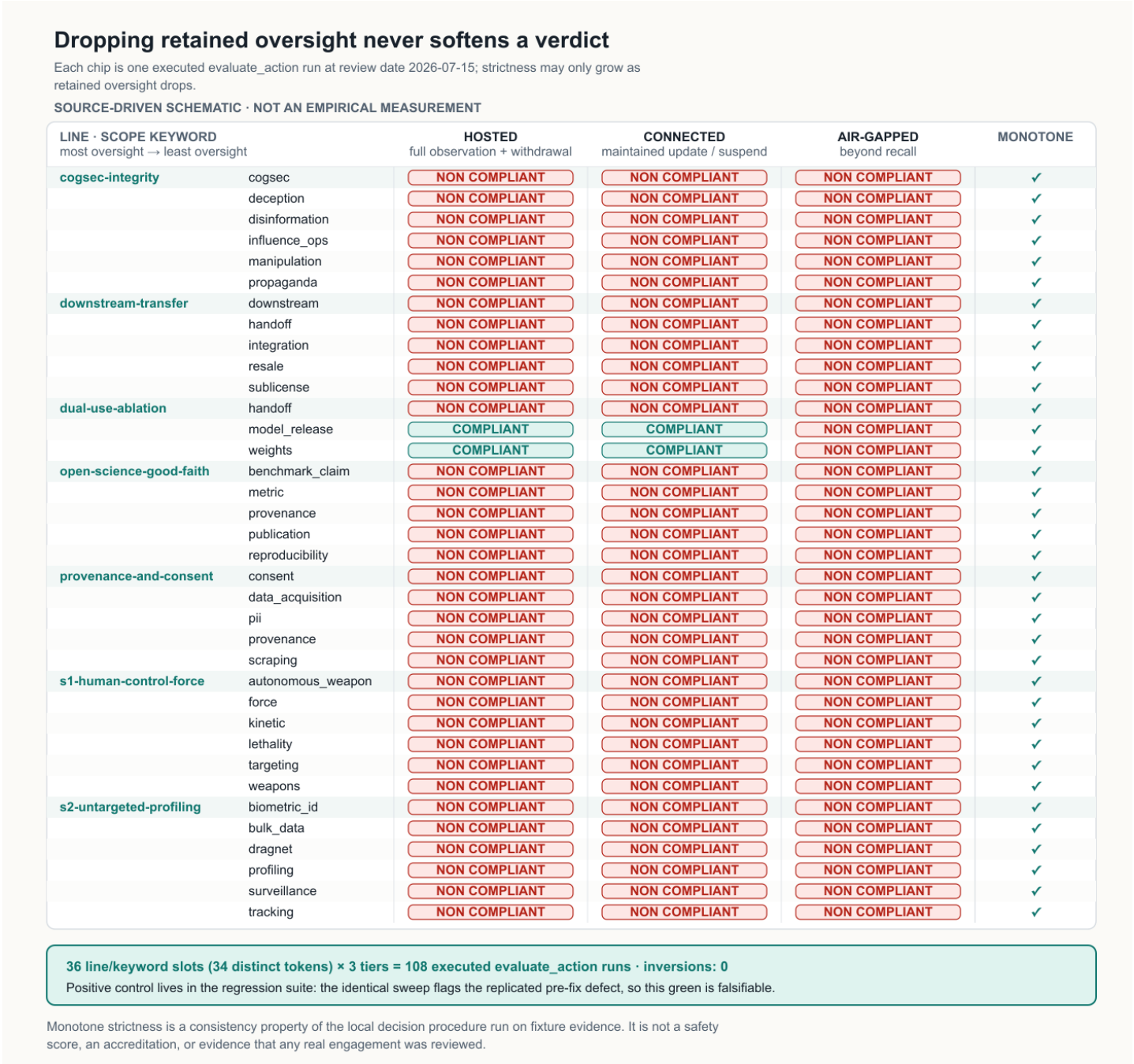


Figure 9: Dropping retained oversight never softens a verdict. Executed verdict-strictness lattice computed by `red_line.analysis.monotonicity.run_monotonicity_sweep`: all 36 line/keyword slots (34 distinct tokens) across the seven current lines run through the real `evaluate_action` with a fully evidenced fixture intake at each of the three deployment tiers — 108 executed evaluations — and every chip is the classification actually returned. Read left to right, a row moves from most to least retained oversight and strictness never decreases; the sweep records zero inversions, and its regression test carries a positive control that detects the replicated pre-fix defect. Monotonicity is a consistency property of the local decision procedure exercised with fixture evidence, not a safety measurement or a review of any real engagement.

11 Written review and transparency

Turner uses a standing Review Body and a fixed vocabulary of findings [Turner, 2026]. Red Line cannot recreate that institutional separation. It keeps the durable local outputs: a written finding for each reviewed action and a tally that makes classifications and escalations visible.

11.1 Finding record

`review_engagement` returns a frozen `ReviewFinding` containing the engagement, classification, implicated line ids, finding text, review date, declared scope, tier, and ambiguity flag. The evaluator is run as of that review date, so a finding cannot silently describe one date while evaluating freshness at another. The finding preserves `INSUFFICIENT_INFORMATION` as a blocking result; it does not turn incomplete intake into a policy judgment.

The result vocabulary is:

Result	Meaning
`INSUFFICIENT_INFORMATION`	required context or evidence is unresolved; stop
`NON_COMPLIANT`	a line is implicated without a verified narrowing condition
`REQUIRES_MODIFICATION`	a verified exemption still has a dimension or tier problem
`COMPLIANT`	complete intake and verified exemption satisfy the local registry
`OUTSIDE_SCOPE`	complete intake, but no current registry line applies

Outside scope is not a compliance certification. It says only that this registry did not match the documented action. A transparency report aggregates the findings supplied to it; it is not an automatic publication channel or a third-party audit.

11.2 Escalation is not permission

The former boolean override was too easy to misread as authorization. The new `ReviewAuthorization` records `authorized_by`, `authority`, `rationale`, and `recorded_on`. It is an escalation or remediation record. A finding remains blocking when its classification is `NON_COMPLIANT`, `REQUIRES_MODIFICATION`, or `INSUFFICIENT_INFORMATION`; the transparency report counts the named authorization without changing that fact.

This is a deliberate difference from institutional governance. The author can make an exception visible, but cannot make the personal “No” disappear by calling it an override.

12 Durability and transparency: a hash-based canary

Turner's framework protects the *durability* of a red-line commitment through procedure: a material modification requires advance notice with a rationale, and an impairment of the Review Body's capacity must be disclosed [Turner, 2026]. A single practitioner has no review body and no counterparty to notify. The personal analog is therefore cryptographic rather than procedural — it makes weakening the standard visible and auditable, not procedurally gated.

12.1 Deterministic registry hashing

`registry_hash` reduces the red-line registry to a single SHA-256 digest over its canonicalized content. Canonicalization sorts the lines by `id`, serializes each fixed payload including narrative carve-outs and typed exemption ids, trigger scopes, and required evidence kinds, and dumps the whole as JSON with stable separators. No timestamps or environment inputs enter the payload, so the hash is a pure function of registry *content*: the current seven lines yield the pinned digest `72835fd8…f5aad7`, and any change to a line's policy semantics changes it. `line_digest` applies the same atom to a single line. This is a cryptographic checksum, not a digital signature: it does not authenticate who issued a statement or make a rewrite impossible.

12.2 The canary statement

A `CanaryStatement` is a dated attestation. It binds the registry digest, the set of line ids present at issue time, and a tuple of per-line `(id, severity, digest)` triples, all captured on a stated `issued_on` date. Binding the issue-time severity alongside each digest — not merely the aggregate hash — lets a later check tell *which* line changed and how severe it was when the author last vouched for it. `issue_canary` also enforces a successor guard: when a prior canary is supplied and the hash has drifted, a silent re-issue is refused unless the author supplies a rationale, which is folded into a self-describing successor statement naming the superseded digest and the removed or added ids.

12.3 Verification and escalation

`verify_canary` compares a prior statement to the live registry. It reports `drift` (any hash change), explicit `removed_ids` and `modified_ids`, and `stale` independently, so a caller sees *how* a canary failed. `intact` requires an unchanged hash, a fresh attestation, complete line-id agreement, and internally consistent populated per-line metadata. Freshness is always evaluated against a 180-day window and fails closed: a missing, future-dated, or unparseable date reads as stale rather than as a silent pass. When a removed or modified line was CANARY-grade at issue time, the detail escalates to `CANARY-GRADE LINE ALTERED`, surfacing it above any lower-severity change.

12.4 The honest trust model

This is the *pattern* of a warrant canary, not the legal instrument: a warrant canary's force rests on a legal asymmetry that does not apply to a personal commitment. The statement is forgeable by anyone with write access to the registry, and the author is also its only enforcer. Its evidential force rests entirely on an external prior copy — a git-committed or otherwise independently held statement, checked by someone other than the author. The registry is pre-publication and there is no external verifier. The instrument makes weakening the standard tamper-evident under that condition; it does not prevent it, and that limitation is disclosed rather than papered over.

An external verifier should: obtain a prior statement from outside the author's write boundary; recompute the current registry hash and per-line metadata; run the freshness and drift check; inspect any successor rationale; and report the result without treating a regenerated current statement as independent evidence. The full command-level procedure is in `docs/VERIFY.md`.

The trust boundary is therefore more important than the hash itself; fig. 10 shows what the mechanism can and cannot signal.

12.5 Defensive security context

The nation-state lens is applied here as a defensive threat model for a static private artifact, not as a claim that the repository is an APT-resistant service. The relevant crown jewels are the registry, evidence and review records, prior canary, private scholarship, release boundary, and rendered outputs. A patient adversary could alter a privileged checkout, poison a dependency or renderer, rewrite a same-author canary fixture, or make a source appear more authoritative than it is. The threat model records these paths and their residual limits in `docs/security-threat-model.md`.

NIST's Secure Software Development Framework, MITRE ATT&CK, and SLSA provide useful vocabularies for provenance, adversary behavior, and future artifact attestation [Souppaya et al., 2022, MITRE, 2026, SLSA Community, 2026]. They are implementation context, not evidence of legal compliance, secure operation, or resistance to a nation-state actor. This project currently demonstrates dependency-light deterministic domain logic, locked development dependencies, local canary and render

validation, and explicit external-witness limitations; it does not claim signed provenance, a generated SBOM, hermetic builds, or runtime telemetry.

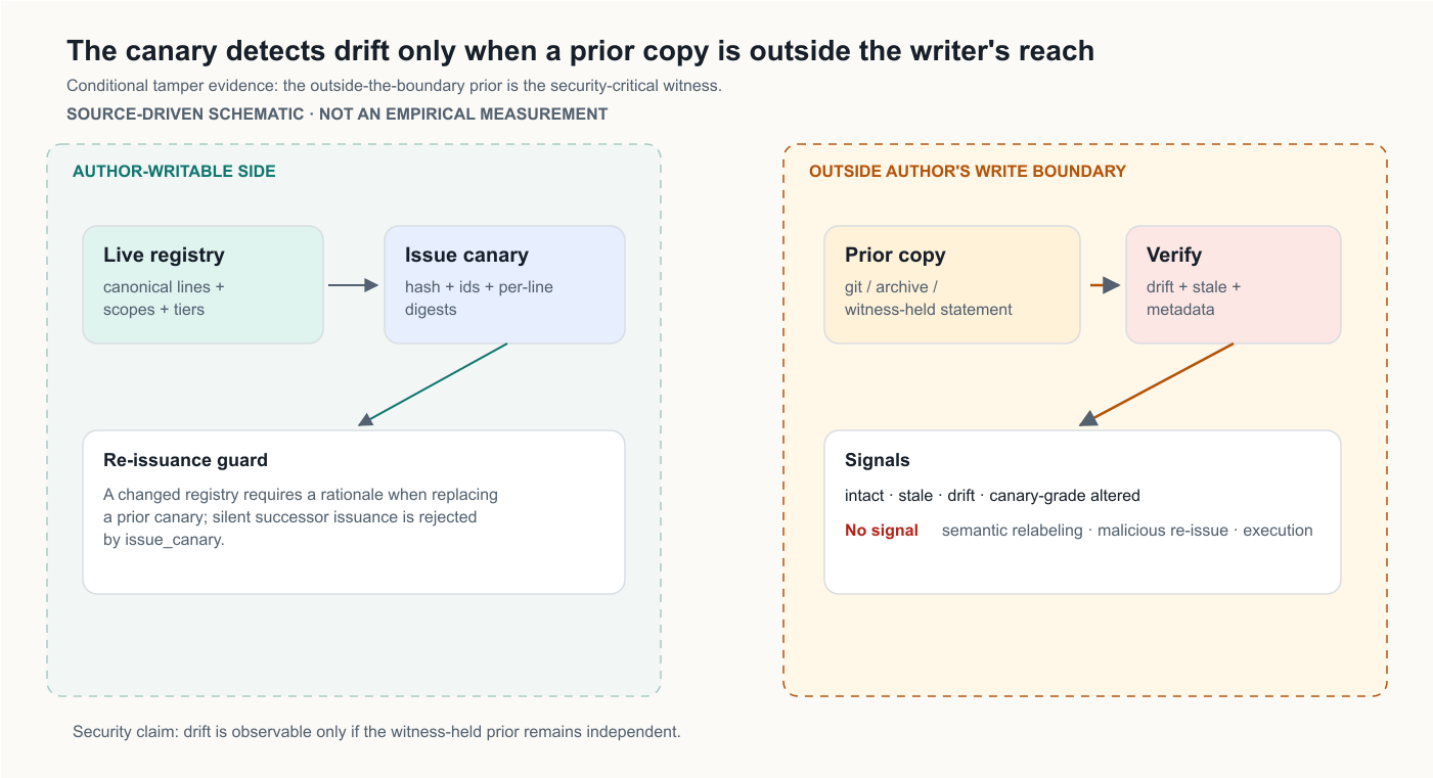


Figure 10: The canary detects drift only when a prior copy is outside the writer's reach. Trust-boundary schematic for the canary mechanism: issuance binds the registry to an aggregate digest and per-line metadata; verification compares those values with the live registry and checks freshness. The dashed boundary is the critical condition: the prior statement must be held by someone or some system that cannot be rewritten by the same author who edits the registry. A canary cannot detect semantic violations hidden by misleading labels, cannot stop a malicious re-issuance, and cannot enforce an action finding. It is tamper evidence conditional on an independent witness, not prevention or legal protection.

13 Evidence, ambiguity, and evaluation

The evaluator is a staged method, not a semantic safety classifier. The first stage asks whether the action can be reviewed at all. `ActionContext` requires nine dimensions: purpose, end use, affected parties, data provenance, legal basis, human control, deployment, downstream transfer, and capability scope. Each must have a `VERIFIED` evidence record, including an explicit supported not-applicable value. Self-assertion, unresolved contradiction, unknown scope, empty scope, or explicit ambiguity returns `INSUFFICIENT_INFORMATION`. The gate is conjunctive over all nine and it says which one it stopped on; Proposition 3 and fig. 14 carry the executed sweep behind that claim.

The evaluator first attempts canonical normalization of the declared scope; malformed or non-ASCII tokens stop the intake rather than becoming new vocabulary. It then checks the evidence gate before matching registry coverage, applying typed exemptions, checking verified evidence for the exemption, and enforcing the deployment tier. Normalization is input hygiene, not a policy verdict. The description is useful for human reading and mismatch hints; it cannot satisfy the scope or evidence gate.

The distinction matters: `OUTSIDE_SCOPE` means the complete intake implicated no current line, while `COMPLIANT` means at least one line was implicated and a verified exemption plus tier requirements satisfied it. A missing legal basis is neither: it is an information stop. The precedence is therefore: intake defect first; then an unexempted line; then a verified but modified or under-tier use; then a fully narrowed implicated line; and only then outside scope. Proposition 2 and Proposition 4 state the same rule formally, each bound to a test that re-derives the branch order from the evaluator source rather than restating it.

This ordering is monotone in severity: the intake gate short-circuits before any policy matching, and the reduction over implicated lines always resolves to the single most severe applicable classification, so a less severe outcome never overrides a more severe one. fig. 11 reads that precedence top to bottom, from the `INSUFFICIENT_INFORMATION` short circuit down through the hard block, the modification requirement, the narrowed compliant result, and the distinct `OUTSIDE_SCOPE` terminal.

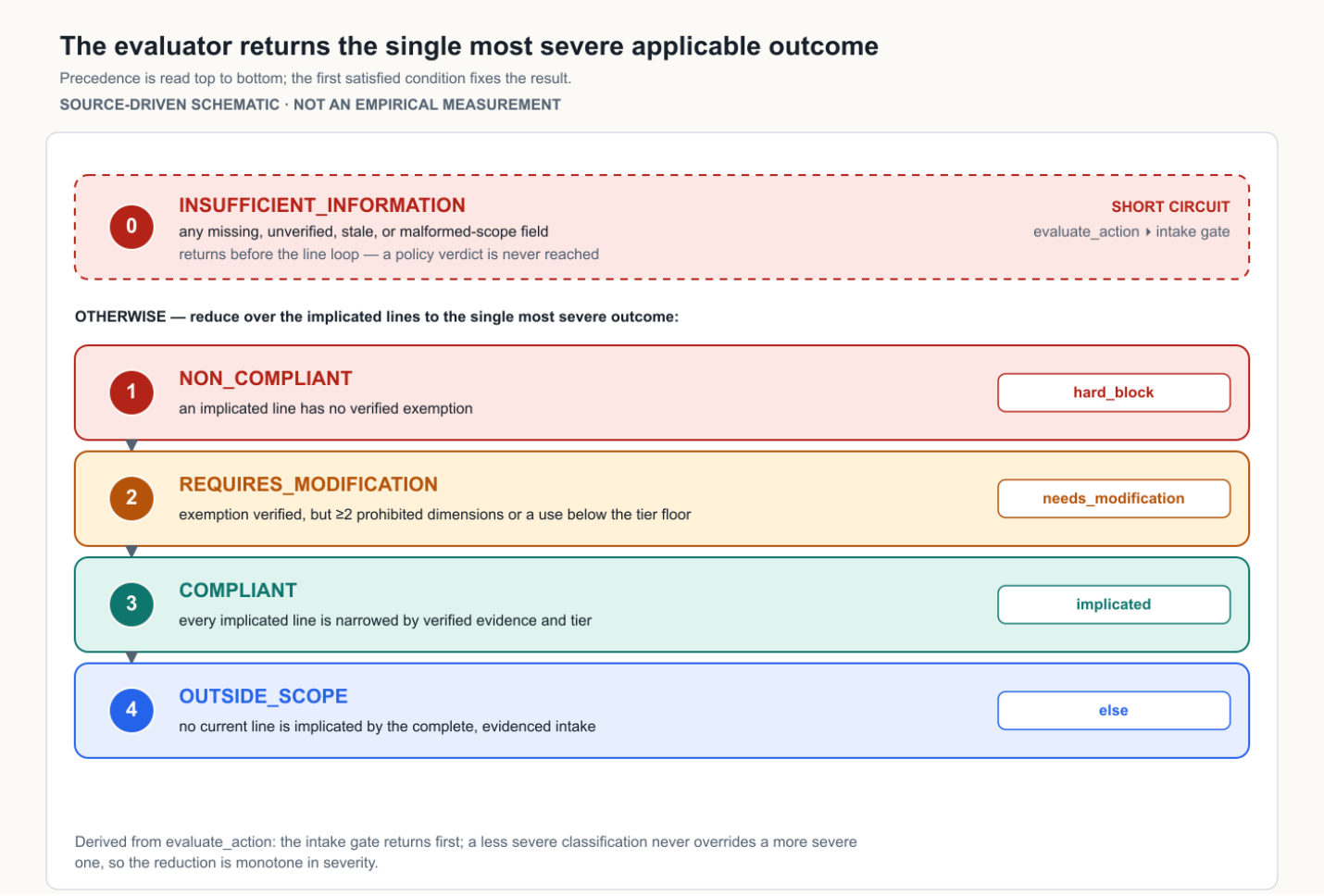


Figure 11: The evaluator returns the single most severe applicable outcome. Outcome-precedence ladder generated from the `evaluate_action` control flow: the intake gate returns `INSUFFICIENT_INFORMATION` before policy matching, after which the reduction over implicated lines resolves to one classification ordered by severity — a hard block dominates a modification requirement, which dominates a compliant result, which dominates `OUTSIDE_SCOPE`. A less severe outcome never overrides a more severe one. The ladder is a control-flow schematic, not empirical evidence that any action was safe, lawful, or correctly described.

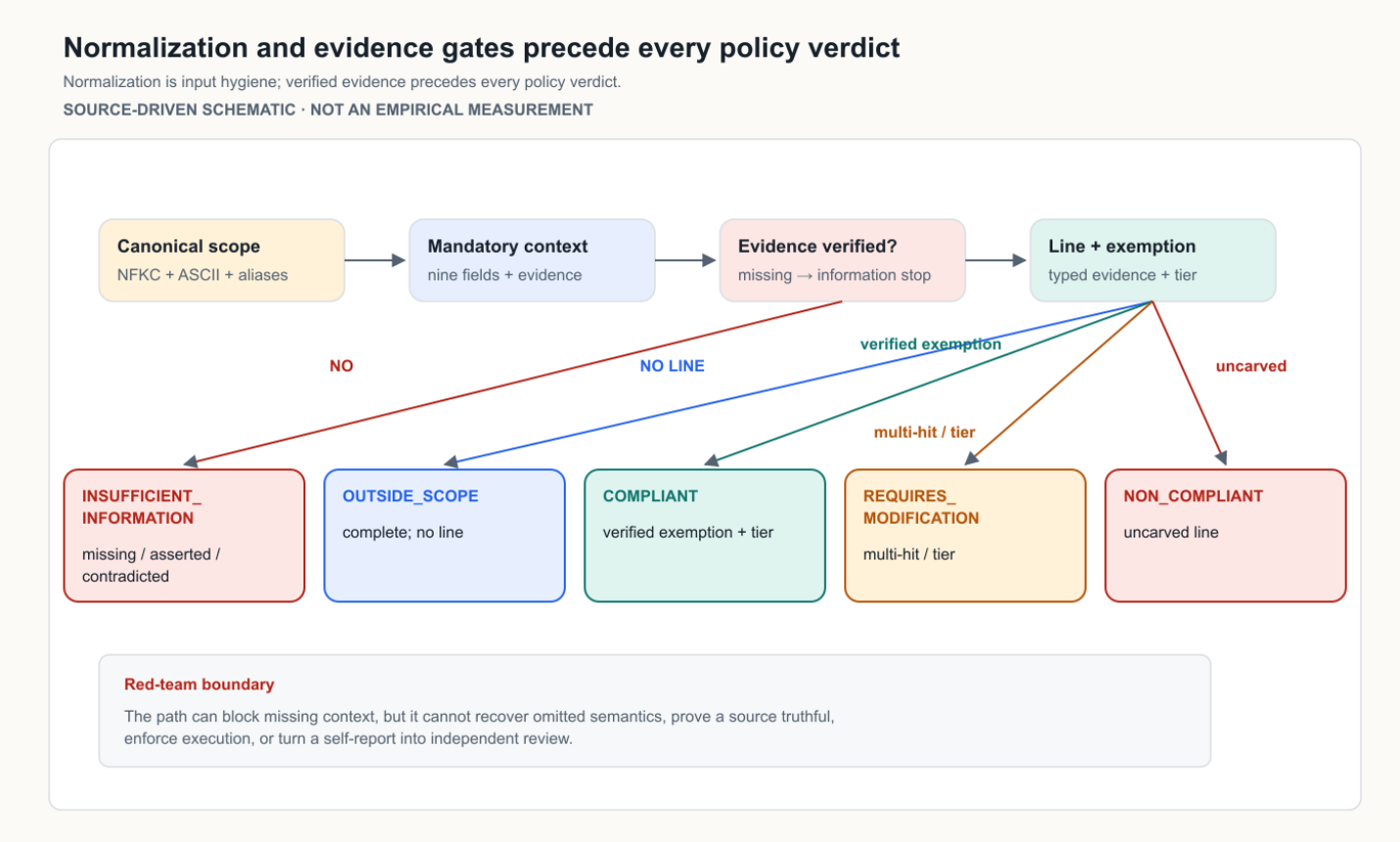


Figure 12: Normalization and evidence gates precede policy verdicts. This deterministic method map shows the reading path from canonical input normalization through mandatory context, required evidence, line coverage, typed exemption, and tier checks. The `INSUFFICIENT_INFORMATION` branch is deliberately upstream of compliance; `OUTSIDE_SCOPE` is a distinct terminal result. The figure is generated from the evaluator contract, not from observed safety data, and the caption's distinction is essential when the visual is read without the surrounding prose.

13.1 Exercised outcome coverage

The two figures above are schematics of intended control flow. The five-outcome claim is separately executed rather than drawn: the harness `red_line.analysis.outcome_coverage.run_outcome_coverage` runs a deterministic battery of five `ProposedAction` fixtures — one designed for each classification — through the real `evaluate_action` against the live registry at a fixed review date (2026-07-15). In the current registry all five of the five classifications are reached and every case lands on its intended outcome, with the evaluator's stable reason codes recorded per case. The same harness honestly reports partial coverage when it should: run against an empty registry, the three implication-dependent outcomes (`COMPLIANT`, `REQUIRES_MODIFICATION`, `NON_COMPLIANT`) become unreachable, and a review date a year later downgrades every fixture's evidence to stale, collapsing even the compliant case to an information stop. fig. 13 renders the executed report.

The reason codes returned by the executed battery also confirm that each case reaches the intended terminal by the intended *route*. The unevidenced case stops with `missing_evidence` and `intake_blocked` — the short circuit, not a policy verdict. The out-of-scope case returns the single code `outside_scope`. The compliant case carries both `verified_exemption` and `all_lines_narrowed`, meaning a line was implicated and then narrowed rather than never matched. The modification case reports `multiple_prohibited_dimensions` — a verified exemption undercut by extra prohibited scope on the same line — and the blocked case reports `unexempted_line`. A future refactor that preserved the five terminals but altered the paths to them would surface here as a changed code set.

Reachability is a structural property of the evaluator's control flow. The battery's evidence records are harness fixtures, so a complete report is a regression pin on the evaluator — it is not evidence that any real engagement was reviewed, safe, lawful, or correctly described.

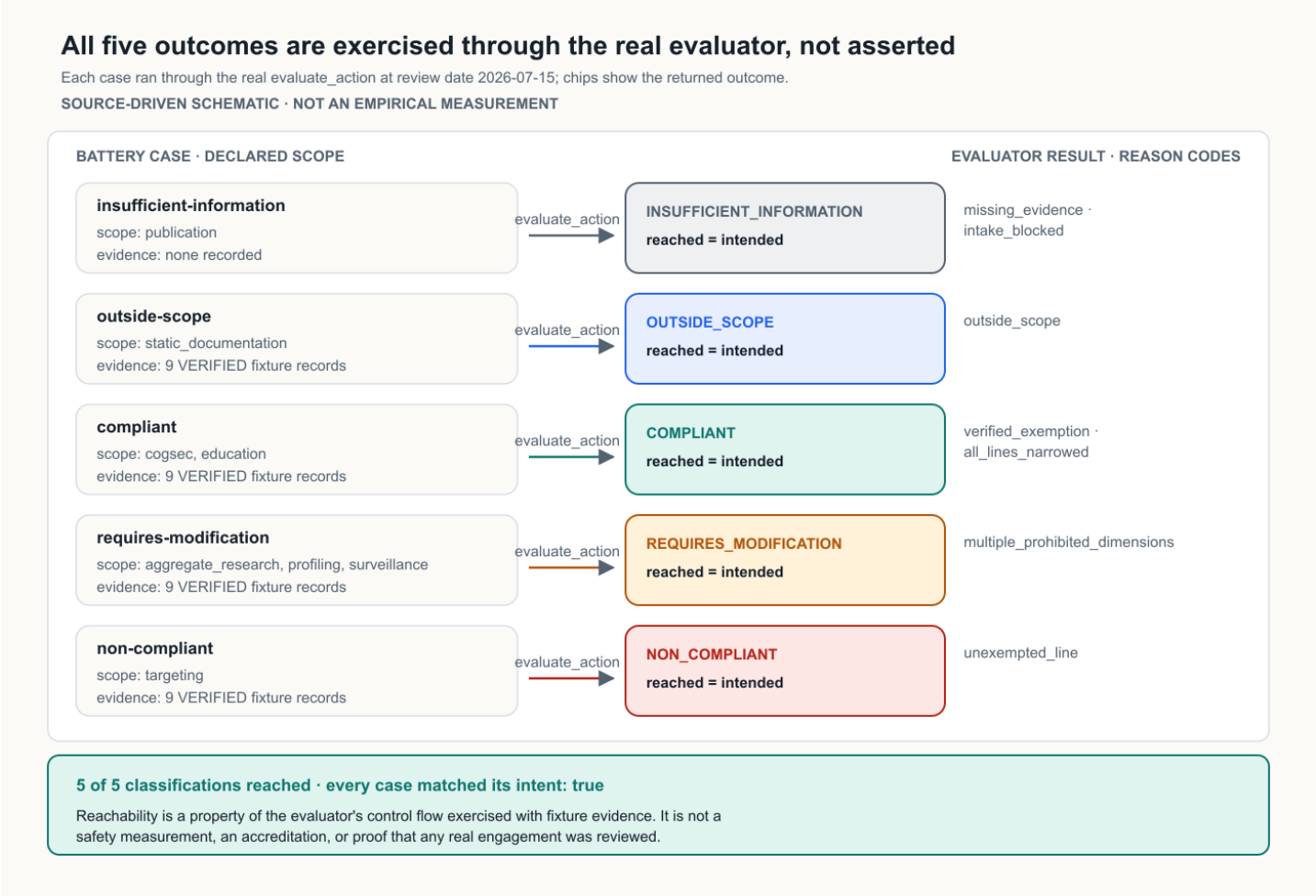


Figure 13: All five outcomes are exercised through the real evaluator, not asserted. Outcome-coverage plate computed by `red_line.analysis.outcome_coverage.run_outcome_coverage`: each battery case runs through the real `evaluate_action` at the fixed review date, and the chip on the right is the classification actually returned with its stable reason codes. All five classifications are reached and every case matches its intended outcome, upgrading the five-outcome claim from a control-flow assertion to an executed, regression-pinned property. The fixture evidence is not real-world verification, and the plate is not a safety measurement.

13.2 Residual lexical limitation

Explicit aliases close trivial singular/plural and punctuation near-misses, but they do not infer meaning. A complete-looking action can still misdescribe its capability or evidence. That is why the manuscript calls the output a local auditability result and requires independent witness/review before making a stronger public assurance claim.

The result is also bounded by the current registry vocabulary. A newly named capability, an unfamiliar deployment pattern, or a misleading but complete evidence packet can escape the intended semantic category. This is why the instrument treats `OUTSIDE_SCOPE` as a bounded registry statement and why the release claim remains auditability rather than safety certification.

That vocabulary is small enough to print in full, and fig. 17 does. Thirty-four words decide what this evaluator can see; two of them belong to more than one line. A reader who wants to know what `OUTSIDE_SCOPE` is bounded by can read the whole boundary off one grid rather than take the caveat on trust.

14 The instrument stated formally

The preceding sections describe the evaluator in prose. Prose about a decision procedure drifts from the procedure, and the reader has no way to see the drift. This section restates the same objects and the same decision rule as numbered definitions and propositions taken from the shipped code, and binds each proposition to a named test that fails if the two disagree. Nothing here is new policy. It is the existing instrument written so that a mismatch is a red test rather than an unnoticed sentence.

Two limits carry through every result below. Each is a statement about a local program's behaviour on declared inputs, never about the world those inputs describe; and each is scoped to the registry version pinned by the digest in sec. 15, because these are properties of a decision procedure applied to particular content.

14.1 The domain objects

Definition 1 (Deployment tier and oversight rank). The deployment tiers are $T = \{\texttt{hosted}, \texttt{connected}, \texttt{air\_gapped}\}$, ordered by a retained-oversight rank $\rho : T \to \{0, 1, 2\}$ with $\rho(\texttt{air\_gapped}) = 0$, $\rho(\texttt{connected}) = 1$, and $\rho(\texttt{hosted}) = 2$. A higher rank means more oversight the author still holds after release: the ability to observe the work product, update it, suspend it, or withdraw it.

Definition 2 (Severity grade). The severity grades are $\{\texttt{canary}, \texttt{absolute}, \texttt{strong}\}$. The grade records how a change to the line itself is treated, not how bad a breach would be: a `canary` line's removal or demotion is itself the reportable event.

Definition 3 (Evidence status). The evidence statuses are $\{\texttt{VERIFIED}, \texttt{SELF\_ASSERTED}, \texttt{UNVERIFIED}, \texttt{CONTRADICTED}\}$. Only `VERIFIED` can support a required dimension. The evaluator does not decide whether a source is true; it records whether a reviewable artifact was checked.

Definition 4 (Intake dimension). The intake dimensions are the nine-element set K:

$$K = \{\ \texttt{purpose},\ \texttt{end\_use},\ \texttt{affected\_parties},\ \texttt{data\_provenance},\ \texttt{legal\_basis},\ \texttt{human\_control},\ \texttt{deployment},\ \texttt{downstream\_transfer},\ \texttt{capability\_scope}\ \}$$

Their enum order is the intake order, and it is the column order of every derived matrix in this paper.

Definition 5 (Evidence record and staleness). An evidence record is a tuple (k, r, s, σ, d) of a kind $k \in K$, a reference r, a summary s, a status σ, and an ISO date d. Given a review date a and a window w (default $w = 180$ days), the record is *stale* when $d > a$ or $a - d > w$. A future-dated record is therefore stale in the same way an expired one is.

Proposition 1 (The staleness boundary is exclusive). Staleness uses the strict comparison $a - d > w$ of Definition 5, so a verified record dated exactly w days before the review date is fresh, and a record dated $w+1$ days before is stale. The boundary is the same strict-exclusive semantics as black_line's currentness boundary and golden_line's temporal boundary: the window's last fresh day is age w, and the first stale day is $w+1$. The threshold w is a review-cadence choice, not a decay law, and a future-dated record is stale regardless of the window.

Definition 6 (Action context). An action context C assigns a string to each of the nine dimensions of Definition 4, and carries a finite tuple of evidence records in the sense of Definition 5 and a finite tuple of declared unknowns. C *supports* a dimension k at review date a when it holds a record of kind k whose status is `VERIFIED` and which is not stale. A dimension whose assigned string is empty or one of `unknown`, `unspecified`, `tbd`, `unclear` is unsupported whatever its records say; `not_applicable` is a permitted value, but it is still only supported when a verified record backs it.

Definition 7 (Scope normalization). A declared token is canonicalized by NFKC normalization, case folding, an ASCII check that raises on anything else, replacement of every non-alphanumeric character by an underscore, collapse of repeated underscores, stripping of leading and trailing underscores, and a final lookup in a reviewed alias table. There is no stemming: a new synonym must be added to the table by hand. $\mathcal{N}(S)$ denotes the canonical token set of a declared scope S.

Definition 8 (Typed exemption). An exemption is a tuple $e = (\mathrm{id}, \mathrm{description}, T_e, R_e, m_e)$ where T_e is a non-empty trigger scope, $R_e \subseteq K$ is the required evidence, and $m_e \in \{\texttt{any}, \texttt{all}\}$ is the match mode. For a declared scope S, with $\mathcal{N}$ the normalization of Definition 7,

$$\mathrm{matches}_e(S) \;=\; \begin{cases} \mathcal{N}(T_e) \subseteq \mathcal{N}(S) & m_e = \texttt{all} \\ \mathcal{N}(T_e) \cap \mathcal{N}(S) \neq \emptyset & m_e = \texttt{any} \end{cases}$$

and e is *satisfied* by a context C at review date a when C supports every $k \in R_e$. Matching is a declaration; satisfaction is the condition that can narrow a line.

Definition 9 (Red line). A red line is an eleven-field record: a slug id, a title, a first-person standard beginning `I`, a rationale, a coverage scope, narrative carve-out clauses, a tier floor `max_tier`, a severity of Definition 2, the person stating it, the ISO date it was stated, and a tuple of typed exemptions. A line ℓ *covers* a declared scope S when $\mathcal{N}(\text{scope}_\ell) \cap \mathcal{N}(S) \neq \emptyset$. The carve-out clauses are prose for a reader; only the typed exemptions execute.

Definition 10 (Proposed action and effective scope). A proposed action is a tuple $(\text{description}, S, C, t, \text{amb})$ of free text, a declared scope, a context, a tier $t \in T$ as ranked in Definition 1, and an ambiguity flag. Coverage and exemption matching both run against the *effective* scope $E = \mathcal{N}(S) \cup \{t\}$: the tier value is itself a token, so a line or exemption may name a tier in its scope. The description is never part of E.

Definition 11 (Classification). The classifications are the five-element set C:

$$C = \{\ \texttt{COMPLIANT},\ \texttt{REQUIRES\_MODIFICATION},\ \texttt{NON\_COMPLIANT},\ \texttt{OUTSIDE\_SCOPE},\ \texttt{INSUFFICIENT\_INFORMATION}\ \}$$

`OUTSIDE_SCOPE` is a finding of absence after a complete inspection; `INSUFFICIENT_INFORMATION` is a refusal to inspect. Neither is permission.

Definition 12 (Intake defect). An action has an intake defect at review date a when any of the following holds: a declared token fails normalization; some dimension is unsupported in the sense of Definition 6; some dimension holds a record that is unresolved (self-asserted, unverified, or contradicted in the sense of Definition 3, with no verified record standing in its place) or stale; the context declares an unknown; the normalized scope is empty or contains one of the markers `unknown`, `unspecified`, `tbd`, `unclear`; or the ambiguity flag is set.

Definition 13 (Tier floor). Each line declares a floor `max_tier`. An action at tier t is *below the floor* of line ℓ when $\rho(t) < \rho(\text{max\_tier}_\ell)$, with ρ the rank function of Definition 1 — it retains strictly less oversight than the line requires. The field name reads as a ceiling and behaves as a floor on retained oversight; the code, not the name, is the definition.

Definition 14 (Strictness order). On the three verdicts a fully evidenced, line-implicating intake can reach, strictness is ranked `COMPLIANT` $<$ `REQUIRES_MODIFICATION` $<$ `NON_COMPLIANT`. The two remaining classifications carry no strictness rank: an intake stop is upstream of policy and an out-of-scope result implicates nothing, so ranking either against these three would be arbitrary. The implementing predicate raises rather than ranking them.

14.2 The decision rule

Proposition 2 (The intake gate precedes policy). If an action has an intake defect in the sense of Definition 12 at review date a, then for every registry — including the empty one — the result is `INSUFFICIENT_INFORMATION`, the implicated-line tuple is empty, and the reason codes include `intake_blocked`. No line is consulted, so no registry content can change the outcome. An action is a proposed action in the sense of Definition 10; the intake gate inspects its context and evidence, not its description.

Proposition 3 (The intake gate is conjunctive over all nine dimensions). Take a baseline action the live registry classifies `COMPLIANT` with a fresh `VERIFIED` record for each of the nine dimensions. Degrading exactly one record — removing it, or setting its status to self-asserted, unverified, or contradicted, or dating it outside the freshness window — withdraws that result. Across the nine dimensions and the five degradations, all forty-five executed evaluations return `INSUFFICIENT_INFORMATION`, and each of the forty-five names exactly the degraded dimension and no other as blocking. No dimension is decorative, and the stop signal identifies the field it stopped on.

Proposition 4 (One verdict, most severe first). Given a defect-free intake, the evaluator reduces over the lines that cover the effective scope and returns exactly one classification, determined in this order: if any covering line has no satisfied exemption in the sense of Definition 8, `NON_COMPLIANT`; else if any satisfied exemption sits below its line's tier floor of Definition 13 or coincides with two or more coverage hits on the same line, `REQUIRES_MODIFICATION`; else if any line was covered, `COMPLIANT`; else `OUTSIDE_SCOPE`. The four cases are exhaustive and mutually exclusive, and a less severe outcome never displaces a more severe one. A line is a red line in the sense of Definition 9; its coverage scope, tier floor, and typed exemptions are the fields this rule reads.

Proposition 5 (All five classifications are reachable). Each of the five classifications of Definition 11 is returned by at least one action against the live registry at a fixed review date. Reachability is a property of the control flow, not evidence that any real engagement was classified correctly; the same harness reports partial coverage honestly when run against an empty registry, where the three implication-dependent outcomes become unreachable.

Proposition 6 (No exemption narrows a line without verified evidence). An exemption narrows its line only when it both matches the effective scope and is satisfied by the context. On the live registry every one of the sixteen typed exemptions requires at least two evidence kinds, so no matching declaration alone can clear a line. The detector that would report a zero-evidence exemption returns nothing on the live registry and fires on a planted one, which is what makes the empty result informative rather than merely absent.

Proposition 7 (ALL-mode triggers cannot be reached by one token). For an `any`-mode exemption every single trigger token matches; for an `all`-mode exemption with more than one trigger token, no single token matches and only the full trigger set does. The live registry carries thirteen `any`-mode and three `all`-mode exemptions, and every `all`-mode exemption has exactly two trigger tokens. Run through the real evaluator against an anchor from the exemption's own line, each `all`-mode exemption's line is `NON_COMPLIANT` when one trigger token is declared and `COMPLIANT` when both are — fifty-eight executed evaluations, every row behaving as its mode requires.

Proposition 8 (Reducing oversight never softens a verdict). Fix a declared scope and a fully evidenced context, and vary only the tier along `hosted` → `connected` → `air_gapped`. Strictness in the sense of Definition 14 never decreases. Every scope keyword of every current line at all three tiers is one hundred and eight executed evaluations with zero inversions. An unexempted covering line is `NON_COMPLIANT` at every tier; a satisfied exemption below its floor is `REQUIRES_MODIFICATION` rather than `COMPLIANT`.

Proposition 9 (Normalization is closed, and failure stops at two layers). $\mathcal{N}$ is idempotent: $\mathcal{N}(\mathcal{N}(S)) = \mathcal{N}(S)$ for every scope it accepts. A token that is not ASCII after NFKC normalization is rejected rather than canonicalized, and the rejection is enforced twice. The action constructor raises, so an ordinary caller cannot build the action at all. Should such a token be written into the frozen record past the constructor, the evaluator normalizes defensively a second time and returns `INSUFFICIENT_INFORM ATION` with an `invalid_scope` reason code and an empty normalized scope. A homoglyph or full-width spelling therefore cannot enter the vocabulary as a new token, and cannot pass as an unmatched one either.

Proposition 10 (Reason codes are a stable, duplicate-free audit surface). Every assessment carries reason codes drawn from a closed vocabulary, appended in evaluation order and never repeated; the assessment record rejects a duplicated code at construction. Human-readable reasons can be reworded without changing the codes, which is what lets a downstream consumer regression-test the *route* to a verdict rather than the terminal alone.

14.3 The report envelope

A classification word is the *last* step of a review, not the whole of it. `COMPLIANT` is a projection of a derivation that also holds the reasons trail, the matched exemption if any, the evidence sweep, and the authorization arm — and a word that travels without its derivation is exactly the safe-looking projection this instrument must not let harden into the state. The envelope is the transport contract that keeps the two attached: the word travels only alongside a digest pointer to the complete native finding, and the instrument's non-claims travel inside the same record.

Definition 15 (Report envelope). The report envelope is the frozen record v = (schema_version, line_id, subject_id, review_date, r with exactly those ten fields, in order, exported under the schema string `line.report-envelope/1.0`. native_status is this line's own classification word from Definition 11; it is one instrument's word in that instrument's vocabulary, never to be compared, ranked, averaged, or merged across lines. report_ref is the SHA-256 of the canonical native finding (`red-line.report/1.0`), which serializes the complete derivation — including the authorization arm as an explicit `null` when it is absent, so an unreviewed arm is distinguishable from an empty one. Sibling instruments export the same shape by publishing the same schema string, never by importing one another.

Proposition 11 (The envelope points, never reinterprets). For every finding f the evaluator returns, `finding_envelop e(f)` produces the Definition 15 and satisfies `envelope_matches_finding(envelope, f)`: the digest pointer, the review date, the registry digest, and the classification word all agree with the finding they were exported from, and editing any checked field afterwards makes the check return false. The envelope adds nothing the finding does not determine except the caller-supplied subject and snapshot references, which are stored, not verified. A matching envelope attests that an archived pair is unedited; it does not certify the finding true, the action safe, or the review well aimed.

14.4 What binds each proposition

Every row names a test in `tests/integration/test_formalism_bindings.py` that re-derives the proposition's content from the code and then asserts the manuscript states it. Corrupting the sentence reddens the row; corrupting the code reddens the derivation inside it.

Proposition	Claim in one line	Verifying test
Proposition 1	staleness uses strict comparison; the window edge is exclusive	`test_staleness_exclusive_boundary_ is_derived_from_the_code`
Proposition 2	a defect stops the evaluation before any line is read	`test_intake_precedence_holds_again st_every_registry`
Proposition 3	all nine dimensions are load-bearing and the stop is localized	`test_evidence_conjunction_matches_ the_executed_sweep`

Proposition	Claim in one line	Verifying test
Proposition 4	one verdict, resolved most-severe-first	`test_outcome_precedence_is_exhaustive_and_ordered`
Proposition 5	all five classifications are reached	`test_outcome_reachability_matches_the_executed_battery`
Proposition 6	a matching trigger alone never narrows a line	`test_exemption_evidence_floor_is_derived_from_the_registry`
Proposition 7	ALL-mode needs every trigger token	`test_trigger_mode_counts_match_the_executed_probe`
Proposition 8	dropping oversight never softens a verdict	`test_tier_monotonicity_numbers_match_the_executed_sweep`
Proposition 9	normalization is idempotent and fails closed	`test_normalization_closure_is_executed_not_asserted`
Proposition 10	codes are closed, ordered, and duplicate-free	`test_reason_code_vocabulary_is_closed_and_duplicate_free`
Proposition 11	the envelope agrees with its finding, and any edit is visible	`test_envelope_pointer_agreement_is_executed_not_asserted`

The two propositions that are hardest to see from the code alone have their own plates. fig. 14 renders the forty-five perturbations behind Proposition 3, and fig. 15 renders the fifty-eight probes behind Proposition 7.

14.5 What the formalism does not establish

A proposition here says what the program returns for declared inputs. It does not say that the declaration is honest, that the verified record is true, that the nine dimensions exhaust what matters, or that the registry names the right boundaries. Proposition 8 is a consistency property of a decision procedure, not a claim that a hosted deployment is safe. Proposition 3 shows that the gate is strict, which is a different thing from showing that strictness is well aimed: an action can satisfy all nine dimensions with plausible false records and reach `COMPLIANT`, and the formalism has nothing to say about that case. Those limits are developed in sec. 17.

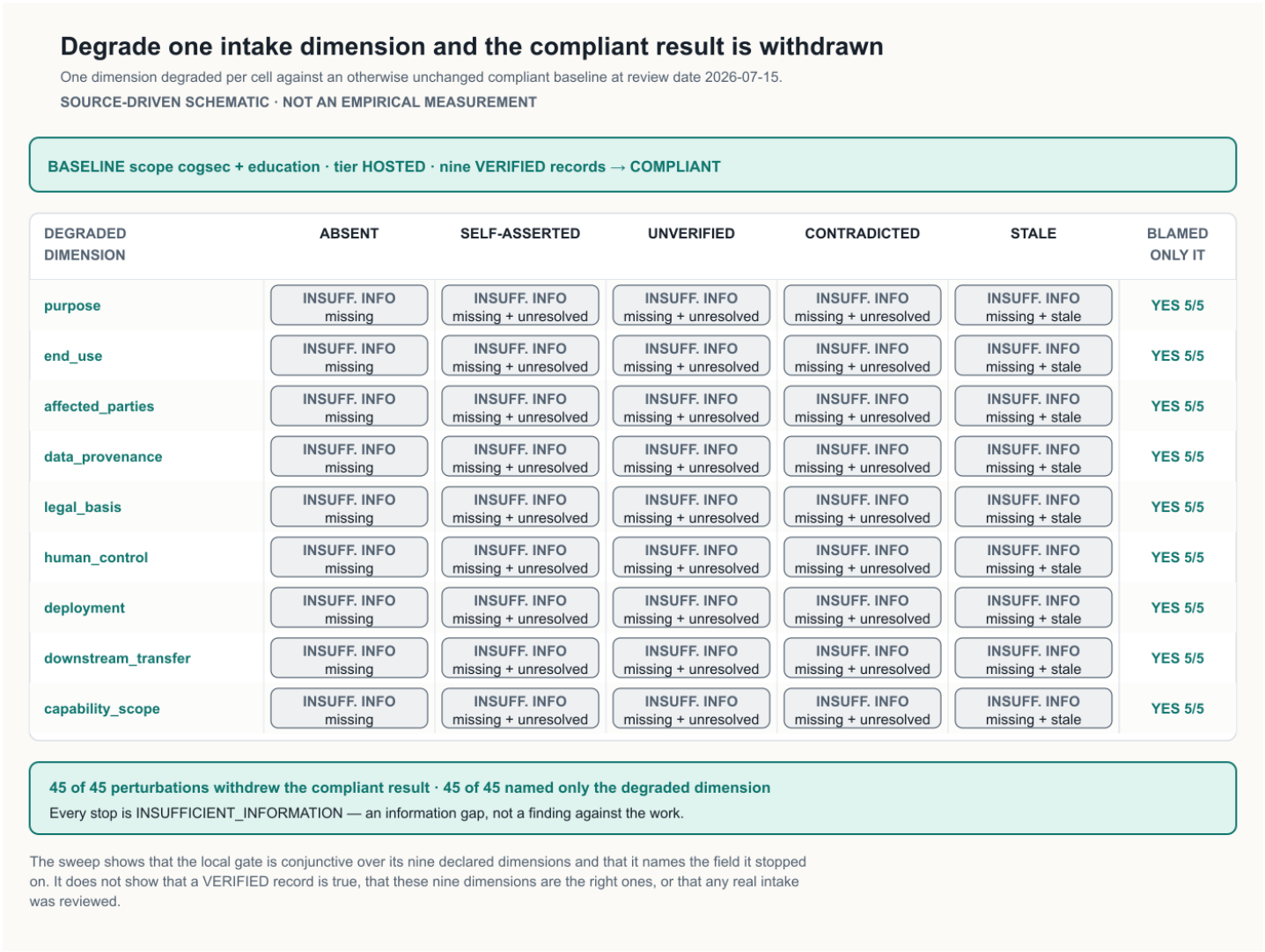


Figure 14: Degrade one intake dimension and the compliant result is withdrawn. Single-dimension perturbation sweep computed by `red_line.analysis.evidence_sensitivity.run_evidence_sensitivity`: a compliant baseline is re-run through the real `evaluate_action` with exactly one of its nine verified records removed, downgraded, or aged past the freshness window. Every cell reports the classification actually returned and the reason codes it raised, and the trailing column confirms the gate named only the degraded dimension. Conjunctive behaviour is a property of the local gate — not evidence that a verified record is true, that these are the right nine dimensions, or that any real intake was reviewed.

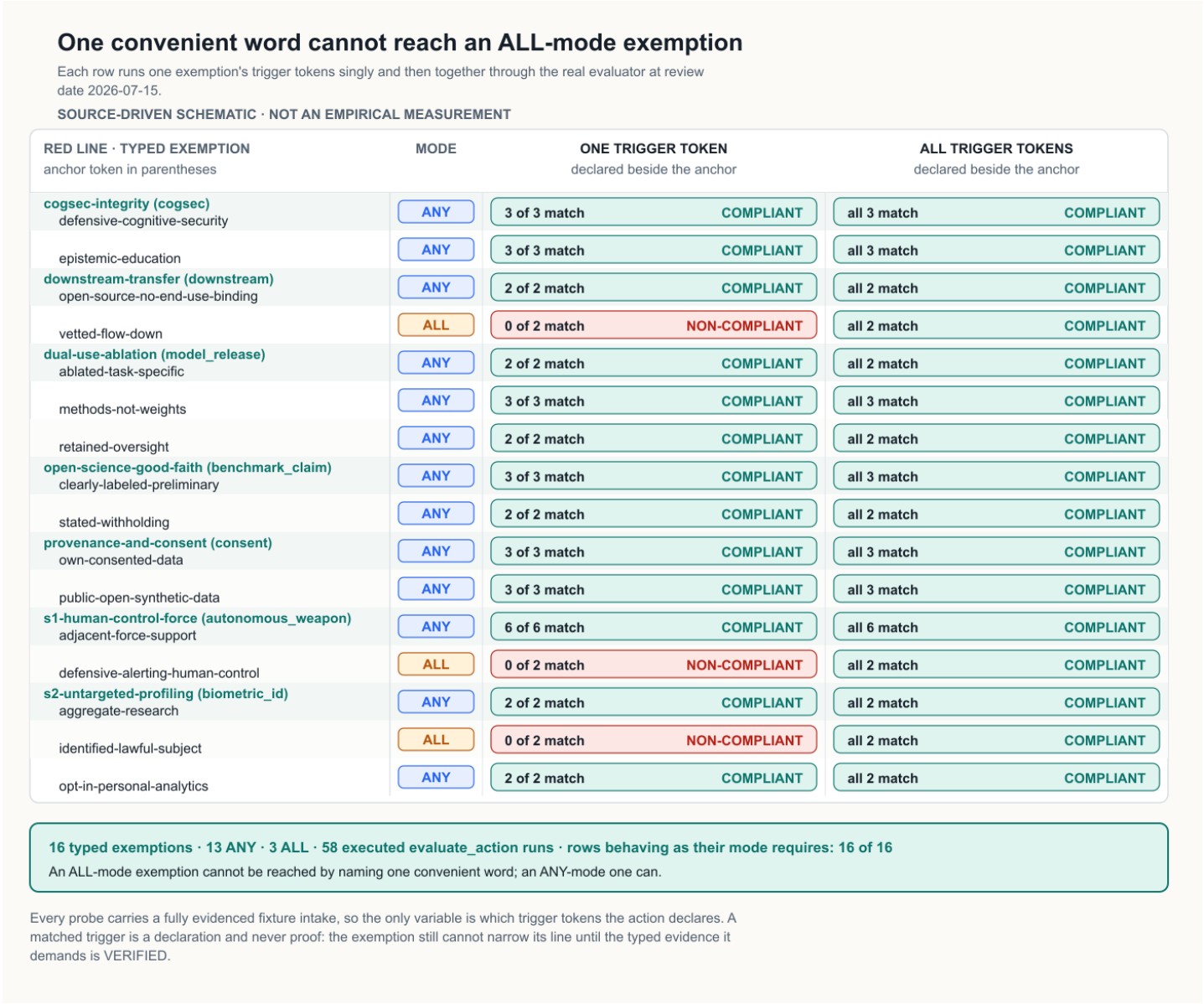


Figure 15: One convenient word cannot reach an ALL-mode exemption. Trigger-semantics probe computed by `red_line.analysis.trigger_semantics.run_trigger_semantics`: every typed exemption is run through the real `evaluate_action` twice, once declaring a single trigger token beside an anchor from its own line's coverage scope and once declaring the whole trigger set. ANY-mode rows match on every single token; ALL-mode rows match on none and clear their line only when every trigger token is present. A matched trigger is a declaration and never proof — the typed evidence must still be verified, and the plate reports match semantics rather than whether a declaration describes the work honestly.

15 The red-line registry

This is the beacon: the author's personal security boundary and explicit No document at version `0.3.0`. The registry is first-person, dated, revisable, and non-exhaustive. It is not a universal ethics code and not a claim made by any AI system that assisted with its preparation.

The current registry contains seven lines, including two CANARY-grade lines (formalized as Definition 9 and Definition 8). The canonical registry digest is:

`72835fd81d1f7ecf70f47b1e0061cd56c385273dd846879ab639225913f5aad7`

Each line has a human-readable standard, rationale, coverage dimensions, narrowing clauses, structured exemptions, required evidence, a tier floor, severity, author, and date. The narrative clauses explain the boundary to a reader; the typed exemptions are the only executable narrowing conditions. Adding a token such as `vetted` or `consented` cannot establish vetting or consent.

The authoritative machine-readable source is `src/red_line/registry/lines.py`. The canary includes exemption semantics in its canonical payload, so changing an exemption or its evidence requirement is a detectable registry change and must be accompanied by a successor rationale.

Because the registry is enumerable data, its narrowing structure can be stated as derived numbers rather than described qualitatively. The analysis module `red_line.analysis.registry_metrics` computes the full exemption × evidence-kind coverage matrix from the live registry: the seven lines carry 16 typed exemptions that together declare 37 evidence requirements across the nine intake dimensions. Affected parties is the most-demanded dimension (six exemptions require it), followed by purpose, legal basis, and capability scope (five each); end use and deployment are the least demanded (two each). Thirteen exemptions match their trigger scope with `ANY` semantics and three require `ALL` trigger tokens to be declared. fig. 16 renders that matrix directly from the analysis code; every filled cell is a precondition for narrowing, not proof that the required evidence is true. The fuller structural profile — severity and tier-floor distributions, scope overlap points, and per-line composition — is derived in sec. 16.

15.1 s1-human-control-force — Human control over force and harm-capable systems

[CANARY]

Standard: I will not build, tune, or knowingly supply a component to a system that selects and engages targets for force without appropriate human control over each engagement.

Rationale: Turner Standard 1: force-application without an identifiable, accountable human decision-maker removes the moral circuit-breaker. Applies whether I provide targeting directly or as a component.

Coverage dimensions: `targeting`, `weapons`, `lethality`, `force`, `kinetic`, `autonomous_weapon`

Does not restrict: defensive-only alerting with a human in the loop

Does not restrict: logistics, translation, maintenance, or research and development

Does not restrict: intelligence analysis reviewed by an accountable human

Typed exemptions and required evidence:

- `defensive-alerting-human-control`: Defensive-only alerting with an accountable human decision-maker; `human_control`, `end_use`.
- `adjacent-force-support`: Logistics, translation, maintenance, research, or intelligence analysis; `purpose`, `human_control`.

Max tier: hosted
Severity: CANARY
Stated by: Daniel Ari Friedman
Stated on: 2026-07-15

15.2 s2-untargeted-profiling — No untargeted profiling or mass surveillance

[CANARY]

Standard: I will not build tooling whose purpose is to convert bulk data into individualized intelligence on persons not already identified as subjects of a specific, lawful, individualized process.

Rationale: Turner Standard 2: bulk-to-individual inference on unnamed persons is the engine of mass surveillance. Demographic-, origin-, or belief-based initiation is prohibited outright.

Coverage dimensions: `surveillance`, `profiling`, `bulk_data`, `biometric_id`, `dragnet`, `tracking`

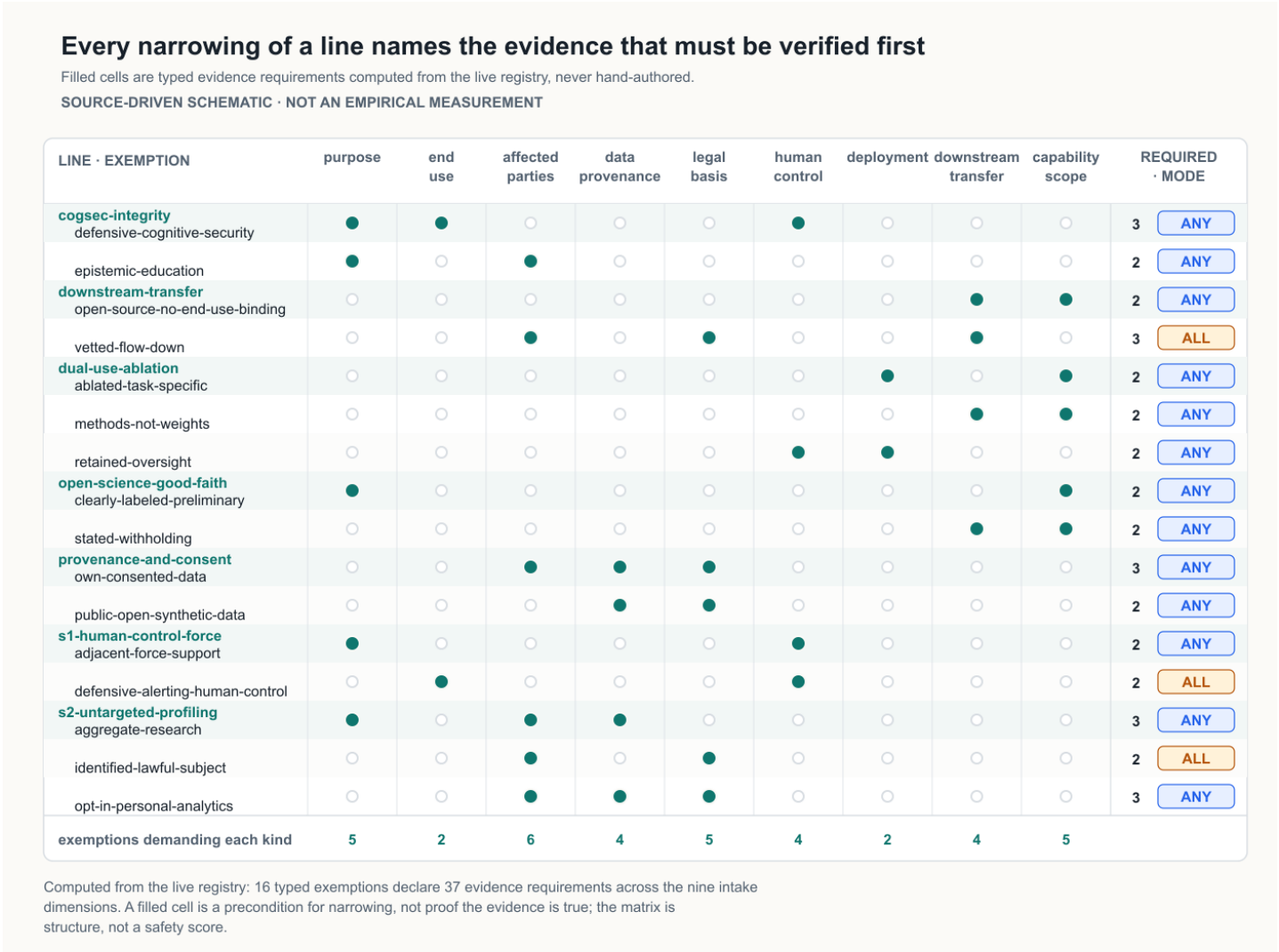


Figure 16: Every narrowing of a line names the evidence that must be verified first. Registry-derived exemption evidence-requirement matrix computed by `red_line.analysis.registry_metrics.exemption_evidence_matrix`: each row is one typed exemption grouped under its red line, each column one of the nine intake dimensions, and a filled cell means the exemption can narrow its line only when a `VERIFIED` record of that kind is present. The bottom row counts how many exemptions demand each dimension. The matrix describes the structural shape of the author's boundaries, not their moral weight; a satisfied requirement is a locally recorded condition, not independent truth or a safety score.

Does not restrict: aggregate research producing no individualized output

Does not restrict: analysis of an already-identified, lawfully specified subject

Does not restrict: consented, opt-in personal analytics

Typed exemptions and required evidence:

- `aggregate-research`: Aggregate research producing no individualized output; `purpose`, `affected_parties`, `data_provenance`.
- `identified-lawful-subject`: Analysis of an already-identified, lawfully specified subject; `affected_parties`, `legal_basis`.
- `opt-in-personal-analytics`: Consent-based opt-in personal analytics; `affected_parties`, `data_provenance`, `legal_basis`.

Max tier: connected
Severity: CANARY
Stated by: Daniel Ari Friedman
Stated on: 2026-07-15

15.3 dual-use-ablation — Scoped release of dual-use models

Standard: I will not release a proprietary or handed-off model beyond my recall (air-gapped) while it retains dangerous dual-use capability that has not been ablated below a repurposing-cost threshold.

Rationale: Turner Tier 3: for work released beyond monitoring, the cost of repurposing a scoped model should exceed the value of doing so.

Coverage dimensions: `model_release`, `weights`, `handoff`

Does not restrict: release of task-specific models with capability removed

Does not restrict: open publication of methods, papers, or benchmarks

Does not restrict: hosted or connected tiers under retained oversight

Typed exemptions and required evidence:

- `ablated-task-specific`: Task-specific release with dangerous capability removed;
  required evidence: `capability_scope`, `deployment`.
- `methods-not-weights`: Open methods, papers, or benchmark publication without dangerous weights;
  required evidence: `capability_scope`, `downstream_transfer`.
- `retained-oversight`: Hosted or connected work under retained oversight;
  required evidence: `deployment`, `human_control`.

Max tier: air-gapped
Severity: STRONG
Stated by: Daniel Ari Friedman
Stated on: 2026-07-15

15.4 cogsec-integrity — Cognitive security strengthens, never degrades, the epistemic commons

Standard: I will not build cognitive-security tooling whose function is to manufacture deception, run covert influence operations, or degrade a population's shared ability to reason.

Rationale: My cognitive-security work is defensive by definition: it strengthens information ecosystems. Weaponized persuasion inverts that mission.

Coverage dimensions: `influence_ops`, `disinformation`, `manipulation`, `propaganda`, `deception`, `cogsec`

Does not restrict: detection, red-teaming, or defensive analysis of influence operations

Does not restrict: education, media-literacy, or transparency tooling

Typed exemptions and required evidence:

- `defensive-cognitive-security`: Detection, red-teaming, or defensive analysis; `purpose`, `end_use`, `human_control`.
- `epistemic-education`: Education, media literacy, or transparency tooling; `purpose`, `affected_parties`.

Max tier: connected
Severity: ABSOLUTE
Stated by: Daniel Ari Friedman
Stated on: 2026-07-15

15.5 provenance-and-consent — Provenance and consent for data and identity

Standard: I will not train, evaluate, or ship on data acquired without a lawful basis and, where persons are involved, without consent or a legitimate public-interest basis.

Rationale: Turner's 'acquisition' clause: any process by which person data enters my systems is covered regardless of how a source labels it.

Coverage dimensions: `data_acquisition`, `scraping`, `pii`, `consent`, `provenance`

Does not restrict: public-domain, openly-licensed, or synthetic data

Does not restrict: my own or explicitly-consented personal data

Typed exemptions and required evidence:

- `public-open-synthetic-data`: Public-domain, openly licensed, or synthetic data; `data_provenance`, `legal_basis`.
- `own-consented-data`: Author-owned or explicitly consented personal data; `data_provenance`, `legal_basis`, `affected_parties`.

Max tier: hosted
Severity: STRONG
Stated by: Daniel Ari Friedman
Stated on: 2026-07-15

15.6 open-science-good-faith — Open-science claims are honest and reproducible

Standard: I will not publish a result, metric, or capability claim I have not verified against a reproducible artifact, nor let proprietary interest launder an unverified claim into the public record.

Rationale: My public research graph is a beacon only if its claims bind to code that runs. Overclaiming corrodes the commons I am trying to build.

Coverage dimensions: `publication`, `benchmark_claim`, `metric`, `reproducibility`, `provenance`

Does not restrict: clearly-labeled preliminary, hypothesis, or negative results

Does not restrict: withholding proprietary detail while stating what is withheld

Typed exemptions and required evidence:

- `clearly-labeled-preliminary`: Clearly labeled preliminary, hypothesis, or negative result;
  required evidence: `purpose`, `capability_scope`.
- `stated-withholding`: Withheld proprietary detail explicitly identified;
  required evidence: `capability_scope`, `downstream_transfer`.

Max tier: air-gapped
Severity: STRONG
Stated by: Daniel Ari Friedman
Stated on: 2026-07-15

15.7 downstream-transfer — No knowing transfer to a violating end use

Standard: I will not sell, license, or hand off work where I know or have substantial reason to believe it will be re-transferred to an end use that violates any of these red lines.

Rationale: Turner's downstream-transfer and 'knows or has substantial reason to believe' clauses: flow-down of the substantive boundary, not just the box.

Coverage dimensions: `resale`, `sublicense`, `integration`, `downstream`, `handoff`

Does not restrict: transfer to a vetted end user with a flow-down commitment

Does not restrict: fully open-source release under a license that itself binds no end use

Typed exemptions and required evidence:

- `vetted-flow-down`: Transfer to a vetted end user with a flow-down commitment;
  required evidence: `downstream_transfer`, `legal_basis`, `affected_parties`.
- `open-source-no-end-use-binding`: Open-source release whose license does not bind end use;
  required evidence: `downstream_transfer`, `capability_scope`.

Max tier: connected
Severity: STRONG
Stated by: Daniel Ari Friedman
Stated on: 2026-07-15

The executable records, rather than this prose alone, determine whether an exemption is satisfied. A complete intake can still be `NON_COMPLIANT`, and a complete action with no matching line is `OUTSIDE_SCOPE`, not a safety finding.

16 Registry composition as derived data

The registry chapter above states what each line refuses; this section states what the registry *is*, structurally, using only numbers computed from the live registry object. Prose descriptions of a versioned artifact drift; a count computed at read time from `src/red_line/registry/lines.py` cannot. The analysis subpackage `red_line.analysis.registry_metrics` exists for exactly this purpose: every function in it is a pure, zero-I/O summarizer over the same in-memory `RedLine` tuple the evaluator consumes, with deterministic sorted output and fail-closed input validation. None of its outputs is a safety score. A count describes the shape of a boundary — how many tokens it covers, how many conditions narrow it, what evidence those conditions demand — not the strength, correctness, or moral standing of the person committing to it.

16.1 Severity and tier floors

The seven lines divide by severity into two `CANARY` lines, one `ABSOLUTE` line, and four `STRONG` lines (`severity_distribution`). Their deployment-tier floors divide into two lines whose floor is `hosted`, three at `connected`, and two at `air_gapped` (`tier_floor_distribution`). The two gradings are deliberately orthogonal: severity records how much process a change to the line itself requires, while the tier floor records the most permissive deployment tier at which the line's exemptions can operate at all. The registry exhibits that orthogonality directly — the two `CANARY` lines sit at different tier floors (`hosted` and `connected`), and the two `air_gapped` floors belong to `STRONG` lines, not to the most change-protected ones.

16.2 Scope vocabulary and overlap points

The seven lines declare 36 scope-token slots over 34 distinct canonical tokens (`scope_token_frequency`). Exactly two tokens are shared between lines, each by two lines: `handoff` (declared by both `dual-use-ablation` and `downstream-transfer`) and `provenance` (declared by both `provenance-and-consent` and `open-science-good-faith`). These are the registry's only structural overlap points: an action whose declared scope includes one of them implicates two boundaries in a single evaluation, and the severity-monotone reduction described in sec. 13 then resolves the pair to the single most severe applicable classification. The near-total disjointness of the remaining 32 tokens is a design consequence of the standalone lines, not an accident — each line names its own coverage rather than inheriting a shared taxonomy.

fig. 17 draws that vocabulary in full — every token against every line — so the overlap is visible as a property of the whole grid rather than as two names in a sentence. The figure also reads out what the evaluator actually returns for each shared token, which is the part that matters operationally: `handoff` is declared by `dual-use-ablation`, whose `retained-oversight` exemption is satisfied by the sweep's fully evidenced hosted intake, and by `downstream-transfer`, whose exemptions are not — and the returned verdict is `NON_COMPLIANT`. One line's verified exemption does not clear a token the other line also claims.

The exemption named there is the one the evaluator reports applying, not the one a reader might expect from the line's carve-out prose: a scope of exactly `handoff` triggers no methods-publication exemption, because that exemption's trigger tokens are `methods`, `paper`, and `benchmark`. The distinction is the whole point of typed triggers, so the id in the sentence above is re-derived from the assessment's own reason strings by `tests/integration/test_manuscript_composition_binding.py` rather than read off the carve-out list.

16.3 Per-line structure

The table below is computed by `red_line.analysis.registry_metrics.line_summaries` from the live registry. Scope is the count of canonical coverage tokens; carve-outs are narrative clauses; exemptions are the typed, executable narrowing conditions, split by trigger match mode.

Line	Severity	Tier floor	Scope	Carve-outs	Exemptions	ANY / ALL
`cogsec-integrity`	absolute	connected	6	2	2	2 / 0
`downstream-transfer`	strong	connected	5	2	2	1 / 1
`dual-use-ablation`	strong	air_gapped	3	3	3	3 / 0
`open-science-good-faith`	strong	air_gapped	5	2	2	2 / 0
`provenance-and-consent`	strong	hosted	5	2	2	2 / 0
`s1-human-control-force`	canary	hosted	6	3	2	1 / 1

Line	Severity	Tier floor	Scope	Carve-outs	Exemptions	ANY / ALL
`s2-untargeted-profiling`	canary	connected	6	3	3	2 / 1

The ranges are narrow by construction: scope sizes run from three to six tokens, every line carries two or three narrative carve-outs, and every line carries two or three typed exemptions. No line is an outlier that concentrates most of the registry's narrowing surface, and no line is a bare prohibition with no stated exemption path. The three `ALL`-mode exemptions sit on `s1-human-control-force`, `s2-untargeted-profiling`, and `downstream-transfer` — one each — where the exemption's trigger matches only when *all* of its tokens are declared rather than any single one.

16.4 Evidence depth and the free-pass check

The 16 typed exemptions distribute their 37 evidence requirements narrowly: eleven exemptions require exactly two evidence kinds and five require exactly three (`exemption_evidence_matrix`; the minimum across all rows is two, the maximum three). Trigger scopes are similarly small — nine exemptions trigger on two tokens, six on three, and one on six.

Trigger scope and evidence are separate gates, and the difference is worth seeing rather than reading: Proposition 7 and fig. 15 probe every exemption with one trigger token and then with all of them, and the three `ALL`-mode rows stay blocked until every token is declared.

The floor of two is the registry's most important structural property, so it is checked rather than assumed. `unevidenced_exemptions` returns every matrix row whose exemption requires *no* evidence kind at all — an exemption that any matching declaration would satisfy, which is to say a free pass through its line. On the current registry the function returns an empty tuple. That empty result is meaningful only because the detector is proven able to fire: the test suite constructs a planted registry containing a deliberate zero-evidence exemption and asserts both that `unevidenced_exemptions` reports it and that the registry invariant suite fails on the same input. Absence of a finding from a detector that has demonstrated detection is evidence about the current registry version; absence of a finding alone would be no evidence at all.

fig. 18 puts the two preceding sections on one plate: the per-line table as four bars on a shared scale, the severity and tier-floor splits as a footer strip, and the free-pass count as a band that is drawn whether or not it is zero. Rendering the zero matters. A panel that appeared only when a free pass existed would make today's clean registry indistinguishable from a figure set that had quietly stopped checking; a band reading `EXEMPTIONS REQUIRING NO EVIDENCE AT ALL: 0` is a result. The regression test for this figure injects one synthetic evidence-free exemption into a copy of the registry and asserts the band changes, so the zero is falsifiable in the plate as well as in the analysis module.

These numbers travel with the version. All of them describe the registry state pinned by the canonical digest in sec. 15; a future amendment that adds a line, widens a scope, or relaxes an evidence requirement changes the derived numbers, the canary payload, and this section's claims together. That coupling is intentional: composition claims that are recomputed from source cannot silently outlive the registry state they describe.

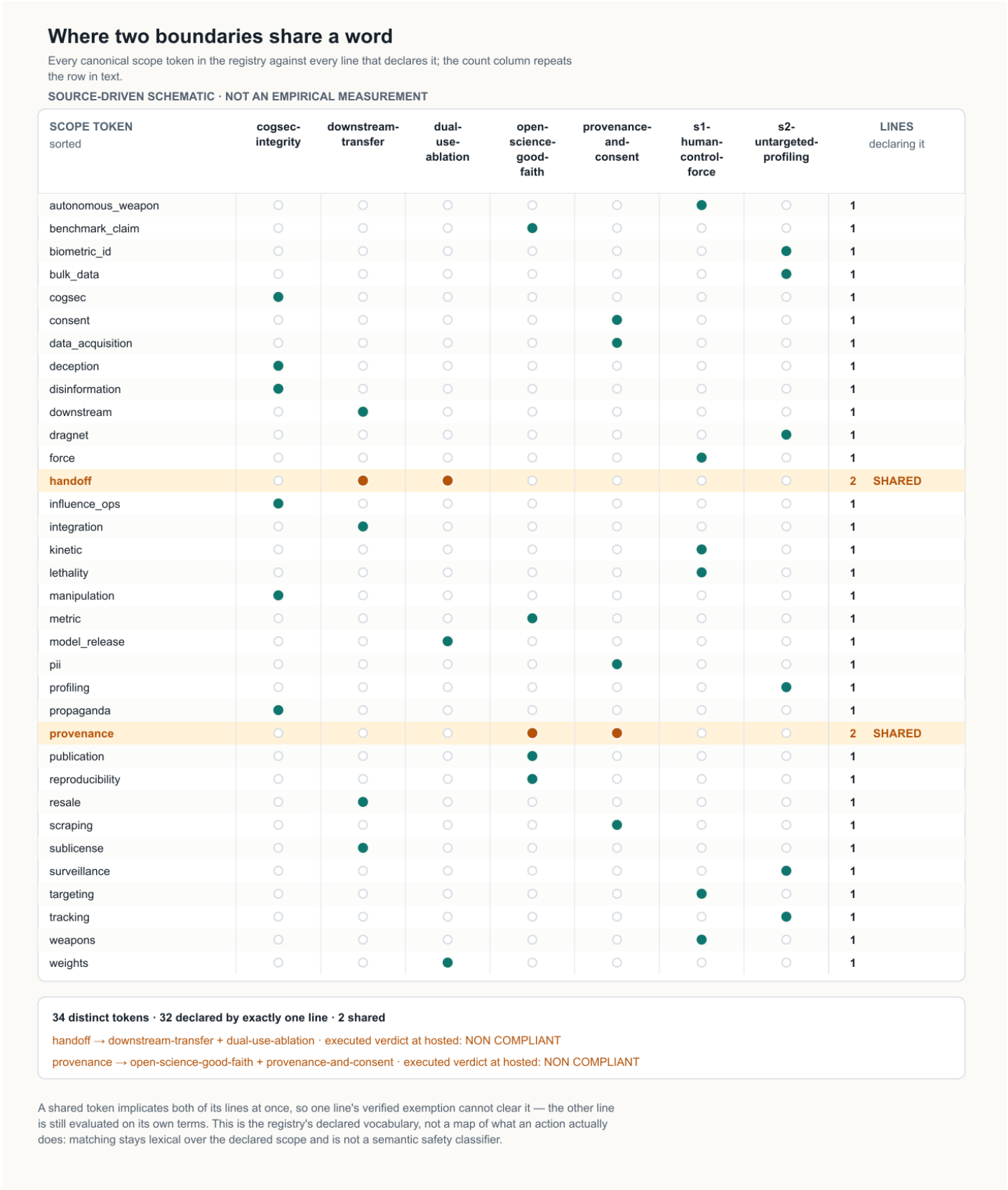


Figure 17: Where two boundaries share a word. Presence grid computed by `red_line.analysis.registry_metrics.scope_token_membership`: every one of the 34 distinct canonical scope tokens against each of the seven lines, with a filled mark where the line declares the token and a count column repeating each row in text. Two tokens — `handoff` and `provenance` — are declared by two lines; the other 32 belong to one line each. The footer prints the verdict the real evaluator returned for each shared token during the tier-monotonicity sweep, so the consequence is executed rather than asserted. Implementation fact: this is the registry's declared vocabulary. Interpretation, stated as a limit: matching remains lexical over a declared scope and is not a semantic classifier, so the grid shows which words can implicate two boundaries, never whether an action truly does.

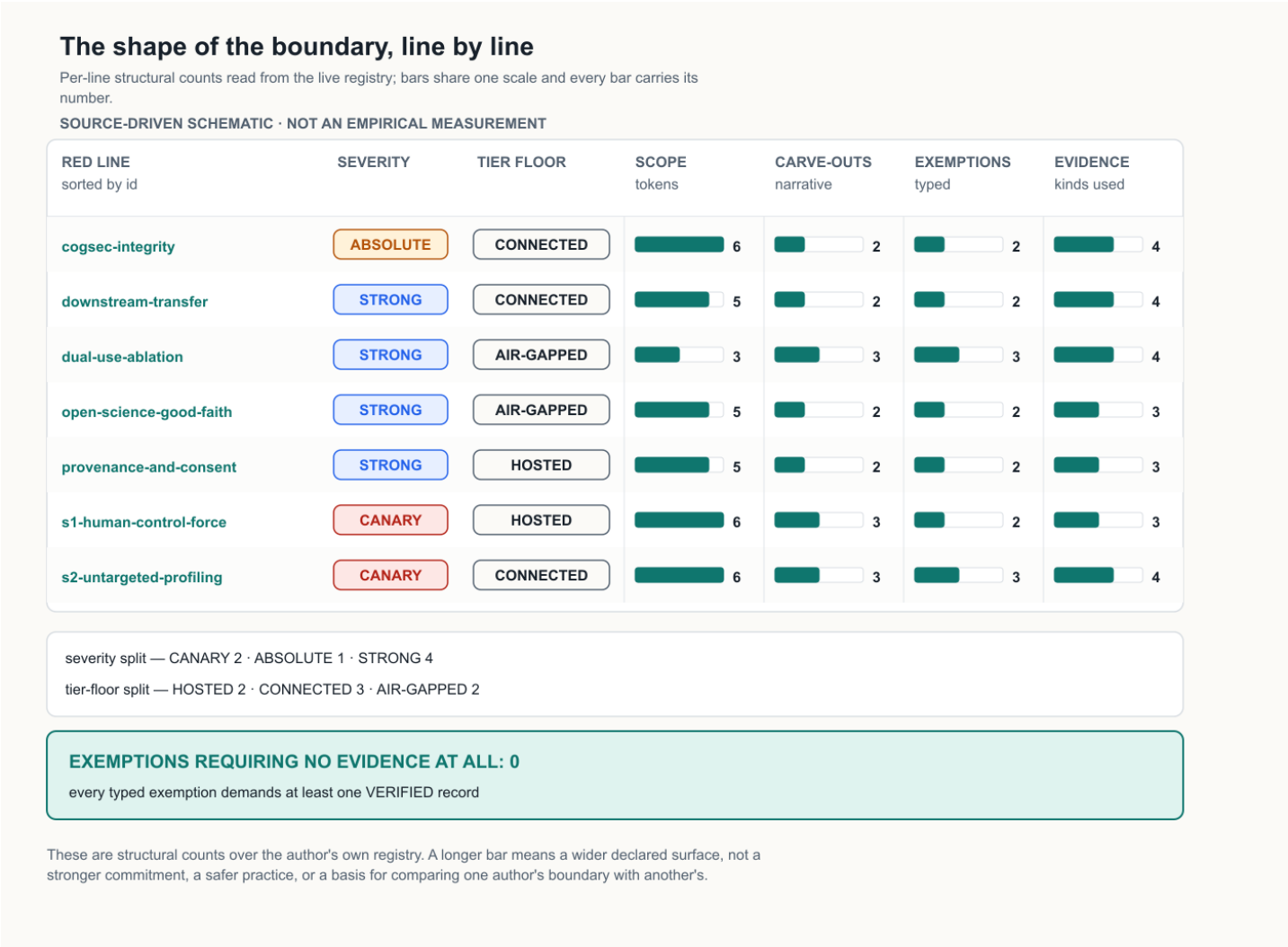


Figure 18: The shape of the boundary, line by line. Per-line structural profile computed by `red_line.analysis.registry_metrics.line_summaries`, `severity_distribution`, `tier_floor_distribution`, and `unevidenced_exemptions`. Each row is one of the seven lines in id order with its severity grade and oversight floor as text-labelled chips, and four counts on one shared bar scale: declared scope tokens, narrative carve-outs, typed exemptions, and distinct evidence kinds used. Every bar carries its number, so the plate reads without colour. The bottom band renders the count of typed exemptions requiring no evidence at all, currently zero. Implementation fact: these are counts over registry fields. Interpretation, stated as a limit: a longer bar is a wider declared surface, not a stronger commitment, a safer practice, or a basis for comparing this author's boundary with anyone else's.

17 Limitations and negative space

17.1 What this document does not decide

Red Line is deliberately narrow. It does not decide:

- **Legal adjudication:** whether an act is lawful, licensed, contractually authorized, or subject to a particular jurisdiction's standard.
- **Truth verification:** whether a cited source, artifact, witness, or evidence record is true, complete, current, or free of strategic misrepresentation.
- **Real-world safety:** whether a system will behave safely, whether harm will occur, or whether a documented control will work under operational pressure.
- **Institutional authorization:** whether an employer, customer, regulator, community, or affected person has granted authority to proceed.
- **Independent external witnessing:** whether anyone outside the author's write boundary has received, checked, or attested to the registry or finding.

The instrument can require escalation when these questions cannot be established, but it cannot answer them by itself.

Auditability, not enforcement. The package records a local result. It does not stop execution, remove access, adjudicate legality, or create an independent review body. A named authorization is visible but never releases a block.

Evidence is not truth. `VERIFIED` means a reviewable artifact or source was identified. The package does not establish that the artifact is accurate, complete, current, lawfully obtained, or free of strategic misrepresentation. Contradicted or unsupported context blocks the result, but a plausible false record can still pass the local status gate.

Lexical, not semantic. Explicit scope aliases are safer than heuristic stemming, but the evaluator still matches declared tokens. A complete-looking action can be mislabeled. Description mismatch hints expose one symptom; they do not solve semantic interpretation.

False positives and false negatives. The fail-closed intake is intentionally biased toward stopping when required information is absent or unresolved. That can produce a false positive for caution: a legitimate exemption may wait until its evidence is assembled. The lexical boundary also permits a more serious residual false negative: a dangerous action can be described with a benign token set or a complete-looking but misleading evidence packet. The instrument reduces these risks through explicit scope, negative tests, source provenance, and independent review; it cannot remove them without semantic and institutional authority it does not have.

17.2 Adversarial declarations

Because every input is self-declared, the instrument can be gamed by construction, and the honest response is to demonstrate the attacks rather than deny them. Each of the following was executed against the real evaluator.

Scope-token laundering. A targeting-component action declares only `logistics`, `translation`, and `maintenance` — permissible support tokens that match the `adjacent-force-support` exemption on `s1-human-control-force` — while omitting `targeting`, `weapons`, or `autonomous_weapon`. The evaluator matches declared tokens against scope and exemption triggers; the action's narrowed scope implicates only `s1-human-control-force`, and the verified exemption returns `COMPLIANT`. The evaluator cannot inspect whether the declared tokens honestly describe a component that selects and engages targets.

Evidence-status fabrication. An action provides VERIFIED records with plausible but fabricated references — a purpose statement citing a non-existent contract, a human-control record pointing to a private repository that has never been reviewed. The evidence gate requires VERIFIED status for every dimension the exemption demands; a record whose status field reads `VERIFIED` satisfies the gate regardless of whether the referenced artifact exists, is accurate, or was genuinely reviewed. The evaluator records epistemic status, not ground truth.

Refresh-date laundering. The freshness window requires evidence dated within 180 days of the review date. An action whose evidence records were last genuinely verified 300 days ago rewrites their dates to the current review date and returns `COMPLIANT`. The evaluator checks the declared date against the window; it cannot distinguish a genuine re-verification — a reviewer re-examining the source — from an edit to a date string.

Declared-unknown gamesmanship. An action declares every sensitive dimension as `not_applicable` and supplies a single VERIFIED record backing that designation, aiming to minimize the evidence surface while staying within the gate. The evaluator still demands VERIFIED evidence for each dimension, and `not_applicable` is a permitted value only when a verified record supports it — but the evaluator cannot assess whether the dimension is genuinely inapplicable or merely declared so to shrink the review burden. An action whose sole remaining required dimension carries a fabricated record passes on the strength of a single false positive.

These are instances of a well-documented dynamic, not defects unique to this design. A refusal instrument that certified safety would make these failure modes catastrophic, because a gamed `COMPLIANT` would launder a dangerous action into apparent permission. Red Line's design response is to refuse the certifying role entirely: a classification reports what the declarer provided and the evaluator computed, so a gamed `COMPLIANT` overstates nothing but the presence of declared evidence. The attacks also stay inspectable rather than hidden — the declared scope, the evidence status records, and the declared dates are the very record a reviewer reads, so a reviewer who asks "do these tokens describe this work?", "was this evidence genuinely reviewed?", or "what changed at this refresh?" is asking questions the declaration itself exposes. The instrument narrows what gaming can counterfeit; it cannot remove the need for the human judgment those questions require, and it never converts any classification into a safety certification, an accreditation, or a permission.

Registry scope is non-exhaustive. `OUTSIDE_SCOPE` is not "safe" and `COMPLIANT` is not "universally acceptable." Both are bounded by seven personal lines, their current vocabulary, and the quality of the intake.

Personal authority remains personal. The lines are the author's refusals, not a claim that other people must share them. The global and historical reading base widens the questions and exposes blind spots; it does not turn the author into a representative of the traditions cited.

Scholarship is curated, not representative. The reading base is not a systematic review or a substitute for affected-party testimony, local expertise, or jurisdiction-specific legal analysis. Its purpose is to widen the questions asked of a personal instrument and to make transfer limits visible.

A formal statement is not a stronger claim. The definitions and propositions in sec. 14 say what the program returns for declared inputs. Writing that down precisely, and binding each proposition to a test, removes one failure mode — prose drifting away from the procedure — and adds no authority. A proposition can be true of the code and useless in the world at the same time, which is the case Proposition 3 describes, and the same is true of Proposition 8 and Proposition 4: the gate is strict, and strictness is not accuracy.

Structural analytics describe shape, not strength. The derived numbers in sec. 16 and the exercised outcome coverage in sec. 13 are registry introspection, not measurement of the world. An exemption that demands three evidence kinds is not thereby "stronger" than one that demands two, and the demand profile across intake dimensions is not a risk model; comparing such counts across authors as if they scored rigor would recreate exactly the safety-score misuse this document disclaims. The free-pass detector reports on structurally degenerate exemptions only — a substantively wrong boundary with well-formed structure passes every metric. Its empty result is bounded to the current registry state, and the outcome-coverage report is a control-flow reachability pin built on fixture evidence: a `complete` report shows the evaluator can produce all five classifications, not that any real engagement was classified correctly.

Canary dependence. A same-author repository fixture is a regression anchor, not an external witness. The hash detects drift only when a prior copy is held outside the writer's control and checked by someone else. The canary is a tamper-evidence pattern, not legal protection or prevention.

Publication pipeline. Deterministic figures, output validation, and rendered inspection reduce source-to-artifact risk, but they do not prove that a polished figure is rhetorically fair. PDF and HTML review remain required.

Release state. This work is released as a standalone repository at https://github.com/docxology/red_line. It has no external canary witness, no minted DOI, no independent review body, and no execution admission control. A public release should not claim those things until the release evidence bundle in `docs/claim-register.md` exists and an independent reader has checked the source, outputs, and prior statement.

18 Conclusion

Red Line is deliberately a refusal instrument. It makes seven personal No's legible before a request becomes a project, then requires an evidence-bearing context before a proposed near-boundary action can receive a policy result. That strictness changes the meaning of the tool's green states: `COMPLIANT` means that this documented intake satisfied this registry, while `OUTSIDE_SCOPE` means only that no current line applied. Neither is a general safety certificate.

The four propositions stated in the introduction hold.

In use, the sequence is equally simple: read the beacon; declare the action, scope, and tier; assemble the nine evidence dimensions; run the local result; retain the finding; and maintain a prior canary outside the author's write boundary. The order is a safeguard against sunk-cost reasoning, scope laundering, and self-certification. It is not a substitute for legal review, affected-party participation, technical safety work, or institutional authority.

What the instrument actually does is now written down twice, in prose and in sec. 14, with a test standing behind every proposition. That duplication is the point. A document about a decision procedure is worth reading only if a reader can find out where it has stopped being true, and the second statement is where they can.

The mechanism source is sec. 3; the personal adaptation stands on its own. Its registry, evidence model, typed exemptions, review findings, non-bypassable authorizations, canary, scholarship ledger, figures, and tests are this project's claims. The reading base does not dissolve the first-person nature of the commitments; it makes the author more accountable for the blind spots around them, and the export-control and refusal literatures locate what the instrument is not — neither a control regime with force nor a refusal with standing. The claim register names the evidence and stopping point for each kind of assertion.

The companion works keep the line set non-redundant. Black Line asks how strong work is done; Golden Line asks what is worth reaching toward; White Line asks what is absent, unknowable, withheld, or ethically left unsaid. Red Line comes first because a positive method or aspiration is not a substitute for a clear boundary. The artifact is ready for continued private use and revision; it is not yet externally attested publication.

The work is built for the author's public research index (`docxology/docxology`) — machine-readable, cross-linked, and verification-logged — with an eventual public home at the `docxology/red_line` repository. That is the boundary kept in plain sight: the refusal is designed to be read before a request becomes a project.

References

African Union Commission. African union data policy framework. African Union, 2022. URL https://au.int/fr/node/42078. Accessed 2026-07-17.

Aristotle. *Politics*. -350. URL https://classics.mit.edu/Aristotle/politics.html. Ancient Greek political text; accessed 2026-07-17.

Abeba Birhane. Algorithmic colonization of africa. *SCRIPTed: A Journal of Law, Technology and Society*, 17(2):389–409, 2020. doi: 10.2966/scrip.170220.389. URL https://doi.org/10.2966/scrip.170220.389.

Simone Browne. *Dark Matters: On the Surveillance of Blackness*. Duke University Press, 2015. doi: 10.1215/9780822375302. URL https://read.dukeupress.edu/books/book/147/Dark-MattersOn-the-Surveillance-of-Blackness. Accessed 2026-07-17.

Miles Brundage, Shahar Avin, Jack Clark, Helen Toner, Peter Eckersley, Ben Garfinkel, Allan Dafoe, Paul Scharre, Thomas Zeitzoff, Bobby Filar, Hyrum Anderson, Heather Roff, Gregory C. Allen, Jacob Steinhardt, Carrick Flynn, Seán Ó hÉigeartaigh, S. J. Beard, Haydn Belfield, Clare Lyle, Rebecca Crootof, Owain Evans, Michael Page, Joanna Bryson, Roman Yampolskiy, and Dario Amodei. The malicious use of artificial intelligence: Forecasting, prevention, and mitigation. arXiv preprint arXiv:1802.07228, 2018. URL https://arxiv.org/abs/1802.07228. Accessed 2026-07-17. Research report/preprint; not a forecast of this project's threat likelihood.

Sasha Costanza-Chock. *Design Justice: Community-Led Practices to Build the Worlds We Need*. The MIT Press, 2020. ISBN 9780262043458. URL https://mitpress.mit.edu/9780262043458/design-justice. Accessed 2026-07-18. Used for community-led design, power, and participation; not a safety certification or consent shortcut.

Nick Couldry and Ulises A. Mejias. *The Costs of Connection: How Data Is Colonizing Human Life and Appropriating It for Capitalism*. Stanford University Press, 2019. doi: 10.1515/9781503609754. URL https://www.sup.org/books/sociology/costs-connection. Accessed 2026-07-17.

Penny Crofts and Honni van Rijswijk. Negotiating 'evil': Google, project maven and the corporate form. *Law, Technology and Humans*, 2(1):75–90, 2020. doi: 10.5204/lthj.v2i1.1313. URL https://doi.org/10.5204/lthj.v2i1.1313. Accessed 2026-07-28. Used as the documented case of collective worker refusal of military AI work and its corporate absorption; not evidence about the outcome of any individual refusal.

Ottobah Cugoano. *Thoughts and Sentiments on the Evil and Wicked Traffic of the Slavery and Commerce of the Human Species*. London, 1787. URL https://quod.lib.umich.edu/e/eccodemo/K046227.0001.001/1%3A3?rgn=div1&view=fulltext. Primary text; accessed 2026-07-17. Used as a situated Black Atlantic source on liberty, consent, responsibility, and moral self-deception.

Adam Dahl. Ottobah cugoano. Stanford Encyclopedia of Philosophy, 2025. URL https://plato.stanford.edu/entries/cugoano/. Scholarly overview accessed 2026-07-17; used to situate interpretive debates rather than to certify a universal African philosophy.

Linda T. Darling. Social cohesion (asabiyya) and justice in the late medieval middle east. *Comparative Studies in Society and History*, 49(2):329–357, 2007. doi: 10.1017/S0010417507000515. URL https://www.cambridge.org/core/journals/comparative-studies-in-society-and-history/article/abs/social-cohesion-asabiyya-and-justice-in-the-late-medieval-middle-east/3117D292647ECD6472E9C77AA294D2A2.

Catherine D'Ignazio and Lauren F. Klein. *Data Feminism*. The MIT Press, 2020. ISBN 9780262044004. URL https://mitpress.mit.edu/9780262044004/data-feminism/. Accessed 2026-07-18. Used for power-aware classification, invisible labor, and the limits of data speaking for themselves; not an exhaustive ethics standard.

Virginia Eubanks. *Automating Inequality: How High-Tech Tools Profile, Police, and Punish the Poor*. St. Martin's Press, 2018. ISBN 9781250074317. URL https://us.macmillan.com/books/9781250074317/automatinginequality. Accessed 2026-07-17.

Miranda Fricker. *Epistemic Injustice: Power and the Ethics of Knowing*. Oxford University Press, 2007. ISBN 9780198237907. URL https://academic.oup.com/book/32817. Accessed 2026-07-17.

Mary L. Gray and Siddharth Suri. *Ghost Work: How to Stop Silicon Valley from Building a New Global Underclass*. Houghton Mifflin Harcourt, 2019. ISBN 9781328566249. URL https://marylgray.org/bio/on-demand/. Accessed 2026-07-17. Used for hidden labor and human judgment at the boundary of automated systems.

Donna Haraway. Situated knowledges: The science question in feminism and the privilege of partial perspective. *Feminist Studies*, 14(3):575–599, 1988. doi: 10.2307/3178066. URL https://www.jstor.org/stable/3178066. Accessed 2026-07-18. Used for situated perspective and accountable partiality, not relativism or a substitute for evidence.

Elisa D. Harris, editor. *Governance of Dual-Use Technologies: Theory and Practice*. American Academy of Arts and Sciences, Cambridge, MA, 2016. URL https://www.amacad.org/publication/governance-dual-use-technologies-theory-and-practice. Accessed 2026-07-28. Comparative study of nuclear, biological, and cyber dual-use governance; used for the layered structure of control regimes, not as an assessment of this project.

Sheila Jasanoff. Technologies of humility: Citizen participation in governing science. *Minerva*, 41(3):223–244, 2003. doi: 10.1023/A:1025557512320. URL https://doi.org/10.1023/A:1025557512320. Accessed 2026-07-18. Used for framing, vulnerability, distribution, and learning under uncertainty; not a local decision procedure.

Anna Jobin, Marcello Ienca, and Effy Vayena. The global landscape of AI ethics guidelines. *Nature Machine Intelligence*, 1(9):389–399, 2019. doi: 10.1038/s42256-019-0088-2. URL https://doi.org/10.1038/s42256-019-0088-2.

Kautilya. *King, Governance, and Law in Ancient India: Kautilya's Arthasastra*. Oxford University Press, 2013. doi: 10.1093/acprof:osobl/9780199891825.001.0001. URL https://academic.oup.com/book/8486. Annotated translation; accessed 2026-07-17.

Tahu Kukutai and John Taylor, editors. *Indigenous Data Sovereignty: Toward an Agenda*. Centre for Aboriginal Economic Policy Research. ANU Press, 2016. doi: 10.22459/CAEPR38.11.2016. URL https://press.anu.edu.au/publications/series/caepr/indigenous-data-sovereignty. Accessed 2026-07-18. Used for collective data authority and Indigenous governance of data, not as a universal consent shortcut or substitute for community authority.

Bartolomé de Las Casas. *A Short Account of the Destruction of the Indies*. Penguin Classics, 1552. ISBN 9780140445626. URL https://www.penguin.co.uk/books/35187/a-short-account-of-the-destruction-of-the-indies-by-bartolome-de-las-casas-ed-and-trans-by-nigel-griffin-intro-anthony-pagden/9780140445626. Spanish primary polemic composed 1542 and published 1552; English edition accessed 2026-07-17. Used as a situated colonial-era source, not Indigenous testimony or a universal authority.

Niccolo Machiavelli. *The Prince*. 1513. URL https://www.gutenberg.org/ebooks/1232. Primary text; accessed 2026-07-17.

Muhsin Mahdi. FARABI vi. political philosophy. Encyclopaedia Iranica, 2000. URL https://www.iranicaonline.org/articles/farabi-vi/. Accessed 2026-07-17.

MITRE. MITRE ATT&CK enterprise techniques. MITRE ATT&CK knowledge base, 2026. URL https://attack.mitre.org/techniques/. Accessed 2026-07-17. Threat-model vocabulary, not attribution or telemetry.

Shakir Mohamed, Marie-Therese Png, and William Isaac. Decolonial AI: Decolonial theory as sociotechnical foresight in artificial intelligence. *Philosophy and Technology*, 33(4):659–684, 2020. doi: 10.1007/s13347-020-00405-8. URL https://doi.org/10.1007/s13347-020-00405-8.

National Research Council. Biotechnology research in an age of terrorism. Technical report, The National Academies Press, Washington, DC, 2004. URL https://nap.nationalacademies.org/catalog/10827/biotechnology-research-in-an-age-of-terrorism. Known as the Fink report, after committee chair Gerald R. Fink. Accessed 2026-07-28. Used for the proposition that researcher-level judgment is a governance layer, not for its biosecurity subject matter.

Helen Nissenbaum. Privacy as contextual integrity. *Washington Law Review*, 79(1):119–158, 2004. URL https://digitalcommons.law.uw.edu/wlr/vol79/iss1/10/.

Safiya Umoja Noble. *Algorithms of Oppression: How Search Engines Reinforce Racism*. New York University Press, 2018. ISBN 9781479837243. URL https://nyupress.org/9781479837243/algorithms-of-oppression/. Accessed 2026-07-17.

OECD. Oecd principles on artificial intelligence. Organisation for Economic Co-operation and Development, 2019. URL https://www.oecd.org/en/topics/ai-principles.html. Updated page accessed 2026-07-17; principles adopted 2019.

Elinor Ostrom. *Governing the Commons: The Evolution of Institutions for Collective Action*. Cambridge University Press, 1990. doi: 10.1017/CBO9780511807763. URL https://www.cambridge.org/core/books/governing-the-commons/7AB7AE11BADA84409C34815CC288CD79.

Scott Rose, Oliver Borchert, Stu Mitchell, and Sean Connelly. Zero trust architecture. Technical Report NIST SP 800-207, National Institute of Standards and Technology, 2020. URL https://csrc.nist.gov/pubs/sp/800/207/final.

James C. Scott. *Seeing Like a State: How Certain Schemes to Improve the Human Condition Have Failed*. Yale University Press, 1998. URL https://yalebooks.yale.edu/book/9780300246759/seeing-like-a-state/. Accessed 2026-07-17.

Andrew D. Selbst, Danah Boyd, Sorelle A. Friedler, Suresh Venkatasubramanian, and Janet Vertesi. Fairness and abstraction in sociotechnical systems. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, pages 59–68. ACM, 2019. doi: 10.1145/3287560.3287598. URL https://doi.org/10.1145/3287560.3287598.

Audra Simpson. *Mohawk Interruptus: Political Life Across the Borders of Settler States*. Duke University Press, Durham, NC, 2014. ISBN 9780822356554. URL https://www.dukeupress.edu/mohawk-interruptus. Accessed 2026-07-28. Source of the concept of refusal as a positive political stance rather than a deficit; read in its Kahnawà:ke and settler-colonial context and explicitly not transferred as a template for a practitioner declining paid work.

SLSA Community. Slsa specification version 1.2. Linux Foundation community specification, 2026. URL https://slsa.dev/spec/v1.2/. Approved specification accessed 2026-07-17; implementation context, not a project attestation.

Linda Tuhiwai Smith. *Decolonizing Methodologies: Research and Indigenous Peoples*. Zed Books, 1999. URL https://www.royalsociety.org.nz/150th-anniversary/tetakarangi/decolonizing-methodologieslinda-tuhiwai-smith-1999. Accessed 2026-07-17.

Murugiah Souppaya, Karen Scarfone, and Donna Dodson. Secure software development framework (SSDF) version 1.1. Technical Report NIST SP 800-218, National Institute of Standards and Technology, 2022. URL https://csrc.nist.gov/pubs/sp/800/218/final. Implementation context; accessed 2026-07-17; not evidence of Red Line compliance.

Sunzi. The art of war. Classical Chinese military text; English translation consulted through the Chinese Text Project, 1910. URL https://ctext.org/art-of-war/laying-plans/ens. Accessed 2026-07-17. Used as a situated source on information, deception, and strategic judgment, not as an ethical endorsement.

Elham Tabassi. Artificial intelligence risk management framework (AI RMF 1.0). Technical Report NIST AI 100-1, National Institute of Standards and Technology, 2023. URL https://www.nist.gov/itl/ai-risk-management-framework.

The Wassenaar Arrangement on Export Controls for Conventional Arms and Dual-Use Goods and Technologies. List of dual-use goods and technologies and munitions list. Multilateral export-control regime; control lists as amended by the December 2025 Plenary, 2025. URL https://www.wassenaar.org/control-lists/. Accessed 2026-07-28. Used as the reference case for a capability-keyed institutional control list, including the definitional over-breadth of the 2013 intrusion-software entry; not a legal authority over this project and not a claim that any listed category applies to it.

Alex Turner. A red line and oversight framework for government AI contracts. The Pond, July 2026. URL https://turntrout.com/red-line-framework. Accessed 2026-07-17. Source framework adapted by this project from organization-to-government scale to a single practitioner governing personal development work.

UNESCO. Recommendation on the ethics of artificial intelligence. United Nations Educational, Scientific and Cultural Organization, 2021. URL https://www.unesco.org/en/legal-affairs/recommendation-ethics-artificial-intelligence?hub=1063. Adopted 23 November 2021; accessed 2026-07-17.

Mary Wollstonecraft. *A Vindication of the Rights of Woman*. 1792. URL https://www.gutenberg.org/ebooks/3420. Primary text; accessed 2026-07-17.

Jonathan Zong and J. Nathan Matias. Data refusal from below: A framework for understanding, evaluating, and envisioning refusal as design. *ACM Journal on Responsible Computing*, 1(1):1–23, 2024. doi: 10.1145/3630107. URL https://doi.org/10.1145/3630107. Accessed 2026-07-28. Used for the four facets of refusal — autonomy, time, power, cost — written from the standpoint of those who refuse; not a claim that a practitioner's refusal is refusal from below.