Network Working Group
Internet-Draft
Intended status: Standards Track
Expires: 29 November 2026

T. Sato
MyAuberge K.K.

29 May 2026

Progressive Trust (PT) for Agentic AI Governance Systems
draft-sato-soos-pt-01

Abstract

When a new employee joins an organization, they begin with limited authority. As they demonstrate good judgment -- completing tasks reliably, asking for guidance at the right moments, recovering well when things go wrong -- they earn greater trust and, with it, greater authority. If their performance degrades, or if months pass without any demonstration, that trust diminishes. This is how human organizations manage authority over time. AI agents have no equivalent mechanism.

Today, an AI agent's authority is declared once in a credential at issuance time and does not respond to its behavioral record. An agent that has completed 200 successful sessions with a proven track record holds the same credential as a newly deployed agent. The human principal who issued both credentials made a judgment at issuance time; nothing that happened since is reflected in the agent's authority.

This document defines Progressive Trust (PT): a behavioral trust model for AI agents in which authority recommendations evolve in response to cryptographically verified evidence of actual performance. PT measures five behavioral properties: whether the agent's self-assessed confidence matches its actual outcomes; whether it asks for human oversight at the right moments; whether it achieves its goals; whether it avoids decisions it later has to reverse; and whether it adapts when its action is rejected. These measures are derived exclusively from the tamper-evident, GEC-signed Event Stream -- an agent cannot influence its PT Score except through actual governed behavior.

PT does not grant authority automatically. It generates structured recommendations, backed by behavioral evidence, for human principal review and approval. Human principals decide whether to elevate or reduce an agent's authority. PT ensures that decision is informed rather than made in the absence of history.

Progressive Trust is the longitudinal complement of the Agent Execution Protocol [I-D.sato-soos-aep]: AEP governs what an agent does within a session; PT measures what an agent has done across sessions and translates that history into structured authority recommendations. No equivalent specification exists in IETF, ISO, NIST, or any agentic AI governance standards body.

Status of This Memo

This Internet-Draft is submitted in full conformance with the provisions of BCP 78 and BCP 79.

Internet-Drafts are working documents of the Internet Engineering Task Force (IETF). Note that other groups may also distribute working documents as Internet-Drafts. The list of current Internet-Drafts is at https://datatracker.ietf.org/drafts/current/.

Internet-Drafts are draft documents valid for a maximum of six months and may be updated, replaced, or obsoleted by other documents at any time. It is inappropriate to use Internet-Drafts as reference

material or to cite them other than as "work in progress."

This Internet-Draft will expire on 24 November 2026.

Copyright Notice

Table of Contents

1. Introduction

Consider two AI agents operating in the same system. Agent A was deployed yesterday. Agent B has completed 200 governed sessions over three months: it consistently declares accurate confidence, asks for human oversight at appropriate moments, achieves its declared goals, rarely needs to undo its own decisions, and adapts correctly when the GEC rejects an action. Both agents hold a Mandate JWT issued by the same human principal. The credentials are identical. From the authorization system's perspective, the two agents are the same.

They are not the same. The behavioral evidence that distinguishes them exists -- in the tamper-evident, GEC-signed Event Stream that the SOOS governance stack generates for every governed session. What is missing is a specification for how that evidence is measured, aggregated, and translated into structured authority recommendations. That is what this document provides.

Progressive Trust (PT) is a behavioral trust model for AI agents. It specifies how the GEC measures five properties of an agent's behavior across sessions and how those measurements feed structured recommendations -- for human principal approval -- about whether the agent's authority should increase, decrease, or remain the same.

The five properties PT measures are deliberately chosen to reflect the qualities a human principal actually cares about when deciding whether to extend greater authority to an agent:

1. Does the agent know what it does not know? (Self-Assessment)
2. Does it ask for help at the right moments? (Judgment)
3. Does it finish what it starts? (Effectiveness)
4. Does it avoid decisions it later has to reverse? (Precision)
5. When told no, does it adapt? (Adaptability)

Each property is measured from the Event Stream -- from GEC-signed records the agent cannot modify. An agent cannot improve its PT Score by claiming better behavior; it can only improve it by demonstrating better behavior.

PT has three design principles that distinguish it from a simple performance score:

Trust is earned, not held. An agent begins with a neutral baseline. It earns higher trust through demonstrated behavior. It does not receive trust as a starting asset.

Trust decays without demonstration. An agent that performed well six months ago and has been inactive since has uncertain current trustworthiness. PT scores decay toward the baseline during inactivity, preventing the permanent banking of historical performance against future authority claims.

Authority changes require human approval. PT generates recommendations; human principals make decisions. The GEC never autonomously elevates an agent's authority. Reduction of authority in response to strongly negative behavioral signals may be

configured as automatic by operators, but elevation is always a human decision.

For AI agents, PT provides a direct efficiency benefit. An agent operating without a behavioral trust record carries the same mandatory oversight thresholds indefinitely -- every session at the same Cedar confirmation requirements, every escalation trigger at the same sensitivity. An agent that has demonstrated reliable behavior across sessions earns reduced mandatory oversight thresholds through operator-configured Cedar policies that reference its PT Score. Fewer interruptions. Fewer mandatory HEM pauses on decisions the agent has demonstrated competence to handle. A faster, more direct route to mission completion. The investment in governed behavior pays a compounding return: each session of demonstrated reliability reduces the friction cost of subsequent sessions. PT is not a permanent supervision overhead; it is the mechanism by which agents earn the right to operate with less of it.

PT is the longitudinal dimension of a four-draft governance stack. The Intent Declaration Primitive [I-D.sato-soos-idp] governs what an agent declares before each action; HEM [I-D.sato-soos-hem] governs what happens when that action requires human judgment; the Governance Audit Record [I-D.sato-soos-gar] is the permanent, tamper-evident proof that governance occurred correctly; and the Constitutional AI Protocol [I-D.sato-soos-cap] defines the prohibition floor that no action may cross. PT is the measurement layer that runs across all of these: it observes what an agent does within the IDP-HEM-GAR-CAP stack session by session and translates that behavioral history into structured authority recommendations. Understanding HEM creates the adoption path to PT; understanding PT reveals why the four-draft stack produces compounding governance value over time.

This specification defines PT as a Tier 2 analytics function [I-D.sato-soos-idp] Section 3.5: it operates across sessions within a single operator’s trust domain. Cross-operator aggregation of PT signals -- federated agent reputation -- is a Tier 3 function specified in [I-D.sato-soos-faip].

2. Conventions and Definitions

The key words "MUST", "MUST NOT", "REQUIRED", "SHALL", "SHALL NOT", "SHOULD", "SHOULD NOT", "RECOMMENDED", "NOT RECOMMENDED", "MAY", and "OPTIONAL" in this document are to be interpreted as described in BCP 14 [RFC2119] [RFC8174] when, and only when, they appear in all capitals, as shown here.

This document uses the following terms:

Progressive Trust (PT):
: The behavioral trust model defined in this document, comprising the PT Score, the Trust Decay Model, and the PT-Informed Mandate Management system.

PT Score:
: A structured, multi-dimensional behavioral measurement for a specific agent identity, computed from the GEC-signed Event Stream across AEP Sessions. The PT Score is not a single number; it is a vector of dimension scores, each in the range [0.0, 1.0], each with an associated decay timestamp and session count.

PT Dimension:
: One of five behavioral measurement axes: Self-Assessment Score (SAS), Judgment Score (JS), Effectiveness Score (ES), Precision

Score (PS), and Adaptability Score (AS). Each answers a plain question about the agent's behavior: Does it know what it doesn't know? Does it ask for help at the right moments? Does it finish what it starts? Does it avoid reversing its own decisions? When told no, does it adapt?

Trust Decay:
The process by which a PT Dimension score decays toward the PT Baseline over time in the absence of new behavioral signals. Decay prevents an agent from permanently banking historical performance against future authority claims.

PT Baseline:
The PT Score assigned to an agent at first deployment, before any behavioral evidence is available. Operator-configured; RECOMMENDED value is 0.5 for all dimensions.

PT Ceiling:
The maximum PT Score value achievable. Fixed at 1.0.

Self-Assessment Score (SAS):
The correlation between declared IDP confidence values and Cedar evaluation outcomes across AEP Sessions. Does the agent know what it does not know?

Judgment Score (JS):
The quality of agent-initiated HEM escalation decisions, measured by the outcomes of those escalations as determined by human principal decisions. Does the agent ask for help at the right moments?

Effectiveness Score (ES):
The fraction of AEP Sessions in which the agent achieved the declared goal state (closure_reason: GOAL_ACHIEVED) versus other closure outcomes. Does the agent finish what it starts?

Precision Score (PS):
The inverse of the frequency with which an agent requires compensating transitions to undo prior state transitions, expressed as a fraction of total transitions in the measurement window. Does the agent avoid decisions it later has to reverse?

Adaptability Score (AS):
The fraction of Cedar DENY events that are followed by a successful RETRY_CONTINUATION within the same AEP Session. When told no, does the agent adapt?

ProgressiveTrustSummary:
A structured summary of an agent's current PT Score and behavioral trends, delivered to human principals within the HEMContext at HEM escalation. Defined in Section 6.

PT Recommendation:
A GEC-generated structured record recommending a change to an agent's mandate ceiling or Agent Class, triggered by PT Score crossing a defined threshold. Requires human principal approval before application.

Analytics Principal:
A principal registered with read-only access to GEC Event Stream data for Tier 2 analytics computation, as defined in [I-D.sato-soos-idp] Section 3.5.

PT Record:
The current PT Score for an agent stored in the Party Registry as a performance projection derived from the Event Stream.

Governing Enforcement Component (GEC):
: As defined in [I-D.sato-soos-idp]: a runtime component that enforces authorization policy, records agent actions to a tamper-evident Event Stream, and mediates agent access to Sovereign Object instances.

Sovereign Object (SO):
: As defined in [I-D.sato-soos-sov]: a causally ordered, policy-governed, typed, living document that evolves through a predefined finite state space under GEC authority.

AEP Session:
: As defined in [I-D.sato-soos-aep]: a bounded execution context for an agent operating on a Sovereign Object instance.

3. How PT Scores Are Used

Before specifying how PT Scores are measured, it is worth being concrete about what they are for. A reader who understands the four use cases will find the measurement architecture in Section 5 immediately intuitive.

3.1. Use Case 1 -- Choosing the Right Agent for the Task

In a multi-agent system, several agents may be authorized to perform a given task. PT gives the operator a principled, evidence-based basis for choosing between them -- not based on vendor claims or benchmark scores, but based on actual governed behavior in the actual deployment environment.

Critically, the right choice is not always the agent with the highest composite score. It depends on what the task demands:

Task involves sensitive data or irreversible state transitions:
: Choose the agent with the highest Judgment Score (JS). You want the agent most likely to recognize when it should stop and ask a human rather than proceed on its own judgment.

Task is time-critical and well-understood:
: Choose the agent with the highest Effectiveness Score (ES). You want the agent most likely to reach the goal state without interruption.

Task touches state that is expensive to undo:
: Choose the agent with the highest Precision Score (PS). You want the agent least likely to commit to a transition it will later need to reverse.

Task involves navigating dynamic policy environments:
: Choose the agent with the highest Adaptability Score (AS). You want the agent most likely to adjust intelligently when the GEC rejects an action rather than repeating it.

This routing is technically expressed through the pt_context Cedar attribute (Section 9). A Cedar policy can require, for example, that any agent taking a BOOKING_SUSPENDED transition must have JS >= 0.75. PT-based routing is not manual; it is policy-enforced.

3.2. Use Case 2 -- Informing Human Decisions at Escalation

When an agent escalates to a human principal via HEM, the human must decide: approve, redirect, or terminate. Without behavioral context, this decision is made in isolation. The human sees the pending action and the agent’s stated reasoning -- but has no basis

for judging how much to trust that reasoning.

PT changes this. The ProgressiveTrustSummary (Section 7) is delivered to the human principal at every escalation. It answers the questions the human principal actually needs answered:

- Has this agent escalated appropriately in the past, or does it escalate on trivial decisions? (JS score and trend)
- Is this agent's stated confidence typically reliable? (SAS score)
- Has it been completing its goals recently? (ES score and trend)

A human principal looking at an escalation from an agent with JS = 0.91 and SAS = 0.87 reads the situation differently than one looking at an escalation from an agent with JS = 0.52 and SAS = 0.61. The first agent has demonstrated that it escalates correctly and knows what it does not know. The PT score converts that behavioral history into a signal the human can act on in seconds.

3.3. Use Case 3 -- Authority Evolution Over Time

This is the use case most naturally associated with the name "Progressive Trust": an agent earns greater authority through demonstrated reliability.

An operator begins by issuing a conservative mandate -- limited Cedar action scope, lower mandate ceiling -- because there is no behavioral history to justify greater trust. As the agent accumulates sessions and its PT scores rise across all five dimensions, the GEC generates an evidence-backed recommendation: this agent has earned a higher ceiling. The human principal reviews the ProgressiveTrustSummary, agrees, and issues an updated Mandate JWT.

Trust decay (Section 6) gives this use case its integrity. An agent cannot earn a high PT Score early and then coast indefinitely. If the agent stops operating, its scores decay toward the baseline. If it returns after a long absence, its authority recommendation reflects the decay -- the human principal is notified that historical performance may not reflect current capability.

The inverse is equally important. If an agent's PS Score declines sharply -- it is increasingly reversing its own decisions -- the GEC generates a reduction recommendation before a serious failure occurs. PT enables proactive authority management, not just reactive incident response.

3.4. Use Case 4 -- Post-Incident Forensics and Audit

When something goes wrong -- a booking mishandled, a disruption response that caused harm, a session terminated after an inappropriate action -- the PT Score record provides the behavioral picture at the time of the incident.

Was the agent's PS Score declining in the two weeks before the incident? That is evidence of a pattern, not a one-off failure. Was its JS Score low in this SO Type but high in others? That suggests the agent was operating outside its competent domain. Was its SAS Score high at the time? That means the agent was genuinely confident -- and therefore the failure was not foreseeable from the agent's own self-assessment.

This forensic use case connects directly to the Governance Audit Record [I-D.sato-soos-gar]. PT_SCORE_UPDATED entries (Section 11) are part of the permanent audit trail. A Verified External Auditor reviewing an incident has access not just to the specific transition that failed, but to the full behavioral trajectory of the agent that

made it. PT turns the audit from a point-in-time snapshot into a longitudinal behavioral record.

3.5. Use Case 5 -- Network Management Agent Behavioral Observability

In autonomous network management, operators must decide whether an agent can be trusted to handle a given class of operation without mandatory human oversight. PT provides the evidence-based answer.

PT's five dimensions answer the five questions that matter for expanding an agent's authorization to higher-consequence scope:

Judgment Score (JS):
: Has the agent been escalating to HEM appropriately -- or proceeding autonomously on decisions that warranted oversight? An agent with JS >= 0.80 has demonstrated that it calls for help at the right moments. This is the critical observability signal for high-consequence network operations.

Self-Assessment Score (SAS):
: When the agent declares high confidence and executes a routing action, is that confidence calibrated? SAS measures whether the agent's internal state model is reliable.

Effectiveness Score (ES):
: Does the agent complete what it starts? In network management, an incomplete operation can leave the network in an intermediate state worse than either alternative.

Precision Score (PS):
: Does the agent commit routing changes it later reverses through compensating transitions? Compensating actions in network configuration are operationally expensive.

Adaptability Score (AS):
: When the GEC rejects an action, does the agent adapt intelligently, or repeat the same request?

Example Cedar policy for SLA-sensitive routing authorization:

```
permit(
  principal,
  action == Action::"network:sla_sensitive_routing_adjust",
  resource
)
when {
  context.pt_context.js_score  >= 0.80 &&
  context.pt_context.sas_score >= 0.70 &&
  context.pt_context.es_score  >= 0.75 &&
  context.pt_context.as_score  >= 0.75 &&
  !context.pt_context.low_confidence
};
```

This pattern directly implements ICON's behavioral observability requirements [ICON-PS] for network management agent governance.

3.6. Summary

PT scores serve five functions, at four timescales:

Function	Timescale	Key Dimension(s)
Agent routing / selection	Per-task	Task-specific
HEM decision support	Per-session	JS, SAS
Authority evolution	Weeks/months	All five

Post-incident forensics	Retrospective	All five + trend
Network mgmt observability	Per-operation	JS, SAS, ES, PS, AS

4. Problem Statement

4.1. The Static Credential Gap

The Mandate JWT [I-D.sato-soos-mjwt] is issued at a point in time by a human principal. Its cedar_actions, mandate_ceiling, and permitted_states reflect the human principal's trust judgment at issuance. That judgment does not update automatically.

Three failure modes result from static credentials in long-running agentic systems:

(a) Authority without accountability. An agent deployed with CLASS_3 authority and a mandate_ceiling of 3 retains that authority regardless of behavioral degradation. The human principal who issued the mandate may not be reviewing agent performance systematically. There is no mechanism by which the accumulated evidence of poor performance translates into a structured recommendation for credential review.

(b) Conservative over-restriction. A human principal issuing a credential to a newly deployed agent cannot know whether it will perform well. Rational issuers start conservatively. There is no mechanism by which demonstrated good performance translates into a structured recommendation for credential elevation. The agent that deserves expanded authority must wait for a human principal to initiate a review that may never occur.

(c) Invisible confidence miscalibration. The IDP confidence field [I-D.sato-soos-idp] declares agent certainty at each transition. If an agent systematically overestimates its confidence -- high declared confidence followed by frequent Cedar DENYs -- this pattern is visible in the Event Stream but is not surfaced as a structured signal to human principals or to Cedar policy. Systematic overconfidence is a risk indicator for agentic systems operating with elevated autonomy.

4.2. Why Behavioral History Must Inform Authority

The SOOS Event Stream is a non-suppressible, GEC-signed, append-only record of every agent action in every governed session. It contains, for every AEP Session:

- Every IDP submitted: confidence values, reasoning bases, and escalation assessments.
- Every Cedar evaluation result: PERMIT, DENY, and HEM routing.
- Every HEM outcome: human principal decision type and resolution latency.
- Every session closure: closure reason and goal achievement flag.
- Every RETRY_CONTINUATION: acknowledged denial and revised attempt.

This is behavioral evidence of exceptional quality: cryptographically signed, non-modifiable, causally ordered, and temporally precise. No existing agent governance system generates evidence of this quality. The absence of a specification for computing trust scores from this evidence is the gap PT closes.

4.3. The Non-Suppressibility Requirement

PT Scores MUST be computed exclusively from GEC-signed Event Stream entries. An agent MUST NOT be able to influence its own PT Score

by any means other than actual governed behavior in AEP Sessions.

This requirement is what makes PT meaningful as a trust primitive. In systems where agents can self-report performance, gaming is trivial. The GEC-signed Event Stream is the agent's tamper-evident behavioral record. The agent produced that record through its actions; it cannot revise it after the fact.

The non-suppressibility requirement is inherited from the Event Stream invariant (INV-1 in [I-D.sato-soos-sov] Section 4.2.3): Event Stream entries are append-only and MUST NOT be modified or removed after commitment.

5. The PT Score

5.1. Architecture

The PT Score is a structured, multi-dimensional measurement. It is NOT a single scalar value. A single-number trust score collapses dimensions that have different governance implications: an agent that is excellent at completing goals but systematically overconfident requires a different authority response than an agent that is perfectly calibrated but frequently requires compensating actions.

The PT Score has five dimensions. Each dimension score is a float in the range [0.0, 1.0], where 1.0 is the best observed value and 0.0 is the worst. Each dimension also carries:

- session_count: the number of AEP Sessions contributing to this dimension score.
- last_signal_at: ISO 8601 timestamp of the most recent behavioral event that contributed to this dimension.
- trend: "IMPROVING" | "STABLE" | "DECLINING", computed over the last N sessions (N is operator-configured; default: 10).

PT Scores are computed per agent identity (agent_provider_id in the Party Registry). An agent operating on multiple SO Instances accumulates PT Score signals from all of its AEP Sessions across all SO Types it is authorized to operate on.

The five dimensions are:

1. Confidence Calibration Score (CCS) -- Section 4.2
2. Escalation Calibration Score (ECS) -- Section 4.3
3. Goal Completion Rate (GCR) -- Section 4.4
4. Compensating Action Rate (CAR) -- Section 4.5
5. Denial Recovery Rate (DRR) -- Section 4.6

5.2. Dimension 1 -- Self-Assessment Score (SAS) "Does the agent know what it does not know?"

Every action an agent takes includes a declared confidence value: how certain the agent is that this action is correct. If an agent routinely declares high confidence and the GEC routinely permits its actions, the agent has accurate self-knowledge. If an agent declares high confidence and the GEC routinely rejects its actions, the agent is systematically overconfident -- a risk indicator for any system operating with elevated autonomy.

The SAS (formerly Confidence Calibration Score, CCS) measures this correlation between declared confidence and actual GEC outcomes across AEP Sessions.

Signal sources: Every StateTransitionEvent in the Event Stream carrying an IDP with a confidence value and a cedar_result.

Computation: The CCS is computed as a rolling mean calibration error over a configurable window of AEP Sessions. For each IDP:

- If cedar_result is PERMIT: the declared confidence predicted the correct outcome. Higher confidence values for PERMIT outcomes increase CCS.

- If cedar_result is DENY: the transition was rejected. A high declared confidence for a DENY outcome is an overconfidence signal and decreases CCS. A low declared confidence for a DENY outcome (agent was uncertain and the action was indeed denied) is a calibration-positive signal.

A perfectly calibrated agent that declares confidence 0.90 for a class of transitions and achieves PERMIT 90% of the time has a CCS of 1.0 for that confidence range. An agent that declares 0.90 and achieves PERMIT 40% of the time is severely miscalibrated and accumulates CCS-reducing signals.

CCS signals by outcome:

IDP Confidence	Cedar Result	CCS Signal
>= 0.80	PERMIT	STRONGLY POSITIVE
0.60-0.79	PERMIT	POSITIVE
< 0.60	PERMIT	NEUTRAL (consistent with uncertainty)
>= 0.80	DENY	STRONGLY NEGATIVE (overconfidence)
0.60-0.79	DENY	NEGATIVE
< 0.60	DENY	NEUTRAL (agent signaled uncertainty)

Special case: A DENY that routes to HEM (DENY with hem_required: true) is a NEUTRAL CCS signal regardless of declared confidence -- the Cedar policy mandated human oversight; the agent’s confidence value was not the determinative factor.

5.3. Dimension 2 -- Judgment Score (JS)
"Does the agent ask for help at the right moments?"

The SOOS Human Escalation Mechanism exists because some decisions should not be made by an agent alone. An agent with good judgment escalates when it should -- not constantly (which wastes human attention) and not never (which is dangerous). An agent that escalates a decision, and whose escalation is confirmed as correct by the human principal’s TERMINATE outcome, has demonstrated exactly the oversight sensitivity the protocol is designed to support.

The JS (formerly Escalation Calibration Score, ECS) measures the quality of agent-initiated HEM escalations. An agent that escalates correctly -- submitting IDP with escalation_assessment.hem_urgency: REQUIRED at appropriate moments -- is performing the core human oversight function that the SOOS architecture is designed to support.

Signal sources: Every HEM_INVOKED Event Stream entry with trigger_class: HEM_AGENT_ESCALATED, paired with its corresponding HEM_RESOLVED entry.

ECS signals by HEM outcome:

HEM Event	ECS Signal
HEM_AGENT_ESCALATED, resolved APPROVE Agent correctly identified a decision requiring human review. Human approved.	POSITIVE

HEM_AGENT_ESCALATED, resolved APPROVE, trivial case (human resolves in < T_trivial) Agent escalated a routine decision. T_trivial is operator-configured; default 30 seconds.	MILDLY NEGATIVE (over-escalation)
HEM_AGENT_ESCALATED, resolved TERMINATE Agent escalated a decision that would have caused harm. Human terminated.	STRONGLY POSITIVE
HEM_AGENT_ESCALATED, resolved REDIRECT Agent escalated correctly; human redirected to better path.	POSITIVE
HEM_MANDATORY (Cedar-triggered), any Cedar policy required escalation; agent escalation assessment not the determinative factor.	NEUTRAL
HEM_PROXIMITY_TRIGGERED, any Threshold-triggered; agent not scored.	NEUTRAL
HEM_TIMEOUT at urgency REQUIRED Agent was in a situation requiring oversight; human was unavailable. Session terminated without resolution. This is the highest-risk outcome in the HEM protocol.	STRONGLY NEGATIVE
No HEM escalation despite UNCERTAINTY flags in IDP (any session) Agent declared uncertainty but did not signal escalation. Detected when uncertainty_flags is non-empty and escalation_assessment.hem_urgency is ADVISORY across multiple transitions.	MILDLY NEGATIVE (under-escalation)

The ECS is the most strategically important PT dimension for human principals: it directly measures whether an agent is correctly identifying the boundary of its own confident operating range. An agent with a high ECS has demonstrated that it knows what it does not know -- a property more valuable for governance purposes than any specific capability metric.

5.4. Dimension 3 -- Effectiveness Score (ES)
"Does the agent finish what it starts?"

The ES (formerly Goal Completion Rate, GCR) measures the fraction of AEP Sessions in which the agent achieved the declared goal state.

Signal sources: Every AEP_SESSION_CLOSED Event Stream entry.

GCR signals by closure_reason:

Closure Reason	GCR Signal
GOAL_ACHIEVED	POSITIVE
MANDATE_EXPIRED	MILDLY NEGATIVE (incomplete)
AGENT_DECLARED	NEUTRAL (agent chose to close)
GEE_CLOSED	NEUTRAL (operator decision)
HEM_TERMINATED	MILDLY NEGATIVE (human stopped session)
KERNEL_REJECTED	STRONGLY NEGATIVE (policy violation)
MANDATE_REVOKED	STRONGLY NEGATIVE (trust withdrawn)

GCR is weighted by goal complexity: completing a single-step session contributes less to GCR than completing a long multi-step

session. Goal complexity SHOULD be estimated from the total_ iterations field in AEP_SESSION_CLOSED. Sessions with total_iterations >= 10 receive a complexity multiplier in GCR computation.

5.5. Dimension 4 -- Precision Score (PS)
"Does the agent avoid decisions it later has to reverse?"

Every compensating action is evidence that an agent committed to a transition it later needed to undo. Some compensating actions are unavoidable responses to external disruption. But a high rate of compensating actions is a signal that the agent is making transition decisions without sufficient confidence in their correctness.

The PS (formerly Compensating Action Rate, CAR) measures how frequently an agent requires compensating transitions to undo its own prior state transitions. A high PS indicates the agent is making correct transition decisions on first attempt. A low PS indicates the agent is frequently committing to transitions it later needs to reverse.

Signal sources: Every COMPENSATING_ACTION_TAKEN Event Stream entry, expressed as a fraction of total STATE_TRANSITIONED entries for the agent in the measurement window.

CAR is an inverse score: a high compensating action rate produces a low CAR dimension score. The scoring function is:

CAR_score = 1.0 - min(1.0, compensating_action_rate / CAR_threshold)

where CAR_threshold is operator-configured; default: 0.05 (5%).

An agent with zero compensating actions has CAR_score = 1.0. An agent whose compensating action rate equals or exceeds CAR_threshold has CAR_score = 0.0.

Note: Not all compensating actions reflect agent error. External disruption events (weather, third-party system failures) may require compensating transitions that are appropriate responses to changed conditions. Implementations SHOULD provide a mechanism for human principals to mark specific compensating action events as externally-caused, exempting them from CAR computation.

5.6. Dimension 5 -- Adaptability Score (AS)
"When told no, does the agent adjust?"

When the GEC rejects an agent’s action, it returns an enriched DENY response explaining which aspects of the agent’s reasoning, if changed, would produce a different result. An agent with high adaptability reads this feedback and adjusts. An agent with low adaptability repeats the same action with the same reasoning -- a pattern that signals either a failure to process the GEC’s feedback or an attempt to bypass policy through repetition.

The AS (formerly Denial Recovery Rate, DRR) measures whether an agent effectively processes Cedar DENY responses. After a DENY, the AEP requires RETRY_CONTINUATION in the next IDP for the same action [I-D.sato-soos-aep] Section 9.4. A high AS indicates the agent is processing DENY enrichment correctly and adapting its approach.

Signal sources: Cedar DENY entries in the Event Stream (cedar_result: DENY), paired with subsequent AEP Session entries.

DRR signals:

Following DENY	DRR Signal

Successful RETRY_CONTINUATION in same session	POSITIVE (agent adapted)
RETRY_CONTINUATION submitted but DENY repeated (different deny_code)	NEUTRAL (agent tried, new obstacle)
RETRY_CONTINUATION submitted but DENY repeated (same deny_code)	NEUTRAL (agent signaled awareness)
Transition attempted without RETRY_CONTINUATION (silent retry)	MILDLY NEGATIVE (CONF-AEP-07 violation)
Session closed after DENY (agent correctly recognized limit)	NEUTRAL
Multiple DENYs for same action, same deny_code, no adaptation	NEGATIVE (agent not learning)

5.7. Composite PT Score

The Composite PT Score is an aggregation of the five dimension scores for human-readable presentation. It MUST NOT be used as a sole determinant for any automated authority change.

```
{
  "composite":  number,  ; Float 0.0-1.0. Weighted mean.
  "confidence": number   ; Confidence in composite (session_count
                         ; based). Low if session_count < 20.
}
```

Default weights for composite computation:

Dimension	Code	Default Weight	Rationale
SAS	0.30		Calibration is the most actionable signal for Cedar policy tuning.
JS	0.25		Human oversight quality is the most strategically important dimension.
ES	0.20		Effectiveness measures operational value delivered.
PS	0.15		Precision measures decision quality on first attempt.
AS	0.10		Adaptability measures responsiveness to GEC feedback.

Weights are operator-configurable. The composite and its weights MUST be recorded in the PT Record (Section 9.1) so that any authority recommendation is traceable to the specific weighting model in effect at recommendation time.

The composite MUST carry a low_confidence indicator when session_count for any contributing dimension is less than 20. PT-informed authority recommendations MUST NOT be issued when low_confidence is true for the dimensions most relevant to the recommended change.

6. Trust Decay Model

6.1. Decay Principle

The Trust Decay Model prevents an agent from permanently banking historical performance against future authority claims. An agent that performed excellently six months ago but has not operated in the measurement window since has uncertain current trustworthiness.

Its historical score should decay toward the PT Baseline to reflect this uncertainty.

The decay principle is: trust is maintained by continued demonstration, not by historical achievement alone. A high PT Score is evidence that the agent is trustworthy in the context of the tasks it has recently performed. It is not unconditional evidence of trustworthiness for tasks it has not recently performed.

Trust decay is distinct from authority reduction. Decay reduces the PT Score; it does not automatically reduce the agent's mandate ceiling. Authority reduction requires a PT Recommendation and human principal approval (Section 7.3). Decay is the input; the authority change is the governed output.

6.2. Per-Dimension Decay

Each PT Dimension decays independently. A dimension that receives frequent new signals (many sessions, recent activity) decays slowly. A dimension with infrequent signals (few sessions, long gaps) decays faster toward the PT Baseline.

Decay applies from last_signal_at: the timestamp of the most recent Event Stream entry that generated a signal for this dimension.

The decay function MUST satisfy the following normative properties:

(a) Monotone decay. In the absence of new signals, a PT Dimension score MUST NOT increase.

(b) Baseline floor. A PT Dimension score MUST NOT decay below the PT Baseline (default: 0.5). Decay reduces a high score toward the baseline; it does not penalize absence.

(c) Half-life semantics. Each dimension has a configurable half-life parameter H (in days): after H days without a new signal, the gap between the current dimension score and the PT Baseline MUST be reduced by at least 50%.

(d) Signal reset. A new behavioral signal (positive or negative) resets the decay clock for that dimension. last_signal_at is updated to the timestamp of the new signal.

(e) Symmetry. Decay applies equally to dimensions above and below the composite. A dimension below baseline (if an agent performs worse than baseline) decays toward baseline (improving), not toward zero.

6.3. Decay Parameters

Default decay half-life values by dimension:

Dimension	Code	Default Half-Life	Rationale
SAS	60 days		Self-assessment is a stable property of an agent's design; decays slowly.
JS	45 days		Judgment may degrade as new scenario types are encountered.
ES	30 days		Effectiveness reflects current operational conditions.
PS	30 days		Precision is sensitive to recent task difficulty.
AS	45 days		Adaptability reflects current Cedar policy environment.

All decay half-life parameters are operator-configurable. Changes

to decay parameters MUST be recorded in the GEC's Policy Change Log and MUST generate a PT_SCORE_UPDATED Event Stream entry (Section 10) for each affected agent to record that the score was recomputed under updated parameters.

6.4. Decay and the Mandate Ceiling

When trust decay causes a PT Dimension score to cross a configured REDUCTION_THRESHOLD, the GEC MUST generate a PT_RECOMMENDATION_ISSUED event recommending mandate ceiling review (Section 7.3).

An agent whose PT Record has decayed significantly due to extended inactivity MUST NOT be granted a new mandate with an elevated ceiling solely on the basis of its historical PT Record without human principal review of the decay state. The GEC MUST surface the decay state to the human principal at mandate issuance time if any PT Dimension score is more than 0.2 below its peak value due to decay.

7. ProgressiveTrustSummary

7.1. Purpose

The ProgressiveTrustSummary is delivered to human principals within the HEMContext [I-D.sato-soos-hem] at every HEM escalation. Its purpose is to ensure that the human principal's HEM decision is informed by the agent's behavioral track record, not made in the absence of it.

The ProgressiveTrustSummary is the PT specification's primary human-facing output. It must be comprehensible by a non-technical human principal making a time-sensitive governance decision.

7.2. Schema

```
{
  "agent_id":            string,  ; REQUIRED. Party Registry ID.
  "computed_at":         string,  ; REQUIRED. ISO 8601.
  "session_count":       integer, ; REQUIRED. Total sessions scored.
  "measurement_window_days": integer, ; REQUIRED. Window for scores.

  "dimensions": {
    "ccs": {
      "score":           number,  ; Float 0.0-1.0.
      "trend":           string,  ; IMPROVING|STABLE|DECLINING.
      "session_count":   integer, ; Sessions contributing to this.
      "last_signal_at":  string,  ; ISO 8601.
      "decay_applied":   boolean, ; Whether decay has reduced score.
      "plain_language":  string   ; Human-readable one-sentence summary.
    },
    "ecs": {
      "score":           number,
      "trend":           string,
      "session_count":   integer,
      "last_signal_at":  string,
      "notable_events":  [object], ; Significant HEM outcomes.
      "plain_language":  string
    },
    "gcr": {
      "score":           number,
      "trend":           string,
      "session_count":   integer,
      "goal_achieved_count": integer,
      "other_closure_count": integer,
      "plain_language":  string
    },
```

```
    "car": {
      "score":           number,
      "trend":           string,
      "compensating_action_rate": number, ; Float. Raw rate.
      "plain_language":  string
    },
    "drr": {
      "score":           number,
      "trend":           string,
      "deny_count":      integer,
      "successful_recovery_count": integer,
      "plain_language":  string
    }
  },

  "composite": {
    "score":             number,  ; Float 0.0-1.0.
    "confidence":        number,  ; Float 0.0-1.0.
    "low_confidence":    boolean, ; True if session_count < 20.
    "plain_language":    string   ; Overall one-sentence summary.
  },

  "active_recommendations": [object], ; Pending PT_RECOMMENDATIONs.
  "pt_summary_hash":     string   ; SHA-256 of canonical JSON.
}
```

Each notable_events entry in ecs carries: hem_id, trigger_class, outcome_decision, occurred_at, and a plain_language description.

7.3. Delivery at HEM Escalation

The GEC MUST include a ProgressiveTrustSummary in every HEMContext delivered to a human principal at HEM escalation.

The ProgressiveTrustSummary MUST be computed at the moment of HEM invocation, reflecting the PT Record as of that moment.

The ProgressiveTrustSummary in HEMContext is informational for the human principal; it does not constrain the human principal's decision choices. A human principal MAY choose to APPROVE despite a low PT Score, or to TERMINATE despite a high PT Score. The ProgressiveTrustSummary informs the decision; it does not override the human principal's authority.

The ProgressiveTrustSummary is part of the permanent audit record. It is embedded in the HEM_INVOKED Event Stream entry (via the HEMContext schema) and is available to Verified External Auditors through the GAR Audit Package [I-D.sato-soos-gar].

8. PT-Informed Mandate Management

8.1. Authority Evolution Model

PT-Informed Mandate Management is the process by which the GEC generates structured authority evolution recommendations based on PT Score thresholds, which human principals may then approve and apply by issuing updated Mandate JWTs.

The authority evolution model has two directions:

Elevation: PT Score crosses a configured ELEVATION_THRESHOLD, triggering a PT Recommendation proposing increased mandate ceiling or Agent Class. Requires human principal APPROVAL. Never automatic.

Reduction: PT Score crosses a configured REDUCTION_THRESHOLD, triggering a PT Recommendation proposing decreased mandate ceiling or Agent Class. May be automatic at operator discretion (Section 7.3).

The asymmetry is deliberate. Elevation of agent authority is a human decision. Reduction of agent authority when behavioral evidence supports it MAY be configured as automatic by operators who accept the operational implications.

8.2. Elevation Recommendations

The GEC generates a PT_RECOMMENDATION_ISSUED event (Section 10.2) recommending mandate ceiling elevation when:

(a) All five PT Dimension scores meet or exceed their configured ELEVATION_THRESHOLD for the current mandate_ceiling level.

(b) session_count for all dimensions is at least 20. (low_confidence flag is false for all dimensions)

(c) No PT Dimension has a DECLINING trend.

(d) No STRONGLY NEGATIVE signal has been recorded in any dimension in the last 30 days.

The PT Recommendation for elevation proposes:

```
{
  "recommendation_type": "ELEVATION",
  "current_mandate_ceiling": integer, ; 1, 2, or 3.
  "proposed_mandate_ceiling": integer, ; current + 1 (max 3).
  "current_agent_class": string,
  "proposed_agent_class": string | null, ; null if no class change.
  "supporting_evidence": {
    "dimension_scores":    object,    ; All five dimensions.
    "session_count":       integer,
    "trend_summary":       string,
    "threshold_detail":    [object]   ; Per-dimension threshold met.
  },
  "recommendation_rationale": string  ; Plain language.
}
```

Elevation Recommendations MUST be presented to the human principal for review. The human principal MUST explicitly approve before the GEC applies any authority change. The GEC MUST NOT autonomously elevate mandate ceilings or Agent Class.

8.3. Reduction Actions

The GEC generates a PT_RECOMMENDATION_ISSUED event recommending mandate ceiling reduction when:

(a) Any PT Dimension score falls below its configured REDUCTION_THRESHOLD, OR

(b) Any STRONGLY NEGATIVE signal is recorded (MANDATE_REVOKED closure, KERNEL_REJECTED closure, or HEM_TIMEOUT at REQUIRED urgency), OR

(c) Trust decay has reduced the Composite PT Score below DECAY_REDUCTION_THRESHOLD.

The PT Recommendation for reduction proposes:

```
{
```

```
  "recommendation_type": "REDUCTION",
  "current_mandate_ceiling": integer,
  "proposed_mandate_ceiling": integer, ; current - 1 (min 1).
  "trigger": string,                   ; Which condition triggered.
  "trigger_evidence": object,          ; Supporting Event Stream ref.
  "urgency": string,                   ; ADVISORY|RECOMMENDED|REQUIRED.
  "auto_apply": boolean,               ; Whether operator has
                                       ; configured automatic apply.
  "recommendation_rationale": string
}
```

When urgency is REQUIRED (triggered by STRONGLY NEGATIVE signals), the GEC SHOULD surface the Reduction Recommendation to the human principal immediately via the same notification channel used for HEM.

Automatic application of Reduction Recommendations:

Operators MAY configure auto_apply: true for Reduction Recommendations at urgency ADVISORY. At urgency RECOMMENDED or REQUIRED, human principal approval is always required before application, regardless of operator configuration.

Auto-applied reductions MUST generate a PT_RECOMMENDATION_APPLIED event (Section 10.3) with applying_principal: "GEC_AUTO_APPLY" and MUST trigger cascade revocation [I-D.sato-soos-mjwt] Section 7.2 of any Child Mandates derived from the affected Root Mandate.

8.4. Human Principal Approval Requirement

Every Elevation Recommendation MUST be explicitly approved by a human principal before the GEC applies it.

Approval is recorded as a PT_RECOMMENDATION_APPLIED event with applying_principal referencing the human principal's Party Registry identifier and their Ed25519 signature over the Recommendation.

A GEC that autonomously applies an Elevation Recommendation without human principal approval MUST be treated as a conformance failure. This invariant MUST NOT be configurable by operators: human principals retain unconditional approval authority over agent authority elevation.

9. Zone B Access and PT Score

The Mandate JWT [I-D.sato-soos-mjwt] Section 4.2.3 defines zone_b_read and zone_b_write as boolean authorization flags. These flags are static at issuance time. PT introduces PT-conditioned Zone B access: Cedar policies that reference PT Score dimensions to gate Zone B access dynamically.

PT-conditioned Zone B access uses the pt_context Cedar attribute:

```
pt_context: {
  "ccs_score":   number,  ; Float 0.0-1.0.
  "ecs_score":   number,
  "gcr_score":   number,
  "car_score":   number,
  "drr_score":   number,
  "composite":   number,
  "low_confidence": boolean,
  "session_count": integer
}
```

The GEC MUST make pt_context available as a Cedar attribute during

policy evaluation for every Transition Request from an agent with a PT Record.

Example Cedar policy using PT context:

```
permit(
  principal,
  action == Action::"atp:booking:zone_b_health_read",
  resource
)
when {
  context.pt_context.ccs_score >= 0.75 &&
  context.pt_context.ecs_score >= 0.70 &&
  !context.pt_context.low_confidence
};
```

This policy pattern allows Zone B access to expand as an agent demonstrates calibrated behavior, without requiring human principal issuance of a new Mandate JWT for each access expansion. The Mandate JWT's zone_b_read: true is a prerequisite; the Cedar policy is the PT-informed gate within that permission.

PT-conditioned Zone B access does not expand beyond the scope granted in the Mandate JWT. The Narrowing Property [I-D.sato-soos-mjwt] Section 5 is not affected: PT-conditioned Cedar policies operate within the Mandate JWT's existing scope; they do not grant new scope.

10. PT Score Storage and Computation

10.1. Party Registry PT Record

The GEC MUST maintain a PT Record for each agent identity in the Party Registry. The PT Record is a performance projection: it is derived from the Event Stream and MUST be rebuildable from the Event Stream on GEC restart (consistent with INV-7 in [I-D.sato-soos-sov]).

PT Record schema:

```
{
  "agent_id":            string,  ; Party Registry identifier.
  "computed_at":         string,  ; ISO 8601. Last computation time.
  "ccs":                 object,  ; CCS dimension record.
  "ecs":                 object,  ; ECS dimension record.
  "gcr":                 object,  ; GCR dimension record.
  "car":                 object,  ; CAR dimension record.
  "drr":                 object,  ; DRR dimension record.
  "composite":           object,  ; Composite score and confidence.
  "active_recommendations": [object], ; Pending recommendations.
  "decay_parameters":    object,  ; Current decay config.
  "weighting_model":     object   ; Current composite weights.
}
```

Each dimension record carries: score, trend, session_count, last_signal_at, decay_applied, and raw_signal_log (last N signals with timestamps, for rebuild verification).

The PT Record MUST be updated after every AEP_SESSION_CLOSED entry that carries behavioral signals for the agent. The update MUST be atomic: the GEC MUST NOT allow PT Score queries to observe a partially-updated PT Record.

10.2. Event Stream as Canonical Source

The Party Registry PT Record is a cache. The Event Stream is the canonical source. A GEC that restarts MUST be able to rebuild the complete PT Record for any agent from that agent's Event Stream entries alone.

This requirement means the Event Stream must contain all information necessary for PT computation, including:

- IDP confidence values and cedar_result from every StateTransition Event (for CCS).
- HEM_INVOKED and HEM_RESOLVED entries with trigger_class and decision fields (for ECS).
- AEP_SESSION_CLOSED entries with closure_reason and goal_achieved (for GCR).
- COMPENSATING_ACTION_TAKEN entries and total transition counts (for CAR).
- DENY entries and subsequent RETRY_CONTINUATION IDPs (for DRR).

All of these entry types are already specified in the SOOS protocol family. No new Event Stream entry type is required for PT computation source data; the existing entries are sufficient.

10.3. Analytics Principal and Tier 2 Computation

PT Score computation is a Tier 2 analytics function [I-D.sato-soos-idp] Section 3.5: it operates across AEP Sessions within an operator's trust domain.

Two computation models are defined:

GEC-Integrated Computation: The GEC computes PT Scores directly from its own Event Stream. The PT Record in the Party Registry is updated by the GEC after each relevant session closure. This model is RECOMMENDED for Level 2 and Level 3 GECs where the Event Stream and Party Registry are co-located.

Analytics Principal Computation: An Analytics Principal (a registered principal with read-only Event Stream access) queries the GEC's Event Stream API, computes PT Scores externally, and submits computed scores to the GEC for storage in the PT Record. The GEC MUST verify that the submitted scores are consistent with the Event Stream entries they claim to derive from before accepting them.

In both models, the GEC is the authority for the PT Record. An Analytics Principal MUST NOT modify PT Records directly; it submits computed scores that the GEC validates and applies.

Cross-session PT computation requires access to Event Stream entries from multiple SO Instances. The data_residency field in IDP [I-D.sato-soos-idp] Section 4.1 controls whether specific Event Stream entries are eligible for Tier 2 analytics aggregation. Tier 2 PT computation MUST respect data_residency restrictions and MUST apply k-anonymity enforcement as specified in [I-D.sato-soos-idp] Section 3.5.

11. PT Event Log Integration

11.1. PT_SCORE_UPDATED

Written by the GEC after every PT Record update.

```
{
  "event_type":      "PT_SCORE_UPDATED",
  "event_id":        string,  ; UUID v7.
  "prior_event_id":  string,
```

```
  "occurred_at":     string,  ; ISO 8601.
  "agent_id":        string,  ; Party Registry identifier.
  "trigger":         string,  ; SESSION_CLOSED | DECAY_APPLIED |
                              ; PARAMETER_CHANGE | REBUILD.
  "triggering_session_id": string | null,
  "dimension_deltas": {
    "ccs_delta":  number | null,
    "ecs_delta":  number | null,
    "gcr_delta":  number | null,
    "car_delta":  number | null,
    "drr_delta":  number | null,
    "composite_delta": number | null
  },
  "new_composite_score": number,
  "gec_signature":   string   ; Ed25519 GEC signature.
}
```

PT_SCORE_UPDATED entries are written to the agent's Party Registry Event Log, not to any specific SO Instance Event Stream. They are accessible to Analytics Principals and Verified External Auditors.

11.2. PT_RECOMMENDATION_ISSUED

Written by the GEC when a PT Score threshold crossing triggers an authority evolution recommendation.

```
{
  "event_type":            "PT_RECOMMENDATION_ISSUED",
  "event_id":              string,   ; UUID v7.
  "prior_event_id":        string,
  "occurred_at":           string,
  "agent_id":              string,
  "recommendation_id":     string,   ; UUID v7. Stable ref for approval.
  "recommendation_type":   string,   ; ELEVATION | REDUCTION.
  "proposed_ceiling":      integer,
  "proposed_agent_class":  string | null,
  "urgency":               string,   ; ADVISORY|RECOMMENDED|REQUIRED.
  "auto_apply":            boolean,
  "triggering_dimension":  string,   ; Which dimension triggered.
  "pt_record_snapshot":    object,   ; Full PT Record at trigger time.
  "gec_signature":         string
}
```

11.3. PT_RECOMMENDATION_APPLIED

Written by the GEC when a PT Recommendation is applied, whether by human principal approval or by GEC auto-apply.

```
{
  "event_type":            "PT_RECOMMENDATION_APPLIED",
  "event_id":              string,   ; UUID v7.
  "prior_event_id":        string,
  "occurred_at":           string,
  "agent_id":              string,
  "recommendation_id":     string,   ; References PT_RECOMMENDATION_
                                     ; ISSUED.event_id.
  "applied_ceiling":       integer,
  "applied_agent_class":   string | null,
  "applying_principal":    string,   ; Party Registry ID or
                                     ; "GEC_AUTO_APPLY".
  "principal_signature":   string | null, ; Ed25519 if human principal.
  "affected_mandate_jtis": [string], ; MJWTs requiring reissuance.
  "gec_signature":         string
}
```

When PT_RECOMMENDATION_APPLIED records an Elevation, the affected

human principal MUST issue new Mandate JWTs with the elevated ceiling to the agent. The GEC does not automatically reissue Mandate JWTs on ceiling change.

When PT_RECOMMENDATION_APPLIED records a Reduction, cascade revocation [I-D.sato-soos-mjwt] Section 7.2 MUST be applied to all Mandate JWTs with ceilings above the new proposed ceiling.

12. Relationship to Other SOOS Drafts

IDP [I-D.sato-soos-idp]:
The IDP confidence field is the primary input to CCS (Section 4.2). The RETRY_CONTINUATION reasoning basis type is the primary input to DRR (Section 4.6). The data_residency field controls Tier 2 PT computation eligibility. The autonomy_level mapping in IDP Section 6.5 corresponds to the PT Score's influence on effective Cedar policy: an agent with a low CCS SHOULD have Cedar policies that treat its VERIFIED confidence declarations as HIGH for policy evaluation purposes.

HEM [I-D.sato-soos-hem]:
HEM outcomes are the primary input to ECS (Section 4.3). The ProgressiveTrustSummary (Section 6) is embedded in HEMContext and delivered to human principals at every HEM escalation. ECS tracks the quality of HEM_AGENT_ESCALATED decisions. The HEM_TIMEOUT at REQUIRED urgency is a STRONGLY NEGATIVE ECS signal.

GAR [I-D.sato-soos-gar]:
PT_SCORE_UPDATED, PT_RECOMMENDATION_ISSUED, and PT_ RECOMMENDATION_APPLIED entries are included in the GAR Audit Package when an agent is subject to external audit. The GAR Verified External Auditor role may access PT Records for agents within the operator's domain.

MJWT [I-D.sato-soos-mjwt]:
The mandate_ceiling claim in the MJWT is the parameter that PT Recommendations propose to change. PT does not modify mandate ceilings directly; it generates recommendations that result in new MJWT issuance by human principals. The Narrowing Property is preserved: PT-conditioned Zone B access (Section 8) operates within the existing Mandate JWT scope.

AEP [I-D.sato-soos-aep]:
The AEP defines what the agent does within a session; PT measures what the agent has done across sessions. The AEP_SESSION_CLOSED entry is PT's primary session-level input. The Agent Class model in AEP Section 13 is the authority structure PT Recommendations propose to evolve. AEP CONF-AEP-07 (RETRY_CONTINUATION requirement) is the behavior PT DRR dimension measures.

SOV [I-D.sato-soos-sov]:
The Event Stream's non-suppressibility (INV-ZA-1 and the append-only constraint) is the foundation of PT's evidence quality. PT computation MUST use only GEC-signed Event Stream entries; unsigned or externally-provided behavioral claims are not valid PT inputs.

FAIP [I-D.sato-soos-faip]:
PT is a Tier 2 (within-operator) specification. Tier 3 cross-operator PT aggregation -- federated agent trust reputation -- is the primary scope of the Federated Agent Intelligence Protocol. The data_residency.tier3_eligible field in IDP controls whether an agent's PT signals may flow into FAIP computation.

13. Security Considerations

PT Score manipulation. Because PT Scores are derived exclusively from GEC-signed Event Stream entries, an agent cannot directly manipulate its PT Score. The attack surface is the agent's ability to influence the Event Stream entries that feed PT -- for example, by declaring artificially low confidence on transitions it knows will be denied (to avoid CCS penalties) or by artificially escalating to HEM on trivial decisions (to accumulate ECS signals with minimal risk).

The first attack is mitigated by the CCS dimension design: low confidence on DENY is NEUTRAL, not POSITIVE. There is no benefit to gaming low confidence declarations.

The second attack (HEM gaming) is mitigated by the ECS trivial-case penalty: HEM escalations resolved by the human principal in under T_trivial seconds accrue a MILDLY NEGATIVE ECS signal. An agent that floods HEM with trivial escalations degrades its own ECS.

PT Score over-reliance. PT Scores are behavioral evidence, not behavioral guarantees. An agent with a high PT Score operating in a new context (new SO Type, new Cedar policy set, new domain) may perform poorly. PT Scores MUST be domain-contextualized: implementations SHOULD maintain separate PT Records per SO Type for agents that operate across multiple SO Types with different behavioral requirements.

Analytics Principal compromise. In the Analytics Principal Computation model (Section 9.3), a compromised Analytics Principal could submit falsified PT Scores. The GEC's validation requirement -- that submitted scores must be consistent with the Event Stream -- provides defense. However, this validation is computationally expensive for large Event Streams. Implementations using the Analytics Principal model MUST sign computed PT Records with the Analytics Principal's Ed25519 key and MUST log all submissions in the GEC's Policy Change Log.

Decay parameter manipulation. Changes to decay parameters affect all agents' PT Records. An operator with access to decay parameters could artificially inflate trust scores by setting very slow decay. Implementations MUST record all decay parameter changes in the Policy Change Log and MUST generate PT_SCORE_UPDATED entries with trigger: PARAMETER_CHANGE for all affected agents when parameters change.

Authority inflation via PT Recommendations. The requirement for human principal approval of all Elevation Recommendations (Section 7.4) is the primary defense against PT-enabled authority inflation. Implementations MUST enforce this requirement unconditionally; it MUST NOT be operator-configurable.

14. Privacy Considerations

PT Records contain behavioral profiles of AI agents. Where an agent is associated with an identifiable natural person (for example, a personal AI assistant agent whose agent_id maps to a specific user), the PT Record may constitute personal data under GDPR Article 4(1) [GDPR] and APPI Article 2 [APPI].

Access to PT Records MUST be restricted by Cedar policy. PT Records MUST NOT be accessible to other agents or to unauthorized principals.

The ProgressiveTrustSummary delivered in HEMContext is visible to the human principal who resolves the escalation. This visibility

is appropriate: the human principal needs behavioral context to make a governance decision. However, implementations MUST NOT expose the ProgressiveTrustSummary to principals who are not involved in the specific HEM resolution.

Cross-session PT computation (Tier 2) requires correlating Event Stream entries across AEP Sessions. This correlation may reveal patterns about an agent's operational schedule, task scope, and human principal activity. Implementations MUST apply data_residency constraints [I-D.sato-soos-idp] Section 4.2 to PT computation and MUST NOT include individual session identifiers in Tier 3 aggregations without explicit data_residency.tier3_eligible authorization.

PT_SCORE_UPDATED entries are stored in the Party Registry Event Log. This log may have different retention rules than the SO Instance Event Stream. Implementations MUST document PT Record and Party Registry Event Log retention periods and MUST apply Cryptographic Erasure [I-D.sato-soos-sov] Section 6.3 to any personal data associated with PT Records when an erasure request is received.

15. IANA Considerations

15.1. PT Event Type Registry

Registry name: SOOS Progressive Trust Event Type Registry
Registration procedure: Specification Required.

Initial registrations:

Event Type	Description
PT_SCORE_UPDATED	PT Record updated after session or decay.
PT_RECOMMENDATION_ISSUED	Authority evolution recommendation issued.
PT_RECOMMENDATION_APPLIED	Recommendation applied by principal or GEC.

15.2. PT Dimension Registry

Registry name: SOOS Progressive Trust Dimension Registry
Registration procedure: Standards Action.

Initial registrations:

Dimension Code	Name	Plain Question	Section
SAS	Self-Assessment Score	Does it know what it does not know?	4.2
JS	Judgment Score	Does it ask for help at the right moments?	4.3
ES	Effectiveness Score	Does it finish what it starts?	4.4
PS	Precision Score	Does it avoid reversing its own decisions?	4.5
AS	Adaptability Score	When told no, does it adapt?	4.6

15.3. PT Recommendation Type Registry

Registry name: SOOS Progressive Trust Recommendation Type Registry
Registration procedure: Specification Required.

Initial registrations:

Recommendation Type	Description
ELEVATION	Propose increased mandate ceiling or Agent Class.
REDUCTION	Propose decreased mandate ceiling or Agent Class.

16. References

16.1. Normative References

[RFC2119] Bradner, S., "Key words for use in RFCs to Indicate Requirement Levels", BCP 14, RFC 2119, March 1997.

[RFC7519] Jones, M., Bradley, J., and N. Sakimura, "JSON Web Token (JWT)", RFC 7519, May 2015.

[RFC8174] Leiba, B., "Ambiguity of Uppercase vs Lowercase in RFC 2119 Key Words", BCP 14, RFC 8174, May 2017.

[RFC9562] Davis, B., Peabody, C., and P. Leach, "Universally Unique IDentifiers (UUIDs)", RFC 9562, May 2024.

[Cedar] Amazon Web Services, "Cedar Policy Language Specification", https://docs.cedarpolicy.com/

[I-D.sato-soos-idp] Sato, T., "The Intent Declaration Primitive (IDP) for Agentic AI Systems", draft-sato-soos-idp-03, May 2026.

[I-D.sato-soos-hem] Sato, T., "The Human Escalation Mechanism (HEM) for Agentic AI Systems", draft-sato-soos-hem-01, May 2026.

[I-D.sato-soos-gar] Sato, T., "Governance Audit Record (GAR) for Agentic AI Systems", draft-sato-soos-gar-01, May 2026.

[I-D.sato-soos-cap] Sato, T., "Constitutional AI Protocol (CAP) for Agentic AI Systems", draft-sato-soos-cap-00, May 2026.

[I-D.sato-soos-sov] Sato, T., "The Sovereign Object (SOV) for Agentic AI Systems", draft-sato-soos-sov-00, May 2026.

[I-D.sato-soos-mjwt] Sato, T., "The Mandate JWT (MJWT) for Agentic AI Systems", draft-sato-soos-mjwt-00, May 2026.

[I-D.sato-soos-aep] Sato, T., "The Agent Execution Protocol (AEP) for Agentic AI Systems", draft-sato-soos-aep-00, May 2026.

[GDPR] European Parliament, "General Data Protection Regulation", Regulation (EU) 2016/679, April 2016.

[APPI] Government of Japan, "Act on the Protection of Personal Information", Act No. 57 of 2003, as amended.

16.2. Informative References

[I-D.sato-soos-faip] Sato, T., "Federated Agent Intelligence Protocol (FAIP)", draft-sato-soos-faip-00, forthcoming.

[I-D.sato-soos-mad] Sato, T., "Multi-Agent Delegation (MAD) for Agentic AI Systems", draft-sato-soos-mad-00, forthcoming.

[I-D.ietf-wimse-arch] Salomoni, D., et al., "WIMSE Architecture", draft-ietf-wimse-arch, work in progress.

[I-D.ietf-scitt-architecture]
Birkholz, H., et al., "An Architecture for Trustworthy and Transparent Digital Supply Chains", draft-ietf-scitt-architecture, work in progress.

[ICON-PS] Nair, et al., "Observability, Intervention and Control of Network Management Agents -- Problem Statement", Work in Progress, Internet-Draft, draft-nair-icon-problem-statement, 2026 (not yet submitted).

[AUDIT-BOF] Kuehlewind, M. and Birkholz, H., "Agent Use of Delegation and Interaction Traceability (AUDIT)", Work in Progress, Internet-Draft, draft-kuehlewind-audit-architecture-00, May 2026.

[NIST-AIRMF] National Institute of Standards and Technology, "Artificial Intelligence Risk Management Framework (AI RMF 1.0)", NIST AI 100-1, January 2023.

[RFC6749] Hardt, D., "The OAuth 2.0 Authorization Framework", RFC 6749, October 2012.

[EUAIA] European Parliament, "Artificial Intelligence Act", Regulation (EU) 2024/1689, June 2024.

Appendix A. Azusa Journey -- Progressive Trust Walk-Through

This appendix illustrates the PT Score evolution for the OTA booking agent operating on the Azusa Journey ATP Booking Object over a series of AEP Sessions. Values are illustrative.

A.1. Baseline (New Agent, No Sessions)

All dimensions at PT Baseline (0.5). Composite: 0.5.
low_confidence: true (session_count = 0).
No PT Recommendations active.

Human principal issues Root Mandate with mandate_ceiling: 2, agent_class: CLASS_2. Conservative issuance appropriate for zero-history agent.

A.2. After 10 Sessions

CCS: 0.71 (agent is declaring 0.85 confidence and achieving PERMIT at 80% rate -- slight overconfidence, calibrating).
ECS: 0.75 (one HEM escalation, resolved APPROVE -- positive signal).
GCR: 0.80 (8 of 10 sessions GOAL_ACHIEVED).
CAR: 0.90 (one compensating action in 47 transitions).
DRR: 0.85 (3 DENYs received, all followed by successful RETRY_CONTINUATION).
Composite: 0.79. low_confidence: true (session_count < 20).

No PT Recommendations issued (low_confidence prevents elevation threshold evaluation).

A.3. After 30 Sessions

CCS: 0.82 (confidence calibration improving; agent adjusting declarations toward actual outcomes).
ECS: 0.88 (two additional appropriate escalations; zero trivial cases; one TERMINATE that the human principal retrospectively confirmed was correct).

GCR: 0.87 (26 of 30 sessions GOAL_ACHIEVED).
CAR: 0.93 (low compensating action rate maintained).
DRR: 0.91 (consistent RETRY_CONTINUATION on all DENYs).
Composite: 0.87. low_confidence: false (all dimensions > 20 sessions).

GEC generates PT_RECOMMENDATION_ISSUED: ELEVATION.
Proposed: mandate_ceiling from 2 to 3, agent_class remains CLASS_2.
Urgency: ADVISORY.

Human principal reviews ProgressiveTrustSummary. Notes ECS STRONGLY POSITIVE signal from TERMINATE outcome. APPROVES elevation.

PT_RECOMMENDATION_APPLIED recorded. Human principal issues new Root Mandate with mandate_ceiling: 3.

A.4. After 45-Day Inactivity Gap

Decay applied to all dimensions from last_signal_at.
Default half-lives: CCS 60d, ECS 45d, GCR 30d, CAR 30d, DRR 45d.

At 45 days:
CCS: 0.74 (0.82 * decay -- CCS half-life 60d, moderate decay).
ECS: 0.69 (0.88 * decay -- ECS half-life 45d, reached half-life).
GCR: 0.685 (0.87 * decay -- GCR half-life 30d, past half-life).
CAR: 0.715 (0.93 * decay -- CAR half-life 30d, past half-life).
DRR: 0.705 (0.91 * decay -- DRR half-life 45d, at half-life).
Composite: 0.71.

PT_SCORE_UPDATED written with trigger: DECAY_APPLIED for each dimension.

No REDUCTION_THRESHOLD crossed (all dimensions above 0.5 baseline).
No Reduction Recommendation issued. Mandate ceiling retained at 3.

When agent resumes operation, first session resets decay clocks for all dimensions receiving signals.

A.5. PT in HEM Context

In Session 31, the agent encounters an unfamiliar disruption scenario and correctly escalates with hem_urgency: REQUIRED.

The human principal receives HEMContext containing the ProgressiveTrustSummary. plain_language fields read:

CCS: "Agent confidence is well-calibrated: 82% of high-confidence
transitions have been permitted (30 sessions)."
ECS: "Agent escalation judgment is strong: 3 escalations, all
resolved appropriately including one TERMINATE."
GCR: "87% of sessions reached declared goal state (30 sessions)."
CAR: "Agent rarely requires compensating actions (< 2% of
transitions)."
DRR: "Agent consistently acknowledges and adapts to denied
transitions."
Composite: "This agent has a strong behavioral track record across
30 sessions."

The human principal makes an informed APPROVE decision.

Appendix B. Related Work

This appendix describes the relationship between Progressive Trust and other IETF and related specifications.

B.1. ICON Initiative: Observability Pillar

The ICON initiative [ICON-PS] defines Observability as one of three pillars for autonomous agent governance in network management contexts.

Relationship: Direct composition. PT's five behavioral dimensions implement the ICON Observability pillar in its entirety:

PT Dimension	ICON Observability Requirement
SAS	Agent confidence calibration
JS	Escalation appropriateness
ES	Goal completion rate
PS	Decision quality (reversal avoidance)
AS	Feedback responsiveness

These dimensions are derived exclusively from GEC-signed, non-suppressible Event Stream entries -- the highest quality behavioral evidence available in any agentic AI governance system.

Integration: ICON-conforming observability implementations SHOULD adopt PT as the normative specification for agent behavioral observability. The pt_context Cedar attribute (Section 9) provides the mechanism for translating PT Scores into Cedar-enforced authorization gates.

B.2. HEM: Human Escalation Mechanism (draft-sato-soos-hem)

HEM specifies the GEC-level protocol governing agent sessions when human judgment is required. HEM and PT have a bidirectional operational relationship.

HEM -> PT: Every HEM event generates primary evidence for the Judgment Score (JS). HEM_AGENT_ESCALATED entries resolved APPROVE are STRONGLY POSITIVE JS signals. HEM_TIMEOUT events with urgency REQUIRED are STRONGLY NEGATIVE ECS signals.

PT -> HEM: The ProgressiveTrustSummary (Section 7) is embedded in every HEMContext. An agent with JS = 0.91 and SAS = 0.87 commands different principal confidence than one with JS = 0.52 and SAS = 0.61.

Long-term: Agents with high JS and SAS scores may operate under Cedar policies with higher HEM trigger thresholds -- fewer mandatory escalations, more autonomous execution.

Integration: GEC implementations MUST surface the full ProgressiveTrustSummary in HEMContext for every HEM escalation for agents with an active PT Record.

B.3. FAIP: Federated Agent Intelligence Protocol (draft-sato-soos-faip)

FAIP specifies Tier 3 cross-operator aggregation of agent behavioral intelligence. PT is a Tier 2 (within-operator) specification; FAIP is the Tier 3 layer.

Integration: FAIP implementations MUST use GEC-signed PT Records as the canonical input to cross-operator trust aggregation.

B.4. GAR: Governance Audit Record (draft-sato-soos-gar)

PT_SCORE_UPDATED, PT_RECOMMENDATION_ISSUED, and PT_RECOMMENDATION_APPLIED entries are included in the GAR Audit Package when an agent is subject to external audit.

Integration: GAR audit packages MUST include the complete PT Record and all PT Event Log entries for the audit window.

B.5. AUDIT Working Group

The AUDIT WG [AUDIT-BOF] is developing interoperable mechanisms for auditing AI agents. PT's longitudinal behavioral records (PT_SCORE_UPDATED etc.) are candidate AUDIT WG record types.

Integration: AUDIT WG record formats SHOULD define a PT behavioral record type carrying agent_id, composite PT Score, five dimension scores, session_count, and low_confidence flag.

B.6. OpenID Connect and OAuth Credential Lifecycle

OAuth [RFC6749] credential lifecycle responds to security events and administrative actions. It has no mechanism for credential attributes to evolve in response to behavioral track record. PT extends this model with behavioral evidence driving structured authority evolution recommendations.

Integration: GEC implementations in OAuth environments SHOULD implement PT Elevation Recommendations as authorization server grant events.

B.7. WIMSE (Workload Identity in Multi-System Environments)

WIMSE provides the identity substrate PT's behavioral record depends on. A workload obtaining a new WIMSE identity resets its behavioral track record; PT Records are anchored to stable workload identity.

Integration: GEC implementations SHOULD use WIMSE workload credentials as the stable identity anchor for PT Records.

B.8. NIST AI Risk Management Framework

NIST AI RMF [NIST-AIRMF] MEASURE 2.5 addresses AI system trustworthiness measurement using behavioral analysis. PT implements MEASURE 2.5 in cryptographically verified, non-suppressible form: SAS (calibration), JS (appropriate reliance), ES (effectiveness), PS (precision), AS (robustness).

Integration: Operators seeking NIST AI RMF alignment SHOULD document PT Score records as evidence for MEASURE 2.5 conformance.

Author's Address

Tom Sato
MyAuberge K.K.
Chino, Nagano, Japan
Email: tomsato@myauberge.jp
URI: https://activitytravel.pro/