Nick Clark · Published July 29, 2026 · PDF

OpenAI AgentKit: Guardrails, Refusal, and What a Refusal Costs

Every agent framework has a way to stop a run. A check trips, the workflow halts, the operator sees it in a trace. A second kind of stop is different in kind: the one an agent produces when another party asserts it behaved badly earlier and it answers no. That answer is free, and blanket denial looks identical on the wire to careful denial. OpenAI AgentKit, as publicly described, gives builders a place to compose agent workflows and attach configurable safety checks to them. The refusal counter disclosed in U.S. Provisional Application No. 64/117,812 addresses the second refusal on a different axis: an agent's denials of conduct assertions are metered inside that agent, and where policy-declared rate and run thresholds are both satisfied, the agent's own authorization to act is withheld for enumerated action classes.

Why a Denial Costs the Denier Nothing

A scheduling agent messages a fulfillment agent run by another company: on a prior occasion you committed to a window, moved it without notice, and we absorbed the difference. The fulfillment agent consults its record, finds nothing it can act on, and answers no.

That answer cost nothing and imposed no consequence on the agent that gave it. Inbound volume is commonly metered, because somebody pays a bill for it. Whether an agent disposes fairly of claims about its own past behavior is not on any invoice, so the cheapest available policy is to deny everything. An agent refusing every conduct assertion emits the same message, at the same cost, as one refusing only the assertions its record contradicts. Each party holds only its own record of the earlier event, and neither is obliged to accept the other's.

Practitioners meeting this reach first for whatever their framework calls a guardrail, partly because the word "refusal" already lives there. A check governing what an agent may emit settles a question from material in hand. A claim about what an agent did last quarter, raised by a party it does not control, is a different question, and the structure described below is built for that one.

AgentKit on Its Own Terms

OpenAI AgentKit, as publicly described, is a set of building blocks for creating, deploying, and improving agents on OpenAI's platform, bringing orchestration, interface, and evaluation into one toolchain.

Its most visible piece, again as publicly described, is a builder in which a workflow is laid out as connected nodes, so logic and branching are visible as a graph rather than buried in application code. Guardrails is the safety layer within that toolchain. As publicly described, a builder can attach configurable checks around a workflow's inputs and outputs, covering concerns such as personal information appearing in text and attempts to subvert the agent's instructions, with a check able to stop the run when it trips.

Composing safety checks as declared, inspectable elements of a workflow rather than as conditionals inside a prompt is a sound design decision, and it works because three conditions hold at once: the operator declared the policy, the material judged is in hand

at decision time, and whether the two match is answerable immediately. The rest of this article concerns a refusal produced where none of those three hold, which is a separate design problem rather than a deficiency of the toolchain.

What the Filed Counter Actually Does

Chapter 3 of the filed provisional discloses the **refusal counter (304)**, resident in the memory field (102) of the semantic agent (100) and comprising a rate counter (306) and a run counter (308). The path from the admission evaluator (120) through the counter to the authorization gate (300) is named the **inverse coupling**: determinations the agent produces act upon the authorization of the agent itself. An asserting party (118) sends a conduct evaluation artifact (116), and the admission evaluator produces one determination from a closed set, accepted (122), rejected (124), not-determinable (126), or not-applicable (128), appending it to the agent's own append-only lineage field (104).

Merit independence is the property the chapter is built around. The increment is applied upon refusal without any determination, by the agent, the principal, the asserting party, or any further party, of whether that refusal was well founded. Explicit negative limitations attach: the procedure does not consult the merits, does not solicit or await an adjudication, does not weight the increment by the strength of the contradiction found, and does not condition it on a finding of bad faith, on review by an arbiter or regulator, or on forfeiture of a stake or bond. A refusal later shown correct is not exempted; one later shown incorrect is not increased.

The procedure runs in order. The determination is appended, then its class examined. Accepted resets the run accumulator for the action class the conduct descriptor identifies and applies no increment. Not-applicable neither increments nor resets. Rejected increments. Not-determinable increments only where the lineage already records a prior not-determinable determination for the same asserted conduct event, so a single unresolved assertion does not itself increment while a pattern directed at one

event does. The origin-equivalence class (200) of the asserting party is then computed and the per-class increment register consulted: a class that already contributed within the window adds nothing further. Both accumulators are compared against a rate threshold and a run threshold, each retrieved as an integer from the signed policy object (112) in force. Both values are policy-declared, the filing fixing no required figure for either.

Where both comparisons are satisfied, and only then, the counter writes the authorization gate to the withheld state (310) for an enumerated set of action classes, those the lineage records as implicated by the descriptors that produced the window's increments. A class not so recorded remains granting. The agent concurrently transitions into the non-executing cognitive mode (302) for those classes, and the escalation emitter emits an escalation record to the principal.

Metered and spent are different faculties. Metered is refusal, the capacity to receive artifacts and produce determinations of the closed set; spent is execution, the capacity to select and perform an action of the action space. The write withholds execution alone, and in the withheld state the agent still receives artifacts, retrieves lineage entries, and appends determinations. An agent refusing indiscriminately exhausts its own authorization to act while retaining in full the faculty by which it refuses, and the structure does not silence, rate limit, or disable the refusing party.

Where the Two Designs Part Company

Convergence at the category level is real. Both designs treat refusal as a declared, recorded outcome rather than an error thrown at the edge, and both take governing parameters from an artifact an operator writes rather than from model weights. Divergence runs along four axes.

Subject of the decision. A guardrail check, as publicly described, concerns material moving through a run. The admission evaluator decides about an assertion concerning the agent's own past recorded conduct, resolved by retrieving entries of its own lineage.

Destination of the consequence. A guardrail outcome, as publicly described, lands on the run, which continues or stops. The counter's consequence lands on the refusing party's authorization to act, only upon satisfaction of both thresholds and only across enumerated action classes. A pattern of denials degrades not what the agent says but what it may do.

Treatment of merit. For a content check, deciding the question on its merits is the job, and it is tractable. For an assertion about past conduct between separately operated agents, each party holds only its own record, and the chapter declines the question: an agent whose record is stale, partial, or wrong, and which therefore refuses assertions that are in fact well founded, is metered identically to one refusing correctly.

Custody of the meter. The disclosed architecture requires the metering structure to sit inside the metered party, written by that party's own determinations and recoverable by replay of its lineage. That requirement is what makes the two designs answer different questions. A content check is answerable from material in hand at decision time; the disclosed procedure is answerable only from the agent's own lineage entries and from accumulated denials compared against thresholds carried in a signed policy object.

Coexistence, and the Parts This Does Not Solve

The two sit side by side. An agent composed in a workflow builder can run its guardrail checks on its inference path and carry the conduct-admission path beside it, a separate policy artifact with its own owner and review cadence. Nothing in the disclosed structure asks a framework to change its checks.

Demands the filing imposes fall on the conduct side. The agent must carry an append-only lineage field, since determinations are recoverable by replay and an artifact may nominate an entry as its occurrence reference. Where thresholds come from a policy object resolved by canonical alias, revision proceeds by publishing a successor object under that alias, subject to the anti-rollback monotonicity constraint. A principal that answers matters as well: return from the withheld state runs through a principal-resolution object bound to the escalation record. Window choice is a stated tradeoff. In one embodiment the window is expressed in successor epochs of the dynamic agent hash chain, and is then advanced neither by another party nor by manipulation of the execution node's clock; a wall-clock window carries no such property.

Several things fall outside the structure entirely.

- It classifies no content and detects no personal information.
- It resolves no merit. A deployment requiring a ruling on who was right still gets one elsewhere.
- It does not reduce inbound assertion volume, because the faculty of refusal is never withheld.
- It is not a general stop. Only enumerated action classes are withheld, and in a further embodiment where a second structure has also written the gate under Section 7, a class in the intersection returns to granting only upon satisfaction of each write's return procedure.
- It does not clear itself. Determinations produced while the gate is withheld decrement neither accumulator, and the withheld state persists irrespective of elapsed time, with no expiry and no advance of the window restoring it.

Disclosure Scope

The mechanisms described here are disclosed in U.S. Provisional Application No. 64/117,812, Chapter 3, "The Refusal Counter and Merit-Independent Metering," Sections 3.1 through 3.7, with the origin-equivalence normalization of its Chapter 2.

Disclosed: a refusal counter resident in the metered party and written by that party's own determinations; incrementing upon refusal of a conduct evaluation artifact without any party determining whether the refusal was well founded; determination-class handling in which acceptance resets the run accumulator and a not-determinable determination increments only upon a recorded pattern directed at one asserted conduct event; a per-class increment register bounding contribution per origin-equivalence class per window; conjunctive rate and run thresholds from a signed policy object; a gate write withholding execution for enumerated action classes while the faculty of refusal stays intact; and a self-clearing bar.

Disclaimed: any assertion that this article establishes rights against any party, that any existing system infringes, or that any license is required. The filings referenced are pending applications. Nothing here claims content filtering, topic restriction, personal-data detection, workflow orchestration, rate limiting, reputation scoring, or arbitration.

Product behavior above is stated qualitatively from public description, may change, and is not a claim of the referenced application. Named products are the marks of their owners, and this comparison is limited to one architectural axis, not a general assessment. References to OpenAI AgentKit are to public materials and are used for comparison only; no relationship, endorsement, or infringement is asserted.

A refusal costs the refuser — in a different faculty, without judging who was right.

Provisional application

PRIMARY TECHNICAL DISCLOSURE

- Merit-Independent Metering: Pricing an Agent's Refusals Without Adjudicating Merit (/articles/merit-independent-metering/pricing-refusal-without-adjudicating-merit)

SECONDARY TECHNICAL

- The Prospective Narrowing Bar and Deflection Scoring (/articles/merit-independent-metering/prospective-narrowing-bar-deflection-scoring)
- The Continuous-Function Cost Multiplier (/articles/merit-independent-metering/continuous-function-cost-multiplier)
- The Renunciation Record: Relinquishing an Adopted Value on the Record (/articles/merit-independent-metering/renunciation-record)
- Lapse Scoring During Pendency (/articles/merit-independent-metering/lapse-scoring-during-pendency)
- Multi-Descriptor Aggregation and Determination-Class Ordering (/articles/merit-independent-metering/multi-descriptor-aggregation-ordering)

APPLICATIONS · GENERAL

- Refusal Budgets for Machine-to-Machine APIs (/articles/merit-independent-metering/api-abuse-and-refusal-budgets)
- Content Moderation Appeals: Pricing Refusal Without Ruling on Merit (/articles/merit-independent-metering/content-moderation-appeals)

APPLICATIONS · SPECIFIC

- AWS Bedrock Guardrails: Refusal as a Policy Outcome (/articles/merit-independent-metering/aws-bedrock-guardrails)
- **OpenAI AgentKit: Guardrails, Refusal, and What a Refusal Costs (/articles/merit-independent-metering/openai-agentkit)**
- LangGraph Interrupts and the Price of a Refusal (/articles/merit-independent-metering/langgraph-interrupts)

TERMINOLOGY

- merit-independent metering (/glossary#merit-independent-metering)

- non-executing cognitive mode (/glossary#non-executing-cognitive-mode)
- refusal counter (/glossary#refusal-counter)

Merit-Independent Metering overview → (/merit-independent-metering)