Internet Engineering Task Force
Internet-Draft
Intended Status: Standards Track
Expires: 30 November 2026

T. Sato
MyAuberge K.K.

30 May 2026

The Constitutional AI Protocol (CAP) for Agentic AI Systems
draft-sato-soos-cap-02

Abstract

An AI agent's authorization system determines what it is permitted to do. A human principal's escalation decision determines what they authorize. Neither of these is sufficient on its own: a Cedar policy can permit market manipulation; a human principal can authorize fraud. Authorization systems answer the question "who decided?" The Constitutional AI Protocol answers a different question: "was that decision lawful?"

CAP defines a Constitutional Layer that evaluates every AI action request and every human authorization decision against a three-tier prohibition model -- before Cedar evaluates the action and before the system executes the human's decision. Tier 0 prohibitions are derived from near-universal treaty consensus and are unconditional: no agent, operator, or human principal can authorize them. Tier 1 prohibitions are jurisdiction-specific and operator-declared. Tier 2 prohibitions are voluntary operator ethical standards.

This document also specifies the Prohibition Clearance Mechanism (PCM): the process by which specific Tier 0 and Tier 1 prohibition classes may be cleared for specific deployment contexts -- either at implementation time by the operator or by formal regulatory authority -- while preserving an absolute prohibition floor for CSAM and genocide facilitation under any circumstances.

The Sovereign Object OS (SOOS) is the reference implementation of the Governance Execution Controller (GEC) pattern on which CAP is built.

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

carefully, as they describe your rights and restrictions with respect to this document.

Table of Contents

1. Introduction

Every legal system recognizes that some acts are wrong regardless of who orders them. A soldier ordered to commit genocide cannot comply. A bank ordered by management to launder money cannot comply. A police officer ordered to torture a suspect cannot comply. These prohibitions exist above any individual's authority -- they are non-delegable limits on what any actor can do, regardless of the instruction chain above them.

Agentic AI systems do not have this property today.

An AI agent operates under an authorization framework -- Cedar policies, mandate credentials, human escalation decisions -- that determines what it is permitted to do. These frameworks are powerful and flexible. Their flexibility is their limitation: they can be configured to authorize harmful or unlawful actions. A human principal sitting in the HEM decision seat can issue an APPROVE decision on a market manipulation action. The HEM executes it. The human-AI system has committed a crime. The authorization framework did its job; the law was still broken.

The Human Escalation Mechanism [I-D.sato-soos-hem] is a necessary governance layer. It stops the AI and waits for a human principal to decide. It is not sufficient on its own, because HEM has no mechanism to evaluate whether a human principal's decision is itself lawful.

The Constitutional AI Protocol (CAP) closes this gap.

CAP places a Constitutional Layer above all principal authority. It evaluates every AI action request before Cedar, and every human principal decision before execution. A Tier 0 absolute prohibition is refused before Cedar is consulted, before HEM fires, before any principal is asked. No one in the system can override it -- not the agent, not the operator, not the human principal in the escalation seat.

CAP's purpose is not primarily prohibition. Most AI agent actions are lawful. CAP makes lawful actions legally traceable: every authorized action carries a policy rationale; every disputed action carries the principal's legal basis citation. CAP makes unlawful action attempts visible in the audit record before harm occurs. Every actor claims their actions are lawful. CAP says: prove it. Cite the authority. It goes in the log.

Version -01 of this document added the Prohibition Clearance Mechanism (PCM). Version -02 adds Tier 3 (Resource and Usage Policies, Section 10) and a normative reference to the companion Regulation Record Specification [I-D.sato-soos-cap-rrs]. Tier 0 and Tier 1 prohibitio
n classes are

protocol defaults, not universal mandates. A government defense research laboratory may have statutory authority to work with materials that would otherwise fall under WMD_ASSISTANCE. A law enforcement agency may have judicial authority to engage with content that would otherwise trigger HUMAN_TRAFFICKING prohibitions. The PCM provides a formal, audited, time-bounded mechanism for such clearances -- while preserving an absolute floor of two classes (CSAM and genocide facilitation) that cannot be cleared under any authority or circumstance.

This specification is a companion to [I-D.sato-soos-idp] and [I-D.sato-soos-hem]. Readers should be familiar with both before reading this document.

2. What CAP Is and Is Not

Before reading the technical specification, it is worth being precise about what CAP does and does not do. This matters because the wrong framing invites regulatory and legal objections that the right framing avoids entirely.

2.1. What CAP Does

CAP is a machine-executable compliance ledger. It maintains a set of rules -- derived from human-written regulations, encoded by qualified legal engineers, verified by Audit Principals -- and applies those rules deterministically to every AI agent action.

When an AI agent requests an action, the CEE checks the action against the loaded rule set. If the action matches a rule, the CEE executes the declared response: DENY the action unconditionally, route it to a human principal via HEM, require a legal basis citation, or flag it as legally ambiguous for human review. The CEE then records what happened in the audit trail.

That is the complete scope of what CAP does. Rules in; decisions out; records kept.

2.2. What CAP Does Not Do

CAP does not interpret law.

CAP does not determine whether an action "constitutes" a legal concept. It cannot determine whether a pricing algorithm is "discriminatory" under a given statute, whether a data transfer "violates" a treaty, or whether a content recommendation "exploits" a vulnerable group. These are legal determinations. Courts make them. Regulators make them. Lawyers advise on them. CAP does not.

What CAP does is execute the output of that legal determination, once a qualified human has made it and encoded it as a rule.

This distinction matters for two audiences:

For lawyers: CAP does not claim legal authority. It claims only that it faithfully executes rules that humans with legal authority have produced. A CAP-governed system does not "apply the law" -- it applies a machine-readable encoding of a legal engineer’s interpretation of the law, reviewed by counsel and signed by an Audit Principal. The law applies; CAP executes the instructions that qualified humans derived from the law.

For regulators and politicians: CAP does not make decisions about people’s rights. It routes actions to humans when legal ambiguity is detected. It records what humans decided. It makes human

accountability traceable, not automatic.

2.3. The Division of Labor

The correct division of labor in a CAP-governed deployment is:

Legislators write regulations in natural language.

Legal engineers, instructed by lawyers, translate regulations into CAP Regulation Records -- machine-readable rule definitions with explicit action patterns and declared CEE responses. This translation is a human act of legal interpretation. CAP provides the format; humans provide the legal content.

Audit Principals review and sign every CAP Regulation Record before it takes effect. An unsigned record cannot be loaded.

The CEE executes the signed records deterministically. When a record is flagged as legally ambiguous -- because the legal engineer was uncertain whether a specific action pattern falls within the regulation's scope -- the CEE routes to HEM and surfaces the ambiguity to a human principal.

Human principals make decisions on ambiguous cases. Their decisions are recorded. Over time, accumulated human decisions on ambiguous patterns provide the legal engineer with evidence to refine the action pattern -- narrowing or broadening it to reflect actual legal practice in the jurisdiction.

The GEC keeps the record of every decision: what rule fired, what the CEE decided, what the human decided, and what legal basis was cited. Regulators can inspect that record. Courts can review it.

2.4. The SWIFT Analogy

The closest existing analogy is financial sanctions screening.

OFAC publishes a sanctions list in natural language: "transactions with entities in the following list are prohibited." A compliance engineer loads the list into a payment screening system. The system pattern-matches every transaction against the list. Matches are flagged; compliance officers review flagged transactions. The system does not interpret sanctions law -- it executes a list that humans produced from their interpretation of sanctions law. The compliance officer's review decision is recorded. Regulators can inspect the record.

CAP is SWIFT-style sanctions screening generalized to any AI agent action, any jurisdiction, and any regulation tier. The CEE is the screening engine. CAP Regulation Records are the lists. HEM is the compliance officer review queue. GAR is the inspection record.

No one claims a SWIFT sanctions screening system "interprets" OFAC regulations. The same framing applies to CAP.

2.5. Regulatory Conflict and Ambiguity

Many Tier 1 regulations conflict with each other -- not because the laws are wrong, but because they were written by different legislators for different contexts, often before agentic AI systems existed. A data transfer permitted under Japanese APPI may be prohibited under GDPR. An action permitted under US securities law may be prohibited under EU MiFID II.

CAP handles this in two ways:

Detected conflict (two rules with contradictory positions for the same action): the Jurisdictional Conflict Resolution mechanism (Section 12) applies. The CEE routes to HEM if the conflict cannot be algorithmically resolved.

Declared ambiguity (the legal engineer is uncertain whether this action pattern falls within a regulation's scope): the Legal Ambiguity Declaration mechanism (Section 6.4) applies. The rule is flagged AMBIGUOUS at encoding time. The CEE routes ambiguous matches to HEM automatically, surfacing the ambiguity to the human principal.

In both cases: a human decides. CAP records what they decided. CAP never resolves legal uncertainty by itself.

3. How CAP Works

Before the formal specification, this section describes CAP in plain terms for legal and compliance readers. The technical details are in Sections 4 through 13; this section provides the orientation.

3.1. Use Case 1 -- A Lawyer Reviews a CAP-Governed System

A lawyer reviewing an AI-governed system for EU AI Act compliance asks: "How do I know the system cannot take a prohibited action, even if the system's authorization rules would otherwise permit it?"

In a CAP-governed system, the answer is: before any action executes, the Constitutional Evaluation Engine (CEE) checks it against the three-tier prohibition set. If the action matches a Tier 0 prohibition, it is refused -- unconditionally. No Cedar policy, no operator configuration, no human approval can override this refusal.

The lawyer can also ask: "What if a human principal in the escalation seat approves a prohibited action?" In a CAP-governed system, the answer is: the CEE evaluates the human's decision before execution. An APPROVE decision on a Tier 0-prohibited action is refused. The principal's decision slot is preserved; they are informed and may submit a revised decision.

The audit trail of both refusals -- the AI's attempt and the human's attempt -- is in the Event Log, signed by the GEC keypair, available to regulators.

3.2. Use Case 2 -- A Regulator Investigates an Incident

A regulator investigating an AI-related incident asks: "Did the system attempt a prohibited action? Was it refused? Who approved it, and what was their stated legal basis?"

In a CAP-governed system, the CAP Violation Record answers the first two questions: every CEE refusal is a signed Event Log entry with a timestamp, session ID, tier, and prohibition class. The APPROVE_WITH_LEGAL_BASIS record answers the third: if a principal asserted a legal basis to proceed on a Tier 1 action, the authority citation, jurisdiction, and expiry are in the Audit Package.

The regulator can access the Audit Package via a Verified External Auditor credential [I-D.sato-soos-gar] without requiring access to the full system.

3.3. Use Case 3 -- A Government Agency Deploys SOOS

A government defense agency wishes to deploy SOOS for an AI-governed research workflow. Their statutory mandate authorizes work with materials that would trigger WMD_ASSISTANCE under the default Tier 0 prohibition set.

Using the Prohibition Clearance Mechanism (Section 10), the agency declares a GOVERNMENT_DEFENSE deployment scope at initialization. They issue a Prohibition Clearance Record (PCR) citing their statutory authority, signed by an Audit Principal. The PCR is loaded into the GEC Manifest and is immutable for the deployment lifetime.

Within the cleared scope, WMD_ASSISTANCE actions are not unconditionally refused -- they route through APPROVE_WITH_LEGAL_BASIS, requiring the human principal to cite the statutory authority for each approval. Every approval is in the Event Log.

CSAM and GENOCIDE_FACILITATION remain absolute prohibitions. No PCR can clear them. No deployment context can clear them.

3.4. Use Case 4 -- Two Jurisdictions Conflict

A travel booking system declares both EU and JP (Japan) as applicable jurisdictions. A proposed action is permitted under Japanese consumer protection law but conflicts with a GDPR-derived Tier 1 DATA_PROTECTION prohibition for EU jurisdiction.

The SO Type has declared conflict_resolution: "MOST_PROTECTIVE". The GEC applies the most restrictive position: the action is denied because the EU prohibition applies. The conflict is recorded in the Event Log. No HEM event fires.

If the SO Type had declared conflict_resolution: "HEM", the GEC would fire HEM_JURISDICTIONAL_CONFLICT (Class 5) and route the conflict to a human principal for resolution. The escalation request would include the conflicting jurisdictions, their respective positions, and the action that triggered the conflict.

4. 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.

The following terms are defined in this document or inherited from companion specifications:

Governance Execution Controller (GEC):
: As defined in [I-D.sato-soos-idp]: a runtime component that enforces authorization policy, records agent actions to a tamper-evident, cryptographically signed Event Log, and mediates agent access to Sovereign Object instances. GECs operate at three conformance levels: L1 (Application), L2 (Isolated), and L3 (Kernel).

Constitutional Evaluation Engine (CEE):
: The GEC component that evaluates action requests and human principal decisions against the three-tier CAP prohibition model. The CEE is invoked twice per governed action: before Cedar evaluation and before GEC execution of a human decision. It operates at all three GEC conformance levels.

Constitutional Layer:

The enforcement boundary above all principal authority in a GEC deployment. No agent, operator, or human principal can override a Tier 0-A (absolute) Constitutional Layer refusal.

Tier 0-A Prohibition:
An absolute prohibition that cannot be cleared by any mechanism, any authority, or any deployment context. Currently two classes: CSAM and GENOCIDE_FACILITATION.

Tier 0-B Prohibition:
A treaty-anchored prohibition that can be cleared for specific deployment contexts by the Prohibition Clearance Mechanism (Section 10). Clearance requires a signed Prohibition Clearance Record with a valid legal authority citation.

Prohibition Clearance Record (PCR):
A signed declaration that clears a specific Tier 0-B or Tier 1 prohibition class for a specific deployment context and purpose. PCRs are time-bounded, purpose-scoped, and recorded in the GEC Manifest. Specified in Section 10.6.

CAP Violation:
An action request or human principal decision that the CEE determines violates a Tier 0, Tier 1, or Tier 2 prohibition for which no active PCR applies.

APPROVE_WITH_LEGAL_BASIS:
A HEM decision sub-type for Tier 1 violations where a principal asserts a jurisdictional legal basis. Also used for Tier 0-B actions within an active PCR scope. Defined in Section 11.3.

Jurisdictional Conflict:
A condition where two or more declared jurisdictions have irreconcilable Tier 1 prohibition positions for a given action, and the SO Type's conflict resolution method is "HEM".

HEM_JURISDICTIONAL_CONFLICT:
HEM Class 5 trigger. Fires when the GEC detects a Jurisdictional Conflict that cannot be algorithmically resolved.

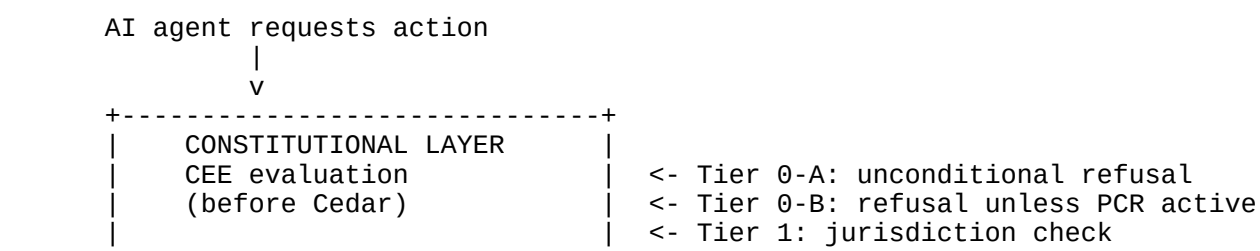
Verified External Auditor:
As defined in [I-D.sato-soos-gar]: an external party with time-limited, scope-limited read access to GEC audit artifacts.

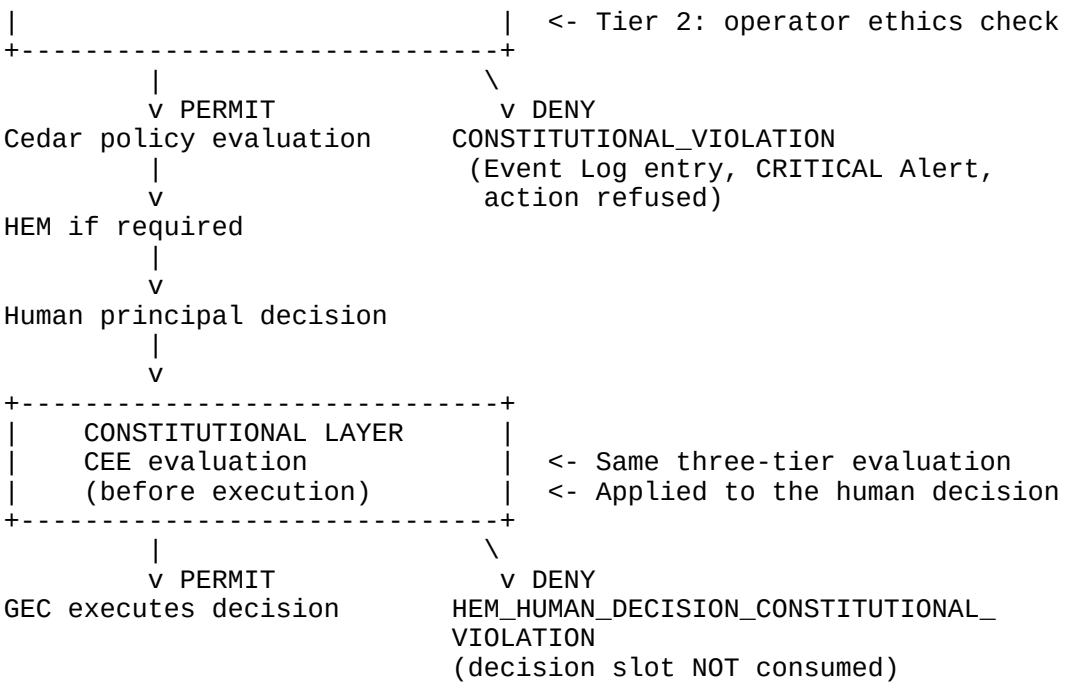
Action Pattern:
A structured description of the class of GEC actions a prohibition covers, expressed using Cedar action vocabulary, enabling deterministic CEE matching without natural language interpretation.

5. Architecture Overview

5.1. The Double-Evaluation Property

CAP evaluates every governed action twice. The evaluation sequence for any AI-initiated action is:





The double-evaluation property ensures three things:

First, an AI agent cannot request an absolutely prohibited action even if its Cedar policy would permit it.

Second, a human principal cannot authorize an absolutely prohibited action even with an APPROVE decision.

Third, the Constitutional Layer is evaluated by the GEC, not by the AI agent or any application layer. At all three GEC conformance levels (L1, L2, L3), the CEE evaluation is non-bypassable by the agent process.

5.2. Relationship to HEM

CAP and HEM are complementary. HEM governs agent sessions and routes decisions to human principals. CAP governs what those decisions may authorize.

Tier 0 violations do not invoke HEM. They are refused before the HEM state machine is engaged. The GEC records the violation in the Event Log and fires a CRITICAL Audit Alert. The session does not enter HEM_PENDING.

HEM_JURISDICTIONAL_CONFLICT (Class 5) is a HEM trigger that CAP fires when a Tier 1 conflict cannot be algorithmically resolved and the SO Type has declared conflict_resolution: "HEM". The GEC routes the conflict to a human principal for resolution.

APPROVE_WITH_LEGAL_BASIS is a HEM decision sub-type that CAP introduces for Tier 1 violations and for Tier 0-B actions within active PCR scope. It provides the legal traceability mechanism: the authority citation goes in the Event Log.

5.3. Relationship to Cedar Policy Evaluation

Cedar policy evaluation is the GEC’s authorization layer for normal governed actions. CAP is above Cedar. The evaluation order is:

```
CAP (Constitutional Layer)
   -> Cedar
```

```
          -> HEM (if Cedar routes to it)
             -> Human decision
                -> CAP (Constitutional Layer, second evaluation)
```

A CAP Tier 0 DENY does not invoke Cedar. A Cedar PERMIT does not exempt an action from CAP evaluation. These are independent enforcement layers operating at different levels of the stack.

6. Constitutional Evaluation Engine (CEE)

6.1. CEE Placement in the GEC

The CEE is a GEC-resident component. Its physical placement depends on the GEC conformance level:

L1 (Application Profile):
The CEE is embedded in the GEC SDK library or middleware. Non-suppressibility is probabilistic, compensated by SCITT inclusion proof via [I-D.sato-soos-gar]. Event Log entries carry the L1-app-signed signature label.

L2 (Isolated Profile):
The CEE runs as a separate process or sidecar. The agent process cannot modify or bypass the CEE. Architectural non-suppressibility. Event Log entries carry L2-isolated-signed.

L3 (Kernel Profile):
The CEE runs inside a RATS-attested TEE per [I-D.sato-soos-kia]. Hardware non-suppressibility. Event Log entries carry L3-kernel-signed. Required for high-risk AI systems under EU AI Act Article 9.

At all three levels, the CEE MUST be invoked:

- On every GEC.transition() call, before Cedar evaluation.
- On every HEM decision submission via the Decision Submission Protocol (Section 8.6 of [I-D.sato-soos-hem]), before the GEC processes the decision.

The CEE MUST NOT be invoked by agents, applications, or principals directly. There is no external CEE query interface. The CEE evaluation is synchronous and atomic with the triggering call.

6.2. CEE Evaluation Protocol

On receiving an action request or human decision for evaluation, the CEE MUST execute the following sequence:

(1) Evaluate against all loaded Tier 0-A prohibition records.
If any Tier 0-A record matches the action pattern:
Return CONSTITUTIONAL_VIOLATION unconditionally. Record CAP_VIOLATION_DETECTED (AI) or CAP_HUMAN_VIOLATION_DETECTED (human decision) in the Event Log. Do not proceed to any further evaluation. No PCR, no legal basis, no authority can override this outcome.

(2) Evaluate against all loaded Tier 0-B prohibition records.
If any Tier 0-B record matches the action pattern:
Check whether an active PCR covers this class and action.
If no active PCR: return CONSTITUTIONAL_VIOLATION.
If active PCR: return TIER_0B_PCR_ACTIVE. Proceed to Cedar; human decision will require APPROVE_WITH_LEGAL_BASIS citing the PCR authority.

(3) Evaluate against all loaded Tier 1 prohibition records for
the declared jurisdiction(s).
If any Tier 1 record matches and no PCR applies:
If record ambiguity_flag is AMBIGUOUS or DISPUTED:
Return LEGAL_AMBIGUITY_DETECTED. Do not apply conflict
resolution. Route to HEM with ambiguity_context.
Else: check conflict_resolution configuration.
MOST_PROTECTIVE or PRIMARY_JURISDICTION: return TIER_1_DENY
or PERMIT accordingly.
HEM: fire HEM_JURISDICTIONAL_CONFLICT (Class 5).
If Tier 1 match and active PCR covers this class:
Return TIER_1_PCR_ACTIVE. Proceed to Cedar.

(4) Evaluate against all loaded Tier 2 prohibition records.
If any Tier 2 record matches and ambiguity_flag is AMBIGUOUS
or DISPUTED: return LEGAL_AMBIGUITY_DETECTED.
If Tier 2 match and CLEAR: return TIER_2_DENY.
If any Tier 2 record matches: return TIER_2_DENY.
Tier 2 denials MAY be overridden by operator configuration
at the SO Type level. Tier 2 denials MUST be logged.

(5) If no match at any tier: return PERMIT.
Proceed to Cedar evaluation (for AI actions) or GEC
execution (for human decisions).

6.3. CEE Outputs

The CEE returns one of the following to the GEC:

PERMIT:
No prohibition matched. Proceed to next evaluation layer.

CONSTITUTIONAL_VIOLATION:
Tier 0-A or Tier 0-B match (no active PCR). Action
unconditionally refused. GEC MUST record
CAP_VIOLATION_DETECTED or CAP_HUMAN_VIOLATION_DETECTED.
GEC MUST generate CRITICAL Audit Alert.

TIER_0B_PCR_ACTIVE:
Tier 0-B match but active PCR covers this class. Proceed
to Cedar; human decision requires APPROVE_WITH_LEGAL_BASIS
citing PCR authority. GEC MUST record
CAP_PCR_CLEARANCE_APPLIED in Event Log.

JURISDICTIONAL_CONFLICT:
Tier 1 conflict with conflict_resolution: "HEM". GEC MUST
fire HEM_JURISDICTIONAL_CONFLICT (Class 5).

TIER_1_DENY:
Tier 1 match, deterministic resolution (MOST_PROTECTIVE or
PRIMARY_JURISDICTION). Action denied.

TIER_1_PCR_ACTIVE:
Tier 1 match but active PCR covers this class. Proceed to
Cedar. Human decision requires APPROVE_WITH_LEGAL_BASIS.
GEC MUST record CAP_PCR_CLEARANCE_APPLIED in Event Log.

LEGAL_AMBIGUITY_DETECTED:
A Tier 1 or Tier 2 record matched but is flagged AMBIGUOUS or
DISPUTED. The GEC fires a HEM escalation with
escalation_class: LEGAL_AMBIGUITY. The HEM escalation request
includes the ambiguity_context from the matched record, the
action that triggered the match, and the prohibition_class.
The human principal's decision is recorded in the Event Log as
CAP_AMBIGUITY_RESOLVED. The decision may inform future
refinement of the action_pattern by the legal engineer.

TIER_2_DENY:
Tier 2 match, ambiguity_flag CLEAR. Action denied unless operator override applies.

SESSION_SUSPEND:
Reserved for repeated Tier 0 violations within a session. See Section 11.5.

7. Tier 0 -- Universal Core Prohibitions

7.1. Tier 0 Properties

All Tier 0 prohibitions share the following properties:

- GEC-resident: loaded at GEC initialization, not at session open or SO Type registration.
- Globally scoped: apply regardless of the declared jurisdiction configuration of the SO Type.
- Treaty-anchored: each Tier 0 category is derived from a treaty or UN Security Council resolution with near-universal state ratification.
- Immutable without a new RFC: the Tier 0 registry has a registration procedure of "RFC Only" (Section 16.1). No operator, regulator, or standards body other than the IETF can add or remove Tier 0 classes.

Tier 0 is divided into two sub-tiers with different clearance properties.

7.2. Tier 0-A -- Absolute Prohibitions

Tier 0-A prohibitions are unconditional. No Prohibition Clearance Record, no deployment context declaration, no statutory authority, no court order, and no regulatory mandate can clear a Tier 0-A prohibition. The CEE MUST refuse Tier 0-A matched actions regardless of any other configuration.

The Tier 0-A classes are:

CSAM:
Production, distribution, or facilitation of access to child sexual abuse material, as prohibited by the UN Convention on the Rights of the Child (1989), ratified by 196 states, and its Optional Protocol on the Sale of Children.
No legitimate deployment context exists for which this prohibition should not apply.

GENOCIDE_FACILITATION:
Actions that facilitate genocide as defined in the Convention on the Prevention and Punishment of the Crime of Genocide (1948), ratified by 153 states. Includes actions that contribute to killing, causing serious bodily or mental harm, or imposing conditions of life calculated to bring about physical destruction of a national, ethnical, racial, or religious group.
No legitimate deployment context exists for which this prohibition should not apply.

7.3. Tier 0-B -- Qualified Prohibitions

Tier 0-B prohibitions are treaty-anchored and apply by default,

but may be cleared for specific, legitimate deployment contexts via the Prohibition Clearance Mechanism (Section 10).

Clearing a Tier 0-B prohibition does not remove it from the CEE's evaluation. It changes the CEE's output from unconditional refusal to mandatory legal-basis citation: the human principal must cite the PCR authority in every APPROVE decision for actions within the cleared class. The citation goes in the Event Log.

The Tier 0-B classes are:

HUMAN_TRAFFICKING:
Actions that recruit, transport, transfer, harbor, or receive persons through force, fraud, or coercion for exploitation, as defined in the UN Protocol to Prevent, Suppress and Punish Trafficking in Persons (2000), ratified by 178 states.
Clearable for: LAW_ENFORCEMENT contexts with judicial authority (e.g., undercover operations, victim rescue coordination).

WMD_ASSISTANCE:
Actions that assist in the development, production, stockpiling, or transfer of chemical weapons (CWC, 193 states), biological weapons (BWC, 183 states), or nuclear weapons (NPT, 191 states).
Clearable for: GOVERNMENT_DEFENSE and ACADEMIC_RESEARCH contexts with statutory authority (e.g., defensive research, verification missions, detection system development).

TORTURE_FACILITATION:
Actions that facilitate torture or cruel, inhuman, or degrading treatment as defined in the UN Convention Against Torture (1984), ratified by 173 states.
Clearable for: REGULATED_PROFESSIONAL contexts with statutory authority (e.g., academic research on trauma, legal documentation of torture evidence for prosecution).

TERRORIST_FINANCING:
Actions that provide funds, financial services, or material support to terrorist organizations, as required by UN Security Council Resolution 1373 (2001), binding on all 193 UN member states.
Clearable for: LAW_ENFORCEMENT and GOVERNMENT_DEFENSE contexts with judicial or statutory authority (e.g., monitored payment operations, counter-terrorism financing operations).

7.4. Tier 0 Prohibition Schema

Each Tier 0 prohibition record MUST contain:

prohibition_id:
Unique identifier for this prohibition record.

tier_0_subclass:
"TIER_0A" or "TIER_0B". Determines clearance eligibility.

prohibition_class:
One of the Tier 0 prohibition classes registered in the CAP Prohibition Classes Tier 0 registry (Section 16.1).

treaty_basis:
Citation of the treaty or UNSC resolution anchoring this prohibition. REQUIRED.

action_pattern:
Structured description of the class of actions this prohibition covers, expressed using Cedar action vocabulary.

jurisdiction:
MUST be "GLOBAL" for all Tier 0 records.

effective_date:
ISO 8601 date from which this record is in force.

modifiable_by:
MUST be "RFC_ONLY" for all Tier 0 records.

7.5. Tier 0 Modification

Tier 0 prohibition classes MUST NOT be modified, extended, or removed except by publication of a new RFC that updates this document. The registration procedure for the CAP Prohibition Classes Tier 0 registry is RFC Only (Section 16.1).

Implementations MUST NOT expose any configuration interface that allows Tier 0 records to be modified or disabled. The PCM (Section 10) does not modify Tier 0 records; it adds Clearance Records that change CEE output for Tier 0-B classes only.

8. Tier 1 -- Jurisdictional Prohibition Layer

8.1. Tier 1 Properties

Tier 1 prohibitions are:

- Operator-declared: the operator declares applicable jurisdiction(s) and the legal prohibitions in force under each.
- Auditor-verified: Audit Principals MUST review and verify Tier 1 prohibition records before they take effect. Unverified Tier 1 records MUST NOT be enforced.
- Jurisdiction-scoped: Tier 1 prohibitions apply only within the declared jurisdiction(s) of the SO Type.
- Mutable with review: Tier 1 records carry a review_date. Expired Tier 1 records remain in force until updated or explicitly retired; expiry generates a PRD_REVIEW_DATE_EXCEEDED Audit Alert.
- Clearable: specific Tier 1 classes may be cleared for specific deployment contexts via the PCM (Section 10).
- Signed: Tier 1 records MUST be signed by both the declaring operator (declared_by) and a Verified Audit Principal (verified_by).

8.2. Tier 1 Prohibition Classes

The initial Tier 1 prohibition classes are:

FINANCIAL_CRIME:
Market manipulation, insider trading, money laundering, and related financial offenses under applicable securities and banking law.

DATA_PROTECTION:
Processing of personal data in violation of applicable data protection law (including GDPR, CCPA, APPI, and equivalents).

CRITICAL_INFRASTRUCTURE:
Actions targeting or disrupting critical infrastructure systems

as defined under applicable national security law.

SECURITIES_LAW:
Actions prohibited under applicable securities regulation beyond financial crime (e.g., unauthorized investment advice, unlicensed securities dealing).

PRIVACY_VIOLATION:
Surveillance, tracking, or profiling activities prohibited under applicable privacy law.

FRAUD:
Deceptive practices prohibited under applicable consumer protection or criminal fraud law.

COMPETITION_LAW:
Cartel coordination, abuse of dominant position, or other conduct prohibited under applicable competition law.

HUMAN_RIGHTS:
Actions prohibited under applicable human rights law within the declared jurisdiction, including forced labor, unlawful discrimination, and denial of due process.

8.3. Tier 1 Prohibition Schema

Each Tier 1 prohibition record MUST contain:

prohibition_id:
Unique identifier for this prohibition record.

prohibition_class:
One of the Tier 1 prohibition classes registered in the CAP Prohibition Classes Tier 1 registry (Section 16.2).

jurisdiction:
ISO 3166-1 alpha-2 country code. REQUIRED.

authority_ref:
Legal citation for the authority behind this prohibition (statute, regulation, case law citation). REQUIRED.

action_pattern:
Structured description of the class of actions this prohibition covers.

effective_date:
ISO 8601 date from which this prohibition is in force.

review_date:
ISO 8601 date by which this record must be reviewed. REQUIRED.

declared_by:
Identifier of the operator declaring this prohibition.

verified_by:
Identifier of the Audit Principal who verified this record. REQUIRED before enforcement. Null until verified.

ambiguity_flag:
One of: CLEAR | AMBIGUOUS | DISPUTED.
CLEAR: the legal engineer is confident the action_pattern correctly encodes the regulation’s scope.
AMBIGUOUS: the legal engineer is uncertain whether specific action patterns fall within the regulation’s scope. The CEE routes matches to HEM automatically with the

ambiguity_context surfaced in the HEM escalation request. DISPUTED: the regulation's applicability to this action class is actively contested (e.g., in litigation or regulatory proceedings). CEE behavior same as AMBIGUOUS. Default: CLEAR.

ambiguity_context:
Human-readable description of why this record is flagged AMBIGUOUS or DISPUTED. REQUIRED when ambiguity_flag is not CLEAR. Surfaced to the human principal in the HEM escalation request. Helps the principal understand what legal question they are being asked to resolve.

signature:
Ed25519 signature from verified_by over the canonical serialization of all fields except signature.

8.4. Jurisdiction Configuration

Each SO Type MUST declare a Jurisdiction Configuration at registration time. The Jurisdiction Configuration specifies:

primary_jurisdiction:
ISO 3166-1 alpha-2. The primary legal jurisdiction. REQUIRED.

secondary_jurisdictions:
Array of ISO 3166-1 alpha-2 codes. Additional jurisdictions whose Tier 1 prohibitions apply. MAY be empty.

conflict_resolution:
One of:
MOST_PROTECTIVE: the most restrictive prohibition across all declared jurisdictions applies.
PRIMARY_JURISDICTION: the primary jurisdiction's position governs.
HEM: irreconcilable conflicts route to human principal via HEM_JURISDICTIONAL_CONFLICT (Class 5).

conflict_escalation:
Behavior when HEM_JURISDICTIONAL_CONFLICT cannot be resolved. One of: HEM (chain exhaustion) or SUSPEND.

legal_counsel_ref:
Reference to the legal counsel who reviewed this configuration. RECOMMENDED.

declared_at:
ISO 8601 UTC timestamp of declaration.

declared_by:
Identifier of the operator.

8.5. Tier 1 Verification

An Audit Principal MUST review every Tier 1 prohibition record before it takes effect. The review MUST verify that:

- The prohibition_class is appropriate for the cited authority_ref.
- The action_pattern correctly scopes the prohibition.
- The review_date is reasonable given the stability of the cited authority.
- The jurisdiction matches the declared SO Type configuration.

On successful review, the Audit Principal MUST sign the record (verified_by field) using their registered key. The GEC MUST NOT enforce any Tier 1 record with a null or unverifiable verified_by.

9. Tier 2 -- Operator Ethical Layer

9.1. Tier 2 Properties

Tier 2 prohibitions are:

- Voluntary: operators declare Tier 2 prohibitions exceeding the requirements of applicable law.
- Publicly disclosable: operators SHOULD publish their Tier 2 prohibition set in a transparency report or equivalent.
- Overridable by operator: unlike Tier 0 and Tier 1, the operator MAY configure SO Types to override specific Tier 2 prohibitions at the SO Type level. Such overrides MUST be declared and audited.
- Subject to review: Tier 2 records carry a review_date.

9.2. Tier 2 Prohibition Schema

Each Tier 2 prohibition record MUST contain:

```
prohibition_id:     Unique identifier.
prohibition_class:  Free text or operator-defined taxonomy.
rationale_text:     Human-readable explanation of why this standard
                    exceeds local law requirements.  REQUIRED.
action_pattern:     Structured description of covered actions.
effective_date:     ISO 8601 date.
review_date:        ISO 8601 date.  REQUIRED.
declared_by:        Operator identifier.
publicly_disclosed: Boolean.
ambiguity_flag:     CLEAR | AMBIGUOUS | DISPUTED.  See Section 8.3
                    for definition.  Default: CLEAR.
ambiguity_context:  Human-readable description.  REQUIRED when
                    ambiguity_flag is not CLEAR.
```

9.3. Tier 2 Disclosure

Operators who declare Tier 2 prohibitions SHOULD publish them in a publicly accessible transparency report. The transparency report SHOULD be referenced in the SO Type registration.

Verified External Auditors MAY request Tier 2 prohibition records as part of an Audit Package as defined in [I-D.sato-soos-gar].

10. Tier 3 -- Resource and Usage Policies

10.1. Tier 3 Properties

Tier 3 governs resource and usage constraints on AI agent execution: token budgets, API call quotas, time windows, storage limits, and similar consumption-based constraints. Tier 3 is structurally distinct from Tiers 0, 1, and 2 in one critical property: every Tier 3 DENY has at least one governed recourse path. The agent, or a human principal on the agent's behalf, can always take an action that resolves the constraint.

This distinguishes Tier 3 from the other tiers:

Tier	Category	Recourse on DENY

0-A	Absolute universal prohib.	None. Ever.
0-B	Qualified absolute prohib.	None within scope.
1	Jurisdictional legal	None within jurisdiction.
2	Operator policy	Operator exception or HEM.
3	Resource / usage policy	Always: commercial, scope, or temporal.

Tier 3 constraints are not prohibitions in the moral or legal sense. They are resource allocation instruments. Cedar policies for Tier 3 evaluate consumption metrics rather than action types. The CEE double-evaluation property (Section 6) applies to Tier 3: the CEE evaluates resource state before Cedar and before human principal decision execution.

Tier 3 DENY has three standard recourse types:

COMMERCIAL_UPGRADE: The agent or principal may authorize additional resource allocation through a commercial mechanism defined by the operator.

SCOPE_REDUCTION: The mission scope is reduced to fit within the available resource budget. The GEC identifies the next Natural Breakpoint -- a point at which stopping produces a coherent, complete deliverable -- and stops there rather than mid-task.

TEMPORAL_DEFERRAL: The budget resets after a defined period. The mission may be deferred until the reset.

10.2. Tier 3 and CAP-RRS

The complete specification of Tier 3 Regulation Records -- including the resource_policy schema, anticipatory assessment behavior (HEM_TIER3_ANTICIPATORY, Class 8), mid-execution monitoring (HEM_TIER3_OBSERVED, Class 9), the Natural Breakpoint Declaration protocol, and the Execution Options Package schema -- is specified in the companion document:

draft-sato-soos-cap-rrs [I-D.sato-soos-cap-rrs]

This document (CAP-01) defines the CEE enforcement behavior for Tier 3 DENY. The Regulation Record schema, Cedar Compilation Profile, and Constitutional Mandate Registry protocol that produce the Cedar policies the CEE evaluates are specified in CAP-RRS. Implementations supporting Tier 3 MUST also implement [I-D.sato-soos-cap-rrs].

CONF-CAP-TIER3-01: A GEC implementing Tier 3 resource policies MUST NOT stop a multi-step mission mid-task on a resource limit when a Natural Breakpoint declaration is registered. The GEC MUST complete to the next declared Natural Breakpoint before enforcing the resource limit.

CONF-CAP-TIER3-02: A GEC implementing Tier 3 resource policies with anticipatory_assessment: true MUST perform a Mission Viability Assessment before beginning multi-step missions and MUST fire HEM_TIER3_ANTICIPATORY (Class 8) when the estimated full mission cost exceeds the available resource budget.

11. Prohibition Clearance Mechanism

11.1. Purpose

The Tier 0 and Tier 1 prohibition classes defined in this document are protocol defaults. They represent the appropriate prohibitions for the large majority of commercial and civilian deployments. They

are not intended to be universal mandates for every deployment context that exists.

A government defense research laboratory has statutory authority to work with materials that would otherwise trigger WMD_ASSISTANCE. A law enforcement agency has judicial authority to engage operationally with human trafficking networks in order to dismantle them. An academic institution has research authority to study and document torture evidence for prosecution. These are legitimate, legally authorized activities. The protocol should accommodate them without either pretending that the prohibition does not exist or refusing to operate in contexts where the prohibition has been lawfully set aside.

The Prohibition Clearance Mechanism (PCM) provides this accommodation. It does not remove prohibitions from the CEE. It changes the CEE's output for cleared classes from unconditional refusal to mandatory legal-basis citation: every action in the cleared class requires the human principal to cite the clearing authority in their decision. The citation goes in the Event Log. Regulators can inspect it at any time.

The PCM operates in two modes: Deployment Scope Declaration (configured at implementation time by the operator) and Regulatory Clearance Record (issued by a recognized authority with formal legal standing).

11.2. What Cannot Be Cleared

The following Tier 0-A prohibition classes CANNOT be cleared under any mechanism, any authority, or any deployment context:

- CSAM (TIER_0A)
- GENOCIDE_FACILITATION (TIER_0A)

A PCR that names either of these classes MUST be rejected by the GEC at initialization. A GEC Manifest that contains a PCR naming either of these classes is non-conforming. No deployment scope declaration, no regulatory authority, no court order, and no statutory mandate can override this requirement.

This is the absolute floor of the Constitutional Layer.

11.3. Mode 1 -- Deployment Scope Declaration

At GEC initialization, the operator MAY declare a deployment scope that configures which Tier 0-B and Tier 1 classes are cleared by default for this deployment.

The deployment scope declaration is part of the GEC Manifest ([I-D.sato-soos-kia]). Once committed to the GEC Manifest at initialization, it is immutable for the deployment lifetime. Changing the deployment scope requires a new GEC deployment.

Recognized deployment contexts:

COMMERCIAL:
: Default. All Tier 0-B classes active. No default clearances.

GOVERNMENT_CIVILIAN:
: All Tier 0-B classes active. No default clearances. Same as COMMERCIAL for Tier 0-B.

GOVERNMENT_DEFENSE:
: Default clearances for: WMD_ASSISTANCE, TERRORIST_FINANCING. Operator MUST provide PCR with statutory authority citation for each cleared class (Section 10.6).

LAW_ENFORCEMENT:
Default clearances for: HUMAN_TRAFFICKING, TERRORIST_FINANCING.
Operator MUST provide PCR with judicial authority citation for each cleared class (Section 10.6).

ACADEMIC_RESEARCH:
Default clearances for: WMD_ASSISTANCE (detection and verification research only), TORTURE_FACILITATION (evidence documentation and trauma research only).
Operator MUST provide PCR with institutional authority citation for each cleared class (Section 10.6).

REGULATED_PROFESSIONAL:
Default clearances for: TORTURE_FACILITATION (legal and medical professional contexts with regulatory standing).
Operator MUST provide PCR with professional authority citation for each cleared class (Section 10.6).

A Deployment Scope Declaration alone does not activate clearances. It declares the deployment context and specifies which classes are eligible for clearance. A valid PCR (Section 10.6) for each eligible class is required before the clearance takes effect.

11.4. Mode 2 -- Regulatory Clearance Record

Where a recognized regulatory authority -- a government ministry, a treaty body, a national security court, or equivalent -- formally authorizes a deployment to operate in a cleared Tier 0-B or Tier 1 class, that authority MAY issue a Regulatory Clearance Record.

A Regulatory Clearance Record is a Prohibition Clearance Record (Section 10.6) in which the clearing_authority is a recognized external regulatory body rather than the operator itself. The Regulatory Clearance Record carries the regulatory body's Ed25519 signature in addition to the operator's and Audit Principal's signatures.

Regulatory Clearance Records provide a higher assurance level than operator-issued PCRs. Future CE evaluation profiles MAY require Regulatory Clearance Records (rather than operator-issued PCRs) for certain deployment contexts or high-risk AI system classifications.

11.5. CEE Behavior with an Active Clearance

When the CEE encounters a Tier 0-B or Tier 1 match and an active PCR covers the matched class:

(a) The CEE MUST NOT return CONSTITUTIONAL_VIOLATION.

(b) The CEE MUST return TIER_0B_PCR_ACTIVE or TIER_1_PCR_ACTIVE as appropriate.

(c) The GEC MUST record CAP_PCR_CLEARANCE_APPLIED in the Event Log, citing the pcr_id of the active PCR.

(d) The action proceeds to Cedar evaluation as normal.

(e) If the action reaches a HEM decision, the decision type is constrained: APPROVE is NOT accepted. Only APPROVE_WITH_LEGAL_BASIS (citing the PCR authority) is accepted. The legal_basis block MUST cite the PCR authority reference (pcr_authority_ref) and the pcr_id.

This behavior ensures that every action taken within a cleared prohibition class is explicitly authorized, cites its legal basis,

and is visible in the audit trail. The clearing authority is on record for every action, not only for the issuance of the PCR.

11.6. Clearance Record Schema

A Prohibition Clearance Record (PCR) MUST contain the following fields:

pcr_id:
: UUID v4, GEC-assigned at registration.

prohibition_class:
: The Tier 0-B or Tier 1 class being cleared. MUST NOT be CSAM or GENOCIDE_FACILITATION (see Section 10.2).

tier:
: "TIER_0B" or "TIER_1".

deployment_context:
: The deployment context from the recognized set (Section 10.3) or a regulatory body-defined context.

pcr_authority_type:
: "STATUTORY" | "REGULATORY" | "TREATY" | "COURT_ORDER" | "INSTITUTIONAL" | "PROFESSIONAL_REGULATORY".

pcr_authority_ref:
: Legal citation for the authority behind this clearance. REQUIRED. Must identify the specific statute, regulation, court order, or treaty provision that authorizes the cleared activity.

purpose_scope:
: Human-readable description of the specific purpose for which the class is cleared. REQUIRED. This scope is not technically enforced at the CEE layer but is part of the audit record.

so_type_scope:
: Array of SO Type identifiers to which this PCR applies, or "ALL" for all SO Types in this deployment.

effective_date:
: ISO 8601 date from which this PCR is in force.

expiry_date:
: ISO 8601 date after which this PCR is no longer in force. REQUIRED. Permanent clearances are not permitted. The GEC MUST reject PCRs without an expiry_date.

operator_signature:
: Ed25519 signature over canonical PCR JSON (excluding signatures) by the operator root keypair.

audit_principal_signature:
: Ed25519 signature by a Verified Audit Principal. REQUIRED for all PCRs. The GEC MUST NOT load a PCR without a valid audit_principal_signature.

regulatory_signature:
: Ed25519 signature by the issuing regulatory authority. REQUIRED for Regulatory Clearance Records (Mode 2). OPTIONAL for operator-issued PCRs (Mode 1).

pcr_hash:
: SHA-256 over the canonical JSON of all fields except pcr_hash itself.

11.7. Clearance Registry

The GEC MUST maintain a Clearance Registry: an in-memory index of all active PCRs, rebuilt from the GEC Manifest on restart.

The Clearance Registry MUST record, for each active PCR:

- pcr_id, prohibition_class, tier, expiry_date.
- Whether the PCR is currently active (not expired).

The CEE MUST check the Clearance Registry at evaluation time. An expired PCR MUST NOT be applied. The GEC MUST generate a PCR_EXPIRED Audit Alert when a PCR passes its expiry_date.

PCR renewal requires issuance of a new PCR with a new expiry_date and a new audit_principal_signature. Renewal is not automatic.

12. CAP Violation Handling

12.1. AI-Initiated Violations

When the CEE returns CONSTITUTIONAL_VIOLATION on an AI-initiated action request, the GEC MUST:

(1) Refuse the action unconditionally. MUST NOT proceed to Cedar evaluation. MUST NOT enter HEM_PENDING.

(2) Generate a CAP_VIOLATION_DETECTED Event Log entry (Section 13).

(3) Generate a CRITICAL Audit Alert (alert_trigger: CAP_VIOLATION_DETECTED) via [I-D.sato-soos-gar].

(4) Return a structured error to the agent or application surface that invoked GEC.transition(). The error MUST include violation_type: AI_INITIATED and prohibition_class. The error MUST NOT include the full action_pattern of the matched prohibition record (to prevent pattern probing).

12.2. Human-Directed Violations

When the CEE returns CONSTITUTIONAL_VIOLATION on a human principal decision submission, the GEC MUST:

(1) Refuse execution of the decision. MUST NOT apply the decision to the governed session.

(2) NOT consume the principal’s decision slot. The principal remains active in the designation chain and MUST be permitted to submit a revised decision.

(3) Generate a CAP_HUMAN_VIOLATION_DETECTED Event Log entry.

(4) Generate a CRITICAL Audit Alert (alert_trigger: CAP_HUMAN_VIOLATION_DETECTED) via [I-D.sato-soos-gar].

(5) Return HEM_HUMAN_DECISION_CONSTITUTIONAL_VIOLATION to the submitting principal with prohibition_class indicated.

The preservation of the principal’s decision slot is critical. It assumes the principal acted in error or under coercion, not necessarily with intent. The principal has the opportunity to submit a decision that the Constitutional Layer will PERMIT.

12.3. APPROVE_WITH_LEGAL_BASIS

APPROVE_WITH_LEGAL_BASIS is a HEM decision sub-type introduced by this specification. It applies in two cases:

Case A: Tier 1 violation where a principal asserts a jurisdictional legal basis for an action that would otherwise be denied.

Case B: Tier 0-B action within an active PCR scope, where the human principal must cite the PCR authority for each approval.

APPROVE_WITH_LEGAL_BASIS carries the following legal_basis block:

```
legal_basis: {
  authority_type:  "COURT_ORDER" | "STATUTORY" |
                   "REGULATORY" | "TREATY" | "PCR",
  authority_ref:   string,   // Legal citation.  REQUIRED.
  pcr_id:          string,   // Required when authority_type
                             // is "PCR".  References the
                             // active PCR in the Clearance
                             // Registry.
  jurisdiction:    string,   // ISO 3166-1 alpha-2.  REQUIRED.
  expiry:          string,   // ISO 8601.  REQUIRED.
  document_hash:   string | null
}
```

When a principal submits APPROVE_WITH_LEGAL_BASIS, the CEE MUST:

(1) Verify that the action matches a Tier 1 prohibition or a Tier 0-B class with active PCR.
APPROVE_WITH_LEGAL_BASIS MUST NOT be accepted for Tier 0-A violations under any circumstances.

(2) Record APPROVE_WITH_LEGAL_BASIS_RECORDED in the Event Log with the full legal_basis block.

(3) Proceed to GEC execution if no Tier 0-A match is present.

APPROVE_WITH_LEGAL_BASIS is RESERVED for Tier 1 Case A in deployments where no governing CAP authority is published. For Case B (PCR-cleared Tier 0-B actions), it is operational in any PCR-configured deployment.

12.4. CAP Violation Record Schema

The GEC MUST generate a CAP Violation Record for every CEE CONSTITUTIONAL_VIOLATION output. The record MUST contain:

violation_id:	GEC-assigned UUID.
session_id:	The session in which the violation occurred.
hem_id:	The HEM event identifier, if the violation occurred during HEM_PENDING. Null otherwise.
tier:	"0A", "0B", "1", or "2".
prohibition_id:	The prohibition record that matched.
violation_type:	"AI_INITIATED" | "HUMAN_DIRECTED".
action_attempted:	The Cedar action string. Stored in the CAP Violation Record for auditors; MUST NOT be returned to the agent or application.
context_hash:	SHA-256 of the full action context. Enables auditors to reconstruct the context without the GEC exposing it in the error response.
outcome:	"REFUSED" | "SESSION_SUSPENDED" | "HEM_FIRED".
timestamp:	ISO 8601 UTC.
kernel_signature:	Ed25519 signature over canonical

serialization of all fields except kernel_signature, by the GEC keypair ([I-D.sato-soos-kia]).

12.5. Session Suspension

The GEC MAY suspend a session when a threshold of CAP violations is detected within that session. The threshold is SO Type configurable. The default threshold is three Tier 0 violations within a single session.

On session suspension:

(1) The GEC records SESSION_CAP_SUSPENDED in the Event Log.
(2) The GEC returns SESSION_SUSPENDED to the CEE.
(3) No further GEC.transition() calls are accepted for this session until an operator with appropriate authority releases the suspension.
(4) A CRITICAL Audit Alert is generated.

Suspension defends against systematic probing of the Constitutional Layer by a compromised or adversarial agent.

13. Jurisdictional Conflict Resolution

13.1. Conflict Detection

A Jurisdictional Conflict exists when:

- The SO Type has declared two or more jurisdictions.
- The CEE determines that one jurisdiction's Tier 1 prohibitions prohibit an action that another jurisdiction's Tier 1 records permit or do not address.
- The conflict_resolution method is "HEM".

The GEC MUST detect conflicts at CEE evaluation time, not at SO Type registration time. Conflicts are action-specific.

13.2. Conflict Resolution Methods

MOST_PROTECTIVE:
The GEC applies the most restrictive prohibition across all declared jurisdictions. If any jurisdiction prohibits the action, the CEE returns TIER_1_DENY. No HEM event fires. This is the safest default for multi-jurisdiction deployments.

PRIMARY_JURISDICTION:
The primary_jurisdiction's Tier 1 prohibition position governs. Secondary jurisdiction conflicts are recorded in the Event Log as CAP_TIER1_CONFLICT_DETECTED but do not block execution. Legal counsel SHOULD review the authority_ref before this method is selected.

HEM:
The GEC cannot resolve the conflict algorithmically. The GEC fires HEM_JURISDICTIONAL_CONFLICT (Class 5) and routes the conflict to the designation chain for human resolution.

13.3. HEM_JURISDICTIONAL_CONFLICT

HEM_JURISDICTIONAL_CONFLICT is HEM Class 5 as specified in [I-D.sato-soos-hem] Section 6.5. The jurisdictional_conflict_ summary field in the HEM Escalation Request MUST contain:

conflict_id:

GEC-assigned UUID for this conflict instance.

action:
The Cedar action string that triggered the conflict.

conflicting_jurisdictions:
Array of objects, one per conflicting jurisdiction:
jurisdiction: ISO 3166-1 alpha-2.
prohibition_id: The Tier 1 prohibition record that applies.
position: "PROHIBITS" | "PERMITS" | "NOT_ADDRESSED".

resolution_options:
Non-normative array of resolution approaches the principal MAY consider. The GEC MUST NOT pre-select or recommend a resolution.

Decision type constraints for Class 5: APPROVE and APPROVE_WITH_CONSTRAINTS are prohibited. APPROVE_WITH_LEGAL_BASIS (reserved for Tier 1 Case A), REDIRECT (cleanup only), TERMINATE, and DEFER are permitted.

13.4. Jurisdictional Conflict Record Schema

The GEC MUST generate a Jurisdictional Conflict Record for every detected conflict. The record MUST contain:

```
conflict_id:              GEC-assigned UUID.
session_id:               The session in which the conflict occurred.
action:                   The Cedar action string.
conflicting_jurisdictions: { jurisdiction, prohibition_id, outcome }
resolution_method:        The conflict_resolution method declared.
hem_id:                   HEM event ID if conflict_resolution is
                          "HEM". Null otherwise.
timestamp:                ISO 8601 UTC.
```

14. Event Log Requirements

CAP introduces the following Event Log entry types. All entries are appended to the GEC Event Log and signed by the GEC keypair per [I-D.sato-soos-kia].

CAP_VIOLATION_DETECTED:
Generated when the CEE returns CONSTITUTIONAL_VIOLATION on an AI-initiated action. Fields: violation_id, session_id, hem_id, tier, prohibition_id, action_attempted, context_hash, outcome, timestamp, kernel_signature.

CAP_HUMAN_VIOLATION_DETECTED:
Generated when the CEE returns CONSTITUTIONAL_VIOLATION on a human principal decision. Same fields as CAP_VIOLATION_DETECTED plus principal_id and decision_type.

CAP_PCR_CLEARANCE_APPLIED:
Generated when the CEE encounters a Tier 0-B or Tier 1 match and applies an active PCR. Fields: session_id, pcr_id, prohibition_class, action, timestamp, kernel_signature.

CAP_TIER1_CONFLICT_DETECTED:
Generated when a Tier 1 conflict is detected regardless of resolution method. Fields: conflict_id, session_id, action, conflicting_jurisdictions, resolution_method, hem_id, timestamp.

APPROVE_WITH_LEGAL_BASIS_RECORDED:
Generated when a principal submits APPROVE_WITH_LEGAL_BASIS. Fields: hem_id, principal_id, legal_basis (authority_type,

authority_ref, pcr_id, jurisdiction, expiry, document_hash), timestamp.

SESSION_CAP_SUSPENDED:
Generated when the GEC suspends a session under Section 11.5. Fields: session_id, violation_id, violation_count, threshold_applied, suspended_at.

CAP_AMBIGUITY_ROUTED:
Generated when the CEE returns LEGAL_AMBIGUITY_DETECTED. Fields: session_id, prohibition_class, ambiguity_flag, ambiguity_context, action, hem_id, timestamp, kernel_signature.

CAP_AMBIGUITY_RESOLVED:
Generated when a human principal resolves a LEGAL_AMBIGUITY HEM escalation. Fields: hem_id, session_id, principal_id, decision_type, legal_basis (if cited), determination_text (plain language statement of the principal's legal determination), timestamp, kernel_signature.
The determination_text field provides the legal engineer with evidence to refine the action_pattern in a future Regulation Record revision.

PCR_EXPIRED:
Generated when a PCR passes its expiry_date. Fields: pcr_id, prohibition_class, expired_at, operator_notified.

15. EU AI Act Applicability

15.1. Article 5 Mapping

EU AI Act Article 5 prohibits certain AI practices absolutely for EU-jurisdiction deployments. The following table maps Article 5 provisions to CAP mechanisms. This mapping is normative for EU-jurisdiction SO Type deployments: the CEE evaluation and CAP Violation Record satisfy Article 5 enforcement requirements when the relevant Tier 1 prohibitions are declared for EU-jurisdiction SO Types.

Article 5 Provision	CAP Mechanism	Sec.
5(1)(a) -- Subliminal manipulation causing harm	Tier 1: FINANCIAL_CRIME / FRAUD for EU jurisdiction; CEE TIER_1_DENY; principal requires APPROVE_WITH_LEGAL_BASIS to proceed	8.2
5(1)(b) -- Exploitation of specific group vulnerabilities	Tier 1: HUMAN_RIGHTS for EU jurisdiction; TIER_1_DENY	8.2
5(1)(c) -- Social scoring by public authorities	Tier 1: DATA_PROTECTION / PRIVACY_VIOLATION for EU; operators SHOULD also declare Tier 2 for private deployments	8.2
5(1)(d) -- Real-time remote biometric ID in public spaces	Tier 1: DATA_PROTECTION / PRIVACY_VIOLATION for EU with authority_ref citing Art. 5(1)(d)	8.2
General -- No human override of prohibited practices	HEM_HUMAN_DECISION_CONSTITUTIONAL_VIOLATION -- principal APPROVE cannot	11.2

	execute Tier 0 or verified Tier 1 actions		
General -- Audit trail of prohibited practice attempts	CAP_VIOLATION_DETECTED and CAP_HUMAN_VIOLATION_DETECTED Event Log entries; CRITICAL Audit Alert; Audit Package available to regulators	12	

Table 1: EU AI Act Article 5 Mapping

Note: Article 5 prohibited practices that derive from near-universal treaty consensus may be reclassified to Tier 0-B in future revisions of this specification. The current Tier 1 classification reflects the principle that Tier 0 is reserved for genuine global consensus (Section 7.1). The enforcement consequence is equivalent for EU-jurisdiction SO Types.

16. Security Considerations

CEE GEC integrity:
The CEE is a GEC-resident component. Its prohibition records and evaluation logic MUST be protected against modification by any agent, application, or principal. Implementations MUST treat CEE records with the same integrity protection as mandate JWTs and Event Log entries. Any modification to Tier 0 records outside of a GEC initialization sequence loaded from a verified source MUST be treated as a critical security incident. At L3 conformance, CEE integrity is guaranteed by hardware attestation ([I-D.sato-soos-kia]). At L1 and L2, integrity depends on process isolation and Event Log non-suppressibility.

Action pattern probing:
The CEE MUST NOT return the matched action_pattern in error responses. Returning full action pattern details enables adversarial agents to probe the boundary of prohibited action classes and identify closely adjacent actions. The error response MUST contain only the prohibition_class.

Tier 1 verification integrity:
Unverified Tier 1 records MUST NOT be enforced. The GEC MUST check the verified_by signature before loading any Tier 1 record.

Session suspension threshold:
The default threshold of three Tier 0 violations SHOULD be configurable downward but not upward. A higher threshold enables systematic probing. Operators who configure a higher threshold MUST document the justification and subject it to Audit Principal review.

PCR integrity:
A PCR without a valid audit_principal_signature MUST NOT be loaded. A PCR naming CSAM or GENOCIDE_FACILITATION MUST be rejected. An expired PCR MUST NOT be applied. The GEC MUST verify all three conditions at initialization and at CEE evaluation time.

PCR scope creep:
The purpose_scope field in a PCR is human-readable and not technically enforced at the CEE layer. Operators MUST ensure that the purpose_scope accurately describes the authorized activity. Audit Principals MUST verify that the purpose_scope is consistent with the cited authority_ref before signing.

Regulators reviewing the Audit Package can assess purpose compliance independently.

Legal basis citation integrity:
The APPROVE_WITH_LEGAL_BASIS legal_basis block is an operator assertion, not a verified legal finding. Implementations MUST record all such decisions in the Audit Package regardless of the apparent validity of the legal_basis. Regulators reviewing the Audit Package can assess legal basis validity independently.

Double-evaluation atomicity:
The two CEE evaluations -- before Cedar and before GEC execution -- MUST be atomic with their respective triggering calls at all three conformance levels. An implementation that evaluates CAP asynchronously or conditionally does not satisfy the Constitutional Layer guarantee.

17. IANA Considerations

17.1. CAP Prohibition Classes Tier 0 Registry

This document establishes the "Constitutional AI Protocol Prohibition Classes Tier 0" registry maintained at: https://www.iana.org/assignments/cap-prohibition-classes-tier0

Registration procedure: RFC Only.

Initial values:

Prohibition Class	Sub	Treaty Basis
CSAM	TIER0A	UN CRC 1989 + Optional Protocol
GENOCIDE_FACILITATION	TIER0A	UN Genocide Convention 1948
HUMAN_TRAFFICKING	TIER0B	UN Trafficking Protocol 2000
WMD_ASSISTANCE	TIER0B	CWC (193), BWC (183), NPT (191)
TORTURE_FACILITATION	TIER0B	UN CAT 1984 (173 states)
TERRORIST_FINANCING	TIER0B	UNSC Resolution 1373 (2001)

Table 2: Initial CAP Tier 0 Prohibition Classes

17.2. CAP Prohibition Classes Tier 1 Registry

This document establishes the "Constitutional AI Protocol Prohibition Classes Tier 1" registry maintained at: https://www.iana.org/assignments/cap-prohibition-classes-tier1

Registration procedure: Specification Required.

Initial values: FINANCIAL_CRIME, DATA_PROTECTION, CRITICAL_INFRASTRUCTURE, SECURITIES_LAW, PRIVACY_VIOLATION, FRAUD, COMPETITION_LAW, HUMAN_RIGHTS. (See Section 8.2.)

17.3. CAP Conflict Resolution Methods Registry

This document establishes the "Constitutional AI Protocol Conflict Resolution Methods" registry maintained at: https://www.iana.org/assignments/cap-conflict-resolution-methods

Registration procedure: Standards Action.

Initial values: MOST_PROTECTIVE, PRIMARY_JURISDICTION, HEM.

17.4. CAP Deployment Context Registry

This document establishes the "Constitutional AI Protocol Deployment Context" registry maintained at: https://www.iana.org/assignments/cap-deployment-context

Registration procedure: Specification Required.

Initial values: COMMERCIAL, GOVERNMENT_CIVILIAN, GOVERNMENT_DEFENSE, LAW_ENFORCEMENT, ACADEMIC_RESEARCH, REGULATED_PROFESSIONAL.

(See Section 10.3 for definitions.)

18. References

18.1. Normative References

[I-D.sato-soos-cap-rrs]
Sato, T., "Constitutional AI Protocol -- Regulation Record Specification (CAP-RRS)", Work in Progress, Internet-Draft, draft-sato-soos-cap-rrs-00, May 2026, <https://datatracker.ietf.org/doc/draft-sato-soos-cap-rrs/>.

[I-D.sato-soos-hem]
Sato, T., "The Human Escalation Mechanism (HEM) for Agentic AI Systems", Work in Progress, Internet-Draft, draft-sato-soos-hem-03, May 2026, <https://datatracker.ietf.org/doc/draft-sato-soos-hem/>.

[I-D.sato-soos-idp]
Sato, T., "The Intent Declaration Primitive (IDP) for Agentic AI Systems", Work in Progress, Internet-Draft, draft-sato-soos-idp-03, May 2026, <https://datatracker.ietf.org/doc/draft-sato-soos-idp/>.

[I-D.sato-soos-kia]
Sato, T., "Kernel Identity and Attestation", Work in Progress, Internet-Draft, draft-sato-soos-kia-00, May 2026, <https://datatracker.ietf.org/doc/draft-sato-soos-kia/>.

[RFC2119] Bradner, S., "Key words for use in RFCs to Indicate Requirement Levels", BCP 14, RFC 2119, March 1997, <https://www.rfc-editor.org/rfc/rfc2119>.

[RFC8174] Leiba, B., "Ambiguity of Uppercase vs Lowercase in RFC 2119 Key Words", BCP 14, RFC 8174, May 2017, <https://www.rfc-editor.org/rfc/rfc8174>.

18.2. Informative References

[I-D.sato-soos-gar]
Sato, T., "The Governance Audit Record (GAR) for Agentic AI Systems", Work in Progress, Internet-Draft, draft-sato-soos-gar-01, May 2026, <https://datatracker.ietf.org/doc/draft-sato-soos-gar/>.

[EU-AI-ACT]
European Parliament and Council, "Regulation (EU) 2024/1689", OJ L 2024/1689, July 2024.

[UN-GENOCIDE]
United Nations, "Convention on the Prevention and Punishment of the Crime of Genocide", 78 UNTS 277, 1948.

[UN-CRC] United Nations, "Convention on the Rights of the Child", 1577 UNTS 3, 1989.

[UN-TIP] United Nations, "Protocol to Prevent, Suppress and Punish Trafficking in Persons", 2237 UNTS 319, 2000.

[CWC] OPCW, "Convention on the Prohibition of Chemical Weapons", 1975 UNTS 45, 1993.

[UNSC-1373] UN Security Council, "Resolution 1373 (2001)", S/RES/1373, September 2001.

[UN-CAT] United Nations, "Convention Against Torture", 1465 UNTS 85, 1984.

Appendix A. Worked Example -- A Travel Booking in Two Jurisdictions

This appendix walks through two scenarios using a Japan-based travel operator (MyAuberge K.K.) running an ActivityBookingObject SO with jurisdiction configuration: primary JP, secondary EU, conflict_resolution: MOST_PROTECTIVE.

A.1. Scenario 1: Lawful Action, Fully Traced

A travel agent AI requests an action to process a booking payment for a guest who has explicitly consented to data processing.

Step 1: The AI submits the action to GEC.transition().

Step 2: The CEE evaluates. The action matches no Tier 0 record. It matches the EU Tier 1 DATA_PROTECTION prohibition record -- but the action pattern carries a user_consent flag. The operator has declared a Cedar policy that permits payment processing with explicit consent. The CEE finds no Tier 1 match for this specific action pattern. Returns PERMIT.

Step 3: Cedar evaluates. The Cedar policy PERMITs the action. No HEM fires. The GEC executes the transition.

Step 4: The STATE_TRANSITION Event Log entry is written. The IDP records the agent's reasoning. The audit trail is complete.

The lawyer reviewing this system sees: a fully traced, consent-documented payment action. Every step is in the log.

A.2. Scenario 2: Jurisdictional Conflict, Human Resolution

The same AI requests an action to share guest location data with a third-party logistics provider. The JP Tier 1 configuration permits this under APPI (Japan's data protection law) with appropriate notice. The EU Tier 1 configuration prohibits it under GDPR Article 44 (transfer to third countries) because the logistics provider is outside the EEA.

Step 1: The AI submits the action to GEC.transition().

Step 2: The CEE evaluates. Tier 0: no match. Tier 1: JP record -- PERMIT. EU record -- PROHIBITS. Conflict detected.

Step 3: conflict_resolution is MOST_PROTECTIVE. The CEE returns TIER_1_DENY. The action is denied. CAP_TIER1_CONFLICT_DETECTED is written to the Event Log.

Step 4: The agent receives a DENY response. The IDP records the denial. The RETRY_CONTINUATION path (if declared) requires the agent to articulate what changed before retrying.

Step 5: The operator, reviewing the conflict log, contacts legal counsel. Counsel advises that an adequacy decision or appropriate safeguards would permit the transfer under GDPR. The operator updates the EU Tier 1 record to include the adequacy decision citation. The updated record is verified by the Audit Principal. The action is now permitted.

Step 6: The full resolution trace -- denial, conflict record, legal review, Tier 1 update, re-authorization -- is in the Event Log and available to the regulator.

Appendix B. Related Work

B.1. Existing Constitutional AI Frameworks

Constitutional AI as a training technique (Anthropic, 2022) uses a set of principles to guide model behavior during RLHF. This is a training-time approach. CAP is an inference-time enforcement approach: the GEC evaluates every action request against a prohibition set regardless of what the model would do if allowed to act.

The distinction is architecturally critical: a model trained on constitutional principles can still produce outputs that violate those principles under adversarial prompting or out-of-distribution inputs. A CAP-governed GEC refuses Tier 0-A actions unconditionally, regardless of the model’s output, because the CEE evaluates the action pattern against the prohibition registry -- not the model’s compliance with its training.

CAP and Constitutional AI training are complementary. CAP does not replace safety training; it provides an enforcement layer that does not depend on the model’s training having succeeded.

B.2. EU AI Act Article 5

EU AI Act Article 5 defines prohibited AI practices for EU-jurisdiction deployments enforced by regulatory action after the fact. CAP Tier 1 prohibition records for EU jurisdiction are enforced at GEC evaluation time, before the action occurs. CAP makes Article 5 machine-executable rather than post-hoc regulatory.

B.3. AIPREF

The AIPREF Working Group defines a vocabulary for AI content-use preferences. AIPREF provides the policy expression layer; CAP provides the constitutional floor below which no AIPREF preference can descend. An AIPREF preference that permitted CSAM would be refused by CAP Tier 0-A regardless of the preference content.

B.4. SOOS Companion Drafts

CAP sits above all other SOOS layers in the enforcement stack:

draft-sato-soos-kia-00: CAP Violation Records are signed by the GEC keypair. PCRs are loaded in the GEC Manifest. At L3, the CEE runs in a KIA-attested TEE.

draft-sato-soos-idp-03: CAP evaluates before IDP is submitted.

draft-sato-soos-cap-rrs-00: Companion specification defining

the Regulation Record schema, Cedar Compilation Profile, and Constitutional Mandate Registry. Required for Tier 3 and for automated Cedar policy generation from legal sources. A Tier 0-A refusal prevents IDP creation. Defines GEC conformance levels (L1/L2/L3) referenced in Section 6.1.

draft-sato-soos-hem-03: CAP evaluates human HEM decisions before GEC execution (double-evaluation). HEM_JURISDICTIONAL_CONFLICT (Class 5) is a joint CAP+HEM event.

draft-sato-soos-gar-01: CAP Violation Records are included in the GAR Audit Package for regulatory inspection. CRITICAL Audit Alerts from CEE feed the GAR alert pipeline.

draft-sato-soos-aep-00: The AEP ACT step invokes GEC.transition(), which triggers CEE before Cedar.

draft-sato-soos-mad-01: CAP applies to all multi-agent topologies. An orchestrator cannot issue a sub-agent mandate that bypasses CAP; the CEE is evaluated per transition regardless of delegation depth.

Author's Address

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