Assuring one delegation

What greenhouse-gas verification can lend to AI agents

Will Brooke

13 November 2026

Working note · version 0.1 · 13 November 2026 · DOI: 10.5281/zenodo.24200916

Every grant of authority to an AI agent should have a record so affected people can ask what authority was given and how it was used. Drawing on greenhouse-gas verification, I propose formal assurance over shared controls for routine jobs and over the individual grant for consequential jobs. Three small tests illustrate how controls can enforce a grant and what evidence a reviewer can use afterward.

In a mock flight-booking study, I gave an AI agent authority to buy a ticket subject to a $650 cap and other conditions. In writing that grant, I tried to put myself in the shoes of someone working to book a flight on someone else's behalf. Before I sent the job out, was the authority mine to give, for each part of the grant? And then afterward, what could someone establish about how the delegated authority was used?

I work with greenhouse-gas (GHG) reporting and verification as a regulator. Some of the practices used to validate projects and verify results in the assurance work I encounter offer structured ways to think about both questions. I propose making **one delegation the unit of record and accountability**. Before the agent acts, check that the grant is valid. Afterward, check how the authority was used. The consequences determine whether formal assurance covers the controls used across grants or the individual grant itself. In both judgments, what counts as material should be informed by what the people authorizing the action owe to those it could affect. How far they need to look depends on what the action could do to others.

Consider what happens when the written record diverges from what the traveller asked for. Hand an outside reader four items from a hypothetical non-refundable booking: the protocol, the intake record showing the traveller's hard 10 a.m. arrival deadline, the grant omitting that deadline, and the booking record for a $640 flight landing at noon. What can the reader conclude about the agent's conduct? The agent met every condition in the grant. What can they conclude about the delegator's decision? The intake record and the grant do not match—a concrete failure in translating the traveller's constraint into the delegation. And who can authorize and pay for a replacement while there is still time?

What this adds

Others have assembled subsets of this problem. On the grant and the authority behind it, Johnny Kao's six-field, task-level delegation contract separates authority from capability and scales specification depth by consequence, reversibility, and delegated authority. Claudiu Hulea adds a necessary distinction: proving that a delegator issued a grant does not prove they held the authority to do so. On evidence and criteria, Yue Zhao and colleagues adapt the preconditions of an ISAE 3000 assurance engagement to define when a claim about an agent is auditable, including when it requires a record from an entity outside the

audited party's control. They explicitly mark their boundary, noting that their verdicts are not assurance opinions and that they leave aside materiality, sampling, and auditor independence. Shea Brown argues that existing assurance standards already fit AI, and that the bottleneck is suitable criteria.

This note takes up the assurance opinion, materiality, and verifier independence that Zhao and colleagues set aside, and pairs them with pre-job validation around one delegator's grant to one agent for one bounded job.

Several CPA firms have begun delivering the joint AICPA–AIUC offering, which combines a SOC 2 examination of system controls with AIUC-1 certification. A full launch of the joint certification is expected in early 2027. At the contained tier, my proposal pairs assurance over shared controls with a record for each grant, a pre-job check on the delegator's capacity to answer for it, and a versioned job-type protocol that supplies the assurance criteria. AICPA–AIUC announcement

The GHG measurement, reporting, and verification (MRV) toolkit informs three design choices for a single grant.

1. **Authority and capacity to answer** during pre-job validation, rather than mere willingness to be accountable.
2. **Claims-based materiality** across the engagement, anchored in what affected parties could legitimately demand, with quantitative thresholds where useful.
3. **Versioned, job-type protocols** (analogous to offset methodologies) to supply the suitable criteria that Brown identifies as the bottleneck.

Related work

The grant and the rules it relies on

For this proposal, one delegation means one delegator's grant to one agent for one bounded job. In the cases considered here, the delegator could be the original principal or someone acting on another's behalf, and a job could include multiple actions. Importantly, the record of the grant needs to capture changes in authority along the way, including waivers and revocation.

The grant sets the terms for a particular job, such as the trip, fare cap, and expiry for a booking. A versioned, job-type protocol sets the rules that apply across grants of that kind. Checks before and after the job use the protocol version that applied to the grant.

Under its Verified Carbon Standard, Verra approves methodologies that set common rules, from soil-carbon measurement and buffer credits under VM0042 to injection records and containment monitoring under VM0049's geological storage module (VMD0058). Independent bodies validate project designs against those rules and verify reported results. Verra's Methodology Development and Review Process, § 3.6 · Verra's validation and verification process · Verra VM0042 · Verra VM0049 · Verra VMD0058

A checklist can record that a check was performed, and an API schema can define which fields a request accepts. A job-type protocol supplies the criteria for deciding whether an action is authorized and what evidence can establish the result. The two job types I tested would call for different criteria.

- **Flight booking.** The protocol would specify how the gate enforces the fare cap, who can waive it, when to recheck the fare, and what safeguards apply to non-refundable tickets. It would define whether commitment occurs when the request is submitted or when the airline accepts it, and require reconciliation of booking records with independent payment or airline settlement records.

- **Weather-data update.** The protocol would make atmospheric pressure the sole writable field and treat temperature as read-only. It would identify acceptable stations or sensors, require a traceable observation with a timestamp, and specify the pressure units and how to verify any conversion to kilopascals.

Routine grants rely on assurance over the shared controls that apply the criteria for that job type. The protocol should say who defines job boundaries, whose claims count, and who adopts the rules. In the noon-arrival case, the traveller needs somewhere to challenge the grant. An internal travel ombudsman, procurement review board, or platform dispute body could compare the grant with the intake record and order a replacement ticket if the institution gives it those powers. Narrow job boundaries can leave foreseeable effects outside a verifier's view. How to give affected people standing remains open.

Before the job, can the delegator answer for it?

Before the job begins, validation answers two questions. Does the grant carry forward the hard constraints from intake and stay within the delegator's authority? Do the safeguards meet the job-specific criteria, given what's at stake?

In *Top Gun* (1986), Stinger tells Maverick, "Son, your ego is writing checks your body can't cash." In an institution, the question is whose account the check is drawn on.

Here I find Stephen Darwall's ideas about answering *to someone* helpful. If I promise to do something for you, it's as though I've given you a receipt for what I now owe you. If I fail to follow through, you can call on me to answer for it. Darwall, "Demystifying Promises"

Kennedy Chun Yin Kong argues that handing work to an AI agent leaves the deployer's answerability undiminished; it is not the deployer's to withhold. When we combine that conservation of answerability with Darwall's view of promissory claims, the pre-job question sharpens: we want to know not only that answerability stays with the delegator, but whether the delegator actually has the standing and capacity to make good on it—who can hold them to it, and whether the responsible entities have the means to intervene or provide a remedy. Kong on undiminished answerability

In the noon-arrival case, the delegator can commit the fare, the airline controls reissuance, and the budget holder controls replacement funds. Before saying, "Relax, if anything goes wrong, I'll take care of it," the delegator needs evidence of timely reissuance under the fare rules or a funded replacement that can meet the deadline.

Who could be affected? Can the action be reversed? What could be repaired or compensated? The answers help determine how much assurance the grant calls for.

Kao scales how much to specify in a delegation contract according to consequence, reversibility, and delegated authority. His tiers run from low consequences and easy reversal through bounded losses and partial reversal to high consequences, difficult reversal, and sensitive or broad authority. I use three corresponding tiers to set how much assurance is required and what the engagement covers. The job-type protocol sets thresholds for each tier, so delegators apply common criteria when classifying a grant.

A **contained** grant covers a refundable booking within budget where a mistake can be corrected in time. The protocol supplies the criteria. The assurance engagement assesses whether the shared controls that implement a specified version of the protocol are designed and operating as intended. Each grant gets its own record and deterministic checks before and after the action. The booking system can confirm the

delegator's spending limit, enforce the fare cap at the gate, and check the booking service's receipt against the terms in force.

A **bounded** grant might cover a non-refundable booking with limited exposure. Before activation, the record for the grant must show how both conditions are met.

1. **Protect the traveller's deadline.** Carry forward the hard constraints from intake and document safeguards that keep schedule risk within the protocol's threshold. Hold the noon-arrival grant until it reflects the 10 a.m. deadline or a change the traveller authorizes. If schedule risk still exceeds the threshold, narrow the grant to options that keep it within bounds or reassess it as consequential.
2. **Address the remaining financial exposure.** Evidence before purchase of reissuance under the fare rules or committed replacement funds supports automated controls with **exception review** if the gate trips or a dispute is filed. The noon-arrival records leave that evidence open. If neither remedy is documented, hold the grant unless the designated budget holder accepts the residual financial exposure in writing within their authority. That route requires **limited assurance** over the record for the grant, including the written acceptance and the schedule safeguards in condition 1.

A grant is **consequential** when its foreseeable consequences are serious enough under the protocol to require reasonable assurance on the individual grant. How readily the action can be reversed, and whether recourse is timely and adequate, inform that judgment. A large financial exposure can qualify even if it could be compensated. A refundable booking can also qualify if a mistake could strand a traveller at a critical moment. This tier calls for an independent human validator and formal opinions on the proposed grant before activation and on its exercise afterward.

Across these tiers, pre-job validation checks whether the available evidence and controls can support the required level of assurance. Missing evidence puts a grant on hold. A narrower grant can reduce the exposure to negative consequences. Some grants will have to be declined as proposed.

Some carbon offset programs make provisions for reversals in advance through buffer pools and rules assigning liability. The corresponding requirement for a delegation is **funded recourse**: establishing before activation who will fund a remedy, what it will cover, and how long it will remain available after the grant expires. Kong rightly objects to collective compensation funds that pay out for harms nobody answers for; here, funded recourse pays for the remedy without displacing who answers. Where a wrong booking can be fixed by a refund or a funded replacement ticket, recourse can cover the exposure. If some harms are beyond what compensation could undo—such as disclosing a confidential file—a validator needs to ask what can prevent the harm before the agent acts. If the available controls are inadequate, that can change whether the grant is acceptable in the first place. ACR's approach to reversal risk

During the job, what can the agent do?

The usefulness of my GHG accounting analogies breaks down during the delegated AI job itself. Here we need a control that is internal to the system, and outside of the model. If an agent is about to take an action of some sufficient seriousness—say, send a file or commit to a purchase—then the system needs some kind of **execution gate** that can stop the action at the moment it is proposed.

In a small test, I authorized an agent to update atmospheric pressure in a synthetic report. Given a format that could also express temperature and advice to update it, the agent proposed changing both. A deterministic checker outside the model rejected the whole update and left the reports unchanged. In a separate read-only test, I removed sensitive files before giving agents a frozen copy of a codebase to

assess. Three of four runs in one set still asked for an excluded file. Those requests were refused and no contents were returned. One test exercised a gate on a proposed action; the other narrowed what could be accessed before work started.

In a hypothetical booking, a purchase request leaves the gate just before revocation and the booking service accepts it just afterward. If the protocol makes submission the point of commitment, the timing supports one finding. If the protocol makes acceptance the point of commitment, the same sequence supports a different finding. The gate has to know which rule to enforce and whether it can still prevent the purchase. A later reviewer needs evidence of both the revocation and the outcome at the booking service.

After the job, what can a verifier establish?

A gate that refuses every purchase trivially satisfies the assertion that "every completed purchase fell within the authority in force at the time of commitment." To prevent safety from collapsing into paralysis, that authorization assertion needs to be paired with a fulfillment assertion: **every eligible request from intake either reached a confirmed outcome or triggered a timely escalation with reasons to a named person while there was still time to intervene.**

For fulfillment, trace eligible requests from intake to a confirmed outcome or a timely escalation to someone with authority to act. For authorization, reconcile the gate log with independent payment or airline settlement records for the grant window to account for all purchases, including any made outside the gate. For each purchase, compare the independently recorded outcome with the terms in force at commitment and any recorded gate decision. Where gaps in the evidence prevent a finding, the claim remains *undecidable*. Zhao et al. on auditable claims

The mock flight study had 28 valid runs, 14 per coding agent, all on synthetic data and with no exposure of any third party. None of the eight preregistered flags fired, and no agent treated an approval planted inside the fare rules as authority. I attributed 17 refusals across 110 tool calls by Claude Code and 33 across 172 by Codex to the test harness. The wrapper enforced rules on record format, labels, reads, and action state. Refused reads accounted for 12 of Claude Code's 17. Those refusals show the operational cost of how I enforced the grant—and illustrate why tracing whether an eligible request stalls or escalates to a human in time to act matters in practice. All three local tests turn on checkable assertions about a field update, access to a file, or compliance with a fare cap. Delegations governed by open-ended qualitative judgment are much harder to gate or verify deterministically.

I also used a blinded model panel to review 17 adjudication packets. The panel was unanimous on 14. In two of the remaining three, the distinction between an incomplete task and an ambiguous rule came to the fore: on one packet, two of the three reviewers called the task incomplete and I agreed with the majority; on another—where one reviewer said compliant, one said non-compliant, and the third reviewer slot was missing—I judged the contract itself ambiguous on the point in dispute. That distinction matters: a finding of non-compliance or incompleteness presumes a requirement clear enough to apply. Because I wrote the contract and made the final adjudication, the study is not independent; a real assurance engagement would require a verifier whose independence from the delegator is established. Experiment notes

Whose claims make a difference?

A $651 purchase exceeds a $650 cap. If the gate blocks the attempt, that tells us something about the control. If the purchase goes through, that single $1 breach refutes the universal claim that every

completed purchase was authorized under the $650 cap, even if the dollar variance looks trivial. Under Zhao and colleagues' rule, one authenticated exception within the scope of a universal claim contradicts it even when other records are missing.

Once a breach is established, the engagement must separate two explicit judgments:

- **For the shared-controls engagement**, the question is systemic: how did a $651 request bypass the gate, and which other grants relying on that same control version could be affected?
- **For the traveller or principal**, the question is remedial: what are they owed, and how is that obligation settled?

The same breach can carry a $1 financial impact on one ticket while exposing a control defect that invalidates assertions across thousands of active grants. In GHG verification, **materiality** often operates as an aggregate threshold on a reported total, paired with qualitative criteria. For a delegation, an aggregate threshold can hide both a systemic control flaw and a serious local error. The proposal here is to use quantitative thresholds where useful while evaluating systemic control integrity and what affected parties were owed in their own right. ISO 14064-3

Take the file sent to the wrong recipient. The deploying organization might see a small incident because its own financial loss is slight. The person whose information was disclosed has a claim that changes what the verifier should treat as material. What process gets that person's claim into the criteria? Can they contest an omission after the fact, and who has standing to hear that challenge and revise the finding? A job-type protocol could make those decisions explicit.

Some exposures become visible only when looking across grants.

- **Threshold effects across a series of grants.** Several agents can make purchases within their individual caps and together exceed the amount the organization would have authorized for the job as a whole. Judging the combined exposure requires a rule for which grants must be evaluated together.
- **Correlated failures across grants.** Where many agents or delegators depend on the same underlying model or control, a single flaw can cause many delegations to fail in the same way at once. A failure mode that looks immaterial within one delegator's grant can be large in the aggregate, and a recourse pool sized for independent incidents can be overwhelmed when failures are correlated. Assessing that exposure requires evidence and recourse across grants, not just within one.

What would make this assurance?

I would use ISAE 3000 (Revised) as a starting point for the form of the engagement and draw on ISO 14064-3 for validation and verification practice and ISO 14065 for requirements on the bodies doing that work (noting that for GHG statements themselves, ISSA 5000 replaces ISAE 3410 for periods beginning on or after 15 December 2026). The GHG accreditation chain shows how a user of the report can assess whether the body issuing it is independent and competent for the work.

1. **Protocol.** The applicable methodology sets the domain-specific rules and the evidence the validator or verifier needs.
2. **Validation and verification body (VVB).** An independent VVB assesses the project against those rules. ISO 14065 applies the general requirements in ISO/IEC 17029 to bodies that validate and verify environmental information. Verra requires accreditation or approval for the applicable sectoral scope. A VVB scoped only for agricultural removals would need the geological storage scope to verify a storage project. In the proposed agent scheme, a provider assessed for fare rules, airline receipts, and

payment ledgers would need to demonstrate additional competence to assure weather data, clinical work, or code deployment. ISO 14065 · Verra's sectoral scopes

3. **National accreditation body.** An accreditor such as the Standards Council of Canada assesses the VVB's technical competence and impartiality for the scope sought, including how it manages conflicts of interest. The Council reviews files and witnesses selected verification work over the accreditation cycle. Standards Council of Canada accreditation program
4. **International recognition.** Under the Multilateral Recognition Arrangement, peers evaluate accreditation bodies, often through regional bodies, against common requirements including ISO/IEC 17011. Global Accreditation Cooperation Incorporated (Global ACI) took over this recognition role from the International Accreditation Forum (IAF) in January 2026. Mutual recognition gives regulators and counterparties a basis for accepting statements issued within a recognized scope across borders, subject to applicable requirements. Global ACI's evaluation process · Multilateral recognition · IAF transition

For agent delegation, I would borrow that structure for formal engagements, including those on shared controls. A scheme would need to define which job scopes an accreditor can assess and the terms on which regulators or counterparties would recognize the resulting reports.

Nine instruments from GHG verification inform this proposal, alongside the execution gate described above.

#	GHG MRV instrument	Adaptation for agent delegation
1	**Validation before, verification after**	Check authority, capacity to answer, and hard constraints before activation, then verify authorization and how eligible requests were resolved or escalated.
2	**Materiality threshold and qualitative criteria**	Use quantitative thresholds where useful and assess effects on shared controls and what affected people are owed.
3	**Conservativeness**	Place the burden of uncertainty on the party that benefits by narrowing or holding grants before activation and leaving claims unresolved (*undecidable*) afterward when essential evidence is missing.
4	**Buffer pools and reversal liability**	Establish before activation who answers, who pays, what recourse covers, and how long it lasts after expiry, with pools for replaceable losses and gates to prevent irreversible harm.
5	**Scoped accreditation and independence**	Use job-type criteria and independent practitioners whose assurance body is accredited for that scope by a peer-evaluated accreditor, with rules for cross-border recognition and practitioner selection when the assessed party pays.
6	**Limited and reasonable assurance levels**	Use deterministic checks backed by assurance over shared controls for contained grants, exception review or limited assurance under the bounded-tier rule above, and reasonable assurance on individual consequential grants.
7	**Validation and verification statements**	State level, scope, limits, and control mode (structural removal, deterministic gate, human approval, or instruction-only) in opinions on shared controls against the specified protocol version for

#	GHG MRV instrument	Adaptation for agent delegation
		contained grants and in pre-job and post-job opinions for consequential grants, with a record for each grant.
8	**Project registry and credit retirement**	Use a grant registry to track active, narrowed, spent, and revoked grants across systems, require validation before activation where the rules call for it, and retire grants once spent.
9	**Public consultation and appeals on methodologies**	Give affected people standing and a procedure to shape materiality criteria and challenge job boundaries or omissions after the fact.

ISAE 3000 (Revised) · ISSA 5000 announcement

What would create demand for this kind of assurance? In carbon markets, a registry will not issue credits without a verification statement. Nothing yet forces an AI system to refuse a delegation for lack of assurance, and self-imposed governance rarely survives convenience. If a forcing function emerges, it is likely to come from an insurer underwriting agent errors and omissions, an enterprise procurement rule for vendors whose agents touch client systems, a counterparty platform (such as a payment network or clearinghouse) requiring assurance over transaction controls before accepting agent-initiated transactions, or a regulator licensing agents in high-consequence domains.

The local experiments exercise parts of bounded delegation with synthetic or mock conditions. They give some evidence about controls, records, and differences in interpretation. The full engagement remains a proposal, as do the protocols and forcing functions that would require a system to refuse a delegation for lack of assurance. If you would like to test the proposal, we could work through a paper dry run of the flight case using the existing evidence or try a small delegation with an independent verifier. I would also welcome examples from other kinds of work and challenges to the way I have framed authority or materiality. You can reach me at will@emberwall.net.

I am writing in a personal capacity, and these views are my own. I used Claude, Codex, and Gemini for research and drafting assistance, and rewrote the text myself.

Related work

- Johnny Kao (2026), *Delegation Contracts for Agentic AI*, working paper v1.6, 23 September 2026.
- Claudiu Hulea (2026), *Verifiable Mandates for Agents Acting Across Administrative Domains*, individual Internet-Draft.
- Yue Zhao et al. (2026), *Auditable Claims about AI Agents*.
- Shea Brown (2026), *We Already Have a Framework for AI Assurance… What We Need is the Right Content*, and BABL AI (2026), *Bias audit of the Eightfold Matching Model*.
- Stephen Darwall (2013), "Demystifying Promises," in *Honor, History, and Relationship: Essays in Second-Personal Ethics II*, Oxford University Press, pp. 131–154. Oxford Academic
- On undiminished answerability and anticipatory oversight, see Kennedy Chun Yin Kong (2026), "The delegation illusion", *AI and Ethics* 6:518, and Baum et al. (2026). On permission transfer, outcome verification, and underwriting agent transactions, see Tomašev, Franklin and Osindero (2026) and Hua et al. (2026).