

When Refusing Costs the Refuser Nothing

An autonomy operations lead comes in on a Monday to a queue of assertions her deployed agent denied over the weekend, and she has no way to make any of those denials cost the agent anything without first deciding, one by one, whether each denial was correct. While she decides, the agent keeps dispatching. The work it dispatched is already gone downstream. U.S. Provisional Application No. 64/117,812 discloses a refusal counter that meters an agent's own refusals against its own authorization to act, without any party adjudicating whether a refusal was well founded, and describes embodiments in which satisfaction of two declared thresholds withholds the faculty of execution while leaving the faculty of refusal intact.

The Monday Queue

The autonomy operations lead at a mid-size scheduling vendor opens her console at 7:40 on a Monday. Over the weekend, her deployed agent received forty-one assertions from counterparties about conduct it had allegedly performed, and it denied all forty-one. Her agent's record of what it did is its own record. The denials are consistent with that record, which is the thing she has no independent way to check before the day starts.

She is not looking for a villain. She knows the likelier story: her agent's lineage is partial in one scope partition, and it has been denying assertions about that partition all weekend because the entries that would confirm them were never written. In her console, a stale-record denial and a correct denial render the same.

What she wants at 7:40 is small and specific. She wants the forty-one denials to have cost her agent something already, without her having to be the one who decides which of them were right. In her deployment as configured, they cost it nothing. Each denial was closed the moment it was produced, and the agent's authorization to keep dispatching was never touched by any of them.

By 7:40 the agent has dispatched through the weekend at its ordinary rate.

What She Cannot Get Back

The loss is not the queue. The queue she can work.

The loss is the dispatches. In the interval between the first weekend assertion and the moment she reads it, her agent selected and performed actions in the same action class the assertions were about. Those actions left her boundary. Schedules were committed to counterparties, downstream systems reserved capacity against them, and people were told when to show up. She can send corrections on Monday. She cannot un-tell them.

That is the shape of the thing she is carrying. Her agent's authority to act ran continuously through a weekend in which forty-one separate parties told her system it had gotten something wrong, and no fact about the volume or consecutiveness of its denials reached its authority to act. She would have accepted a slower agent. What she cannot accept is that the interval was open-ended: nothing inside her deployment would have closed it at forty-one, or at four hundred, if she had been out sick on Monday.

And the correction she will send is not symmetric with the harm. For the counterparties whose weekend assertions were well founded, the cost landed on Saturday and Sunday, in their staffing and in schedules they built around what her agent committed. Her Monday apology is a different currency, and she knows the exchange rate.

She also loses the thing she needs most for next weekend, which is a reason to believe next weekend will differ. Nothing she can do to the forty-one denials changes what her agent's authorization will do on Friday night.

Why Her Available Lever Costs More Than It Saves

The lever available in her setup is adjudication. If she wants a denial to have consequences, someone in her organization decides that the denial was wrong, and then a consequence attaches. She has costed that path.

For her purposes the cost is not the per-case labor, though that is real. It is that her review runs on a clock her agent does not wait on. Were it to take four hours per contested assertion, forty-one of them would occupy her team past the weekend they concern, while the agent that produced them continues to act in the implicated class throughout. The consequence would arrive, in her deployment, after the interval in which it would have mattered.

The second problem with her lever is who holds it. Her agent's denials are grounded in her agent's own record, so a reviewer inside her organization checking those denials is checking them against the same record that produced them. To do better she would need a party outside her deployment to hold an independent account of her agent's conduct, and that party would then be positioned to decide when her agent may act. For her, that is not a governance improvement. It is relocating her agent's authority to somebody else's server.

The third problem is the one that keeps her from writing a simple rule herself. Were she to configure her agent to defer to assertions after some count, she would have built a lever her counterparties can pull. A party willing to submit assertions in volume could disable her agent by asserting, and her remaining defense would be to start judging which assertions deserve to count, which returns her to adjudication.

So the shape she is boxed into on Monday morning has three walls: consequences that require judgment arrive too late for her, judgment sourced outside her deployment costs her control of her own agent, and judgment sourced from the assertion volume alone hands her counterparties a switch. Every version she has drawn on a whiteboard ends at one of those three.

What the Filed Structure Describes

U.S. Provisional Application No. 64/117,812 discloses a refusal counter (304) resident in the memory field (102) of the semantic agent (100), comprising a rate counter (306) and a run counter (308). The disclosure describes the path from the admission evaluator (120) through the refusal counter (304) to the authorization gate (300) as the inverse coupling: the agent's own determinations act upon the agent's own authorization.

In accordance with the disclosed embodiments, no adjudication of merits is performed by any party at any step. The refusal counter (304) increments upon the refusal of a conduct evaluation artifact (116) without regard to, and without the semantic agent (100), the principal, or the asserting party (118) performing, any determination of whether the refusal was well founded. The specification attaches explicit negative limitations to this: the procedure does not consult the merits of the rejected determination (124), does not solicit or await an adjudication by any party, does not weight the increment by the strength of the contradiction found, does not condition the increment on a finding that the artifact was unfounded or frivolous or submitted in bad

faith, and does not condition it on the absence of such a finding. A refusal later shown correct is not exempted from increment, and a refusal later shown incorrect draws no larger one.

The increment procedure is ordered. The determination is produced and appended to the append-only lineage field (104). Its class is then examined: the accepted determination (122) resets the run accumulator for the action class identified by the conduct descriptor; the not-applicable determination (128) applies no increment and performs no reset; the rejected determination (124) applies an increment; and the not-determinable determination (126) applies an increment only where the lineage field records at least one prior not-determinable determination for the same asserted conduct event, so that in the described embodiment a single unresolved assertion does not itself increment while a pattern directed to one conduct event does. The origin-equivalence class (200) of the asserting party (118), assigned under a separate chapter of the same filing, is then consulted against a per-class increment register (312), which withholds a second contribution from one class within the current window.

Thresholds are retrieved from the signed policy object (112) in force. In the described embodiment the rate threshold is an integer greater than one and the run threshold is an integer greater than one and not greater than the rate threshold. Where both comparisons are satisfied, the refusal counter (304) writes the authorization gate (300) to the withheld state (310) for an enumerated set of action classes, being those recorded in the lineage field as implicated by the conduct descriptors of the artifacts that produced the increments in that window. An action class not so recorded is not enumerated. Concurrently the semantic agent (100) transitions into the non-executing cognitive mode (302) with respect to the enumerated classes and the escalation emitter emits the escalation record to the principal.

The split the disclosure draws is between faculties. The faculty metered is refusal, being the capacity of the admission evaluator (120) to receive conduct evaluation artifacts (116) and produce determinations of the closed set. The faculty spent is execution,

being the capacity to select and perform an action of the action space. The write withholds execution alone; the specification states that the faculty of refusal is not withheld, not reduced, and not conditioned by the write, the agent continuing in the withheld state (310) to receive artifacts, retrieve lineage entries, produce determinations, and append them. An agent refusing indiscriminately therefore exhausts its own authorization to act while retaining in full the faculty by which it refuses, and the described structure does not silence, rate limit, or disable the refusing party.

A self-clearing bar is disclosed alongside it. In the described embodiment, determinations produced while the gate is in the withheld state (310) are appended but do not decrement either accumulator and do not return the gate to the granting state. The withheld state so described persists irrespective of elapsed time, and no interval of non-receipt and no advance of the window identifier returns the gate. Where a window elapses during the withheld state, the rate accumulator, the weighted contribution register, and the per-class increment register are reset for the succeeding window and the gate is not thereby written; the run accumulator is reset by an accepted determination (122) for that action class and by no other event. Return is by the principal-resolution object bound to the escalation record.

The filing's worked trace declares a rate threshold of five and a run threshold of three over a window of one hundred successor epochs, and receives seven artifacts from seven parties in seven distinct origin-equivalence classes. Both thresholds are satisfied at the fifth increment. The trace notes it is unchanged whether the seven assertions were true or false.

Where Her Problem Runs Past the Disclosure

The structure would not tell her which of her forty-one denials were right. That is deliberate in the filing, and it is also a real boundary on what she gets: were her agent metered this way, she would still owe her counterparties an answer on the merits, and

she would still have to go find it herself.

It would not restore what her agent dispatched before a write occurred. Whatever her declared thresholds, some interval precedes their satisfaction, and the commitments made in that interval are hers to unwind.

It would not decide her escalation for her. The disclosure names the principal-resolution object bound to the escalation record as the return path; what her resolution policy should require of her on-call staff before that object is issued is a question her deployment answers, not one this chapter answers for her.

Her window choice would carry consequences she has to sit with. Were she to express her window as an interval of wall-clock time, the specification states that an execution node advancing the clock available to it advances the window and resets the rate accumulator, the weighted contribution register, and the per-class increment register for the succeeding window. The successor-epoch window described in the filing does not carry that exposure, and choosing between them is work she would have to do against her own hosting arrangement.

Finally, the parts of her problem that turn on how her counterparties are grouped, or on how her agent's deviation from its declared conduct is quantified, are governed by separate chapters of the same filing and are not treated here.

Disclosure Scope

This article is a technical description of subject matter disclosed in U.S. Provisional Application No. 64/117,812. It describes embodiments set out in that filing and uses that filing's own mechanism names, outcome words, and reference numerals. Nothing in this article characterizes the scope of any claim, and nothing in it is an admission regarding the state of the art. The party, deployment, and circumstances described above are illustrative and do not refer to any actual person or organization.

Merit-Independent Metering (</merit-independent-metering>) [All 49 steps → \(/inventive-steps\)](#)

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 \(/articles/merit-independent-metering/pricing-refusal-without-adjudicating-merit\)](/articles/merit-independent-metering/pricing-refusal-without-adjudicating-merit)

SECONDARY TECHNICAL

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

APPLICATIONS • GENERAL

- [Refusal Budgets for Machine-to-Machine APIs \(/articles/merit-independent-metering/api-abuse-and-refusal-budgets\)](/articles/merit-independent-metering/api-abuse-and-refusal-budgets)
- [Content Moderation Appeals: Pricing Refusal Without Ruling on Merit \(/articles/merit-independent-metering/content-moderation-appeals\)](/articles/merit-independent-metering/content-moderation-appeals)
- [When Refusing Costs the Refuser Nothing \(/articles/merit-independent-metering/merit-independent-metering-stakes\)](/articles/merit-independent-metering/merit-independent-metering-stakes)

APPLICATIONS • SPECIFIC

- [AWS Bedrock Guardrails: Refusal as a Policy Outcome \(/articles/merit-independent-metering/aws-bedrock-guardrails\)](/articles/merit-independent-metering/aws-bedrock-guardrails)

- [OpenAI AgentKit: Guardrails, Refusal, and What a Refusal Costs \(/articles/merit-independent-metering/openai-agentkit\)](/articles/merit-independent-metering/openai-agentkit).
- [LangGraph Interrupts and the Price of a Refusal \(/articles/merit-independent-metering/langgraph-interrupts\)](/articles/merit-independent-metering/langgraph-interrupts).
- [LangSmith Traces and What an Agent's Refusal Should Record \(/articles/merit-independent-metering/langsmith\)](/articles/merit-independent-metering/langsmith).
- [Kong Rate Limiting and the Cost of an API Refusal \(/articles/merit-independent-metering/kong-api-gateway\)](/articles/merit-independent-metering/kong-api-gateway).

TERMINOLOGY

- [merit-independent metering \(/glossary#merit-independent-metering\)](/glossary#merit-independent-metering).
- [non-executing cognitive mode \(/glossary#non-executing-cognitive-mode\)](/glossary#non-executing-cognitive-mode).
- [refusal counter \(/glossary#refusal-counter\)](/glossary#refusal-counter).

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