AI in Safeguards Work: An Interim Protocol

Version 0.1 · Drafted July 2026 · sociable.systems

THE RETIREMENT CLAUSE

This protocol exists because no equivalent public guidance does, and it is written to be retired.

When a safeguards-setting institution publishes operational guidance covering this ground, this protocol will be reviewed against it and retired, revised, or narrowed wherever that guidance supersedes its function. (Operational guidance here means procedures a practitioner can follow and a reviewer can check, applying to E&S deliverables generally rather than to one tool or workflow.) Until that day, this is the discipline, and its version history becomes a record of what practitioners had to improvise in the meantime.

Why an interim discipline, and why now

The safeguards world runs on documents. Impact assessments, resettlement plans, stakeholder engagement records, grievance logs, monitoring reports: the entire architecture of accountability for large projects is, in the end, a chain of documents that someone downstream is expected to trust. AI has now entered the rooms where those documents get made, and it has entered quietly.

Let us be precise about the state of things, because imprecision is how this goes wrong. AI drafting of full ESIAs is not standard practice. It is piloted, partial, uneven, and often undisclosed. Screening tools summarize baseline data. Models draft sections that humans revise. Translation, synthesis and search run through systems teams may not consistently log. The honest description of mid-2026 is a transition caught halfway: too far in for “we don’t use AI” to be true, not far enough for anyone to have built the verification habits the new tools require.

The institutions have noticed, and it matters to be precise about what already exists, because this protocol builds on it rather than pretending an empty field. IAIA’s SP16, “Principles for Use of AI in IA” (https://iaia.org/wp-content/uploads/2025/06/SP16_AI-in-IA.pdf) (February 2025) gives the profession eight adopted principles, including genuinely strong disclosure language (tool, date, manner of use, down to consultation data), and then says plainly of itself that it is not a guidance document on how to apply them. The World Bank is past principles and into production: its Tech-for-ESF initiative, presented at IAIA25 (https://iaia.org/iaia25-conference) (deck (https://2025.iaia.org/presentations/1370_SONG_AI-Powered_Environmental_and_Social.pdf)), has an AI assistant drafting sections of environmental and social review summaries, piloted with more than thirty-five specialists ahead of a scheduled June 2025 launch, alongside geospatial tooling that screens project risks across 190-plus public data layers and internal datasets. The only governance language visible in that public deck is a single sentence about outputs being reviewed and verified. And an IAIA panel in May 2026 (https://iaia.org/from-data-to-decisions-how-ai-is-reshaping-impact-assessment/), with a MIGA seat at the table, was billed in exactly the operational register: AI-enabled processes that are repeatable and auditable.

Meanwhile the cost of the missing discipline stopped being hypothetical. In 2025 a Big Four firm repaid the final instalment of a AU$440,000 Australian government contract (https://fortune.com/2025/10/07/deloitte-ai-australia-government-report-hallucinations-technology-290000-refund/) after its report was found to contain invented citations, including a quote attributed to a Federal Court judgment that did not exist. In April 2026 South Africa withdrew its draft national AI policy (https://www.dlapiper.com/en-us/insights/publications/2026/05/withdrawal-of-south-africa-draft-ai-policy) after at least six of its sixty-seven bibliography sources proved unverifiable or fabricated; the minister's stated most-plausible explanation was AI-generated citations included without verification. In June 2026 another Big Four firm withdrew its own report on agentic-AI excellence (https://www.theregister.com/ai-and-ml/2026/06/12/kpmgs-ai-report-turns-into-a-demo-of-ai-hallucinations/5255029) after an external audit (GPTZero's, widely reported) could verify five of its forty-five citations, while named clients disputed its claims in public. Behind the headline cases sits a quieter curve: the share of research papers carrying at least one fabricated reference (https://www.statnews.com/2026/05/07/lancet-study-finds-steep-rise-fraudulent-citations-academic-papers/) ran roughly one in 2,828 in 2023, one in 458 in 2025, and one in 277 in early 2026. Generative AI is not the whole story behind that curve (paper mills and older kinds of misconduct contribute), which makes the checking problem larger, not smaller.

Every one of those failures happened to organizations with more compliance staff than most safeguards teams will ever have. The lesson is structural rather than moral: in each case, the verification that should have caught the fabrications was absent, ineffective, or insufficiently independent of the work it was checking. The safeguards field is walking into the same weather with thinner coats. A resettlement plan with an invented precedent in it does not embarrass a consultancy; it reshapes what a displaced family is owed.

So the precise state of the field, mid-2026: the profession has principles that decline to be procedures, the lenders are deploying tools ahead of rules, and we found no publicly available operational procedures from any safeguards-setting institution governing how AI use in E&S deliverables must be disclosed, evidenced, bounded, and independently reviewed. (The search behind that sentence is documented in this protocol's prior-art register, and we would genuinely like to be corrected.) The nearest thing to a worked example comes from outside the field entirely: the evidence-synthesis community's joint position statement (https://onlinelibrary.wiley.com/doi/full/10.1002/cl2.70074) (Cochrane and its sister organizations, 2025) already requires that any AI use which makes or suggests judgments be fully and transparently reported. Safeguards work has no public equivalent. This protocol is that missing operational layer, written from the practitioner's side, assuming SP16 where SP16 speaks.

> So this protocol makes a narrow bet. The window between AI entering safeguards work and AI-in-safeguards guidance arriving is exactly the window in which discipline is cheap. Adopt the habits now, while they are voluntary, and the eventual guidance becomes a formality you already comply with. Wait, and the first hostile reader to check your chain of citations will write the guidance for you, in a complaint, at a price you do not set.

What SP16 says, and how this protocol stands on it

SP16 is two pages and eight principles, and it has earned a fair summary rather than a wave. **Responsibility**: humans carry full accountability for AI-assisted work, end to end. **Transparency**: disclose the tool, the date and the manner of use, including where AI generated or analyzed consultation data, with real teeth in the margins (commenters should disclose their own AI use; participants must be told when AI will process their input, and may withdraw). **Integration**: AI may supplement but never replace regulatory or standardized methods unless certified. **Expertise and Oversight**: experts "should conduct

independent verification of the outputs of AI tools." **Awareness of Limitations**: bias, hallucination, sparse non-digitized local knowledge, and ambiguous significance criteria. **Vulnerability and Privacy**: AI must not replace direct communication with affected people, with resettlement and social health data named as especially sensitive ground. **Competence** and **Model Collapse** round out the set.

Three of this protocol's four rules are those principles taken at their word and given procedure. Rule 1 is Transparency operationalized: "disclose the manner of use" becomes a register with a grain. Rule 4 is Expertise and Oversight operationalized: "independent verification" acquires an independence requirement and a procedure, a reviewer outside the authorship of the work running a concrete checklist before the deliverable ships. Rule 3 draws the line that Responsibility and Vulnerability gesture toward: a named list of judgments the machine may inform but never conclude. Rule 2 has no SP16 parent at all. Evidence custody appears nowhere in the principles, and it is the rule every other rule stands on.

Where operational rules of this kind already exist, at other altitudes, they take recognizably this shape. The World Bank's own governance practice, surveying how states govern AI inside public institutions (https://openknowledge.worldbank.org/entities/publication/20a298de-912d-452b-a32d-ab2521429d2f) (2026), lands on the same moves: impact assessment before deployment, documented datasets, independent evaluators, and appeal processes that require "clearer documentation and traceability of inputs." Canada's Directive on Automated Decision-Making (https://www.tbs-sct.canada.ca/pol/doc-eng.aspx?id=32592) has bound federal agencies to that shape since 2019. The safeguards field is simply the domain where no public equivalent yet exists. This protocol looks the way it does because this is what such rules look like wherever someone finally writes them.

The rules

Four rules. Each is stated in one sentence, followed by its teeth. They are deliberately few: a protocol a busy team cannot hold in memory is a template, and a template is where discipline goes to be performed rather than practiced.

> RULE 1
>
> ### Disclosure at touch-point grain
>
> *Every deliverable carries a register of where AI touched it.*
>
> Deliverable-level disclosure ("AI tools were used in the preparation of this report") is the new "we are committed to responsible AI": true in the way that sentence is true, and useless to every reader who matters. Touch-point grain means the register names the section, the nature of the touch (drafting, screening, synthesis, translation, search), the tool used, and the human who reviewed that touch. The lender's reviewer, the inspection panel, the procurement desk and the litigator will all eventually ask the same question: which parts did the model write? A team that can answer from a register looks disciplined. A team that has to reconstruct the answer from memory looks like the June 2026 headlines.
>
> The register is one table. Maintained as the work proceeds it should be lightweight; reconstructed after the fact it is unaffordable. That asymmetry is the whole argument. (A ready-to-use template ships with this protocol's operational pack; any table carrying these columns satisfies the rule.)

RULE 2

Evidence custody at decision-trace grain

Conclusions must be independently traceable from preserved evidence and decision records, by someone who was not there.

The evaluation field's own gold standards require conclusions to be described and explained; they do not consistently require custody at the grain needed to reconstruct the full path from source record to accepted conclusion. That gap predates AI by decades, and AI synthesis inherits it silently: automate the summarizing and the discretion the standard was always carrying becomes invisible, instant, and impossible to appeal. Note what this rule does not promise: hosted AI systems cannot honestly guarantee a literal re-run, because models update silently, sampling varies, and retrieval sources drift. What can be promised, and what this rule requires, is custody of the path. The sources consulted, the prompts used, the model and version, the raw outputs actually relied on, and the human decisions that turned output into conclusion, preserved at a grain where an outsider can reconstruct and audit how the team moved from evidence to accepted claim.

This is the rule teams resist, because it feels like storing exhaust. It is the rule that decides everything downstream. Custody is what turns "trust our process" into "check for yourself," and no other rule in this protocol survives contact with a hostile reader unless this one is being followed.

RULE 3

The never-machine-settled list

Material safeguards judgments are never concluded by automated screening; the machine may inform the human process they exist to trigger, and may never conclude it.

The safeguards standards carry their discretion in specific phrases. "Where feasible." "Commensurate with." Consent that "does not require unanimity." Read for enforcement, each phrase is a gate: a point where the standard hands judgment to a person, in a process, with affected people in the room. When an AI screening tool pre-settles feasibility, or pre-scopes what is commensurate, the gate still gets recorded as passed. The judgment simply happened earlier, in a system, with nobody watching and nobody to appeal to.

A judgment is material where it determines rights, eligibility, impact significance, mitigation adequacy, alternatives, compensation, consent, consultation scope, or the need for further human inquiry. (Mundane drafting conveniences that touch none of these are not the target.) Each adopting team writes its own list at two tiers: a public list of judgment categories, and a project-specific list disclosed to the client and reviewers, since project detail can carry sensitivities the public list should not. The distinctions that will be tested in practice: triage is not determination, a recommendation is not a decision, and human confirmation is not human judgment. A gate passed by rubber stamp is a gate this rule counts as machine-settled, and Rule 4's reviewer checks for exactly that.

RULE 4

The hostile read before submission

No AI-assisted deliverable ships to a lender, regulator or public process without an adversarial pass by a reviewer independent of its authorship.

Friendly review checks whether the document is good. Hostile review checks whether it survives: whether the citations exist, whether the numbers survive recomputation, whether the commitments are enforceable as worded, whether the disclosure register matches the document it describes. The reviewer must not have authored or approved the content under review, must hold the competence to judge it, and must have standing to record unresolved defects without pressure to clear the deliverable. Proportionality applies: where material judgments (Rule 3) or high-stakes deliverables are involved, the reviewer is organizationally independent or external; for routine assistance, a competent internal reviewer outside the authorship chain will do. The year's public failures show what happens when verification is absent, ineffective, or insufficiently independent. The method costs a fraction of the deliverable, and far less than the withdrawal.

(The adversarial method this protocol assumes is public and can be watched running at sociable.systems/stress-test. Any equivalent method satisfies the rule. The rule is the independence, and the check.)

Conditions of use

The four rules assume ground beneath them. Adopting teams also commit, proportionately to risk:

Data boundaries. No personal data, Indigenous knowledge, grievance records, or health and resettlement data enters an AI tool without a named lawful basis, and material classified confidential or sensitive enters only an institutionally approved system under an appropriate processing agreement and access controls. Approval covers purpose limitation, minimization, retention and deletion, cross-border transfer, and whether prompts or outputs may feed vendor training. The Rule 1 register records which tools are approved for which classes of data.

Proportionality. Custody and review burdens scale with consequence; translation polish is not compensation eligibility. Each team writes its tiers down, the project lead records and approves the tier before AI use begins, and Rule 4's reviewer may require a higher one. Where no tier is written, the heaviest applies by default.

Affected-person rights. SP16's notice provisions, operationalized: people are told before participation when AI will or may process their input, what that processing entails, and what choices are available, including withdrawal where applicable. A meaningful non-AI channel for participation and grievance stays open, and using it must not result in reduced access, consideration or remedy.

Error handling. When an AI-linked defect is found after submission: preserve the evidence, notify the client, correct or withdraw the affected content, and record what changed. A quiet fix is its own finding.

What this protocol does not claim

It does not claim AI should be kept out of safeguards work. The tools are in the buildings, and some of what they do (holding institutional memory that outlives staff rotation, making decades of commitments and grievance records findable) is work the field has needed done for decades.

It does not claim these four rules are sufficient. They are the minimum that makes AI-assisted work checkable, which is the precondition for every finer-grained standard to come.

It does not claim to be first. IAIA's SP16 principles precede it and are assumed throughout; where SP16 states the what, this protocol supplies the how, which SP16 expressly declined to write. (Readers reaching for the World Bank's 2021 note on AI data-collection tools (https://documents1.worldbank.org/curated/en/099062923153017633/pdf/P17468208b2baa04d0aa1805f462523e46c.pdf), or IFC's 2020 draft technology code (https://openknowledge.worldbank.org/bitstreams/08bac134-ba44-5a0e-a9a3-5f73b8e73d95/download) for investee companies, are reaching for different artifacts: the first governs building survey tools, the second governs portfolio companies' products. Neither binds the practitioner producing an E&S deliverable.)

And it does not claim authority it lacks. This is a practitioner's document, versioned and dated, standing in for guidance that institutions with actual mandates have yet to write. Its ambition is to be superseded.

Adoption statement

(One page, signable, citable in bids. This statement may be cited alongside IAIA SP16, whose principles it operationalizes.)

[Firm / team name] adopts the Interim Protocol on AI in Safeguards Work, v0.1.

In all environmental and social deliverables we produce:

1. We disclose AI involvement at touch-point grain, in a register accompanying each deliverable.
2. We preserve evidence and decision records at decision-trace grain, such that our conclusions can be independently reconstructed and audited by a party who was not present.
3. We maintain a public list of judgment categories that are never concluded by automated screening, and a project-specific list disclosed to each client.
4. We submit AI-assisted deliverables to an adversarial review independent of their authorship before they ship.
5. We apply the protocol's conditions of use, proportionate to risk.

```
Signed: _____    Date: _____
Register of adopting teams: sociable.systems/protocol
```

Version and changelog

`10 Jul 2026` Prior-art audit applied

`11 Jul 2026` Four-model review panel + second pass applied

Version 0.1, July 2026; revised 11 July 2026 after a four-reviewer adversarial pass (the review record is preserved). Revisions will be numbered and dated; the current version lives at sociable.systems/protocol. Corrections and dissent are

welcome and will be credited: a protocol that cannot survive a hostile read of itself has no business recommending one.