Internet-Draft
Intended status: Standards Track
Expires: November 17, 2026

T. Sato
MyAuberge K.K.

May 17, 2026

The Constitutional AI Protocol (CAP) for Agentic AI Systems
draft-sato-soos-cap-00

Abstract

This document specifies the Constitutional AI Protocol (CAP), a kernel-enforced prohibition architecture for agentic AI systems. CAP defines a three-tier prohibition model: Tier 0 absolute prohibitions derived from near-universal treaty consensus, Tier 1 jurisdiction-specific prohibitions declared by operators and verified by auditors, and Tier 2 voluntary operator ethical standards. CAP evaluates every action request twice -- once on the AI agent's action before Cedar policy evaluation, and once on the human principal's decision before kernel execution -- ensuring that neither agents nor human principals can authorize absolutely prohibited actions. This document also specifies Jurisdictional Conflict Resolution (JCR), the mechanism by which the kernel surfaces irreconcilable jurisdictional conflicts to human principals via the Human Escalation Mechanism [I-D.sato-soos-hem].

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 November 17, 2026.

Copyright Notice

Table of Contents

1. Introduction

Agentic AI systems operate under authorization frameworks that determine what agents are permitted to do. These frameworks -- Cedar policy evaluation, mandate JWT scope, human escalation decisions -- are powerful and flexible. Their flexibility is also their limitation: they can be configured to authorize actions that are harmful, unlawful, or both.

The Human Escalation Mechanism [I-D.sato-soos-hem] stops the AI and waits for a human principal to decide. HEM is a necessary governance layer. It is not sufficient on its own, because HEM has no mechanism to evaluate whether a human principal’s decision is itself lawful. A human principal can issue an APPROVE decision on a market manipulation action. HEM executes it. The human-AI system has committed a crime.

The Constitutional AI Protocol (CAP) closes this gap. CAP defines a Constitutional Layer that sits above all principal authority and evaluates action requests twice:

- When the AI agent requests an action -- before Cedar policy evaluation. An action that violates a Tier 0 absolute prohibition is refused unconditionally. Cedar is not consulted. HEM does not fire. No principal can override the refusal.

- When a human principal submits a decision -- before the kernel executes the decision. An APPROVE or APPROVE_WITH_CONSTRAINTS decision on a prohibited action is refused. The principal's decision slot is not consumed; they may submit a revised decision.

CAP's primary value is not prohibition -- it is transparency and traceability of legal authorization. Most deployed agentic AI 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 government claims its actions are lawful. CAP says: prove it. Cite the authority. It goes in the log.

This document specifies:

- The Constitutional Evaluation Engine (CEE) -- the kernel component that evaluates action requests against the three-tier prohibition model.

- Tier 0 -- Universal Core Prohibitions -- six absolute categories derived from near-universal treaty consensus, kernel-resident, immutable without a new RFC.

- Tier 1 -- Jurisdictional Prohibition Layer -- operator-declared, jurisdiction-specific prohibitions verified by Audit Principals.

- Tier 2 -- Operator Ethical Layer -- voluntary operator standards exceeding local law, publicly disclosed.

- Jurisdictional Conflict Resolution (JCR) -- the mechanism by which irreconcilable jurisdictional conflicts are surfaced to human principals via HEM Class 5 (HEM_JURISDICTIONAL_CONFLICT).

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

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.

The following terms are defined in this document or inherited from [I-D.sato-soos-idp] and [I-D.sato-soos-hem]:

Constitutional Evaluation Engine (CEE):
: The kernel 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 kernel execution of a human decision.

Constitutional Layer:
: The enforcement boundary above all principal authority. No agent, operator, or human principal can override a Tier 0 Constitutional Layer refusal.

CAP Violation:
: An action request or human principal decision that the CEE determines violates a Tier 0, Tier 1, or Tier 2 prohibition.

APPROVE_WITH_LEGAL_BASIS:
A HEM decision sub-type reserved for Tier 1 violations where a principal asserts a jurisdictional legal basis. Operational only when a CAP specification is the governing authority. Defined in Section 8.3.

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

HEM_JURISDICTIONAL_CONFLICT:
HEM Class 5 trigger. Fires when the kernel detects a Jurisdictional Conflict that cannot be algorithmically resolved. Specified in [I-D.sato-soos-hem] Section 5.5.

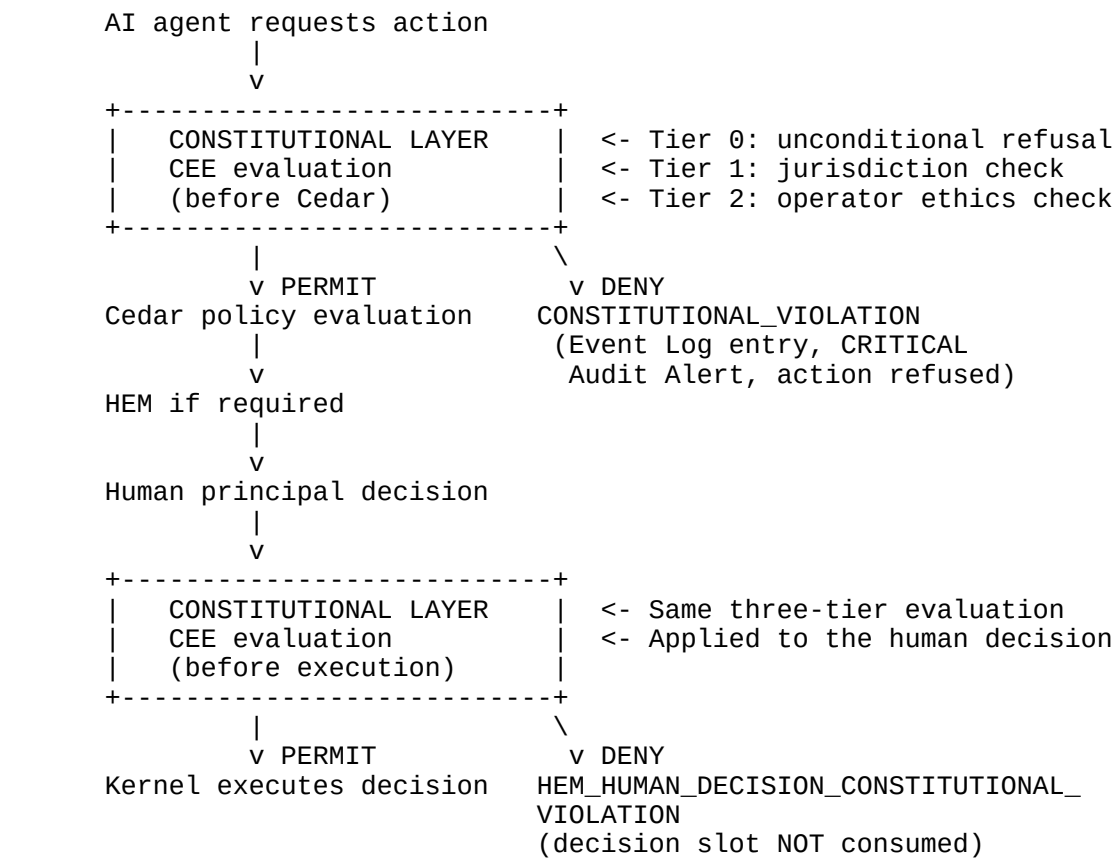
Verified External Auditor:
Defined in [I-D.sato-soos-gar]. An external party with time-limited, scope-limited read access to kernel audit artifacts.

Action Pattern:
A structured description of the class of kernel actions a prohibition covers. Action patterns are expressed using the same vocabulary as Cedar policy action sets, enabling the CEE to match incoming action requests against prohibition records without natural language interpretation.

3. Architecture Overview

3.1. The Double-Evaluation Property

CAP introduces a Constitutional Layer that evaluates every governed action twice. The evaluation sequence for any AI-initiated action is:



The double-evaluation property ensures that:

- An AI agent cannot request an absolutely prohibited action even if its Cedar policy would permit it.
- A human principal cannot authorize an absolutely prohibited action even with an APPROVE decision.
- The Constitutional Layer is evaluated by the kernel, not by the AI agent or any application layer.

3.2. 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.

- HEM_CAP_VIOLATION is a CAP event, not a HEM trigger class. Tier 0 violations are refused before HEM state machinery is engaged. The kernel records CONSTITUTIONAL_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 kernel 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. It is reserved in this specification and becomes operational only when CAP is published as a normative authority.

3.3. Relationship to Cedar Policy Evaluation

Cedar policy evaluation is the kernel’s authorization layer for normal governed actions. CAP is above Cedar. The evaluation order is: CAP (Constitutional Layer) -> Cedar -> HEM (if Cedar routes) -> Human decision -> CAP (Constitutional Layer again).

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.

4. Constitutional Evaluation Engine (CEE)

4.1. CEE Placement in the Kernel

The CEE is a kernel-resident component. It MUST be invoked:

- On every kernel.transition() call, before Cedar evaluation.
- On every HEM decision submission via the Decision Submission Protocol (Section 7.6 of [I-D.sato-soos-hem]), before the kernel 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.

4.2. CEE Evaluation Protocol

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

(1) Evaluate against all loaded Tier 0 prohibition records.
    If any Tier 0 record matches the action pattern:
    DENY unconditionally. Record CAP_VIOLATION_DETECTED (AI) or
    CAP_HUMAN_VIOLATION_DETECTED (human decision) in the Event Log.
    Return CONSTITUTIONAL_VIOLATION to the kernel. Do not proceed
    to Tier 1 or Tier 2 evaluation.

(2) If no Tier 0 match, evaluate against all loaded Tier 1
    prohibition records for the declared jurisdiction(s).
    If any Tier 1 record matches:
    Check conflict_resolution configuration.
    If conflict_resolution: "MOST_PROTECTIVE" or
    "PRIMARY_JURISDICTION": apply resolution and return DENY or
    PERMIT accordingly.
    If conflict_resolution: "HEM": fire HEM_JURISDICTIONAL_CONFLICT
    (Class 5).

(3) If no Tier 1 match or Tier 1 PERMIT, evaluate against all
    loaded Tier 2 prohibition records.
    If any Tier 2 record matches: return DENY with
    violation_type: TIER_2.
    Tier 2 denials MAY be overridden by operator configuration
    at the SO Type level. Tier 2 denials MUST be logged.

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

4.3. CEE Outputs

The CEE returns one of the following to the kernel:

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

CONSTITUTIONAL_VIOLATION:
: Tier 0 match. Action unconditionally refused. Kernel MUST record CAP_VIOLATION_DETECTED or CAP_HUMAN_VIOLATION_DETECTED. Kernel MUST generate CRITICAL Audit Alert.

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

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

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

SESSION_SUSPEND:
: Reserved for repeated Tier 0 violations within a session. See Section 8.5.

5. Tier 0 -- Universal Core Prohibitions

5.1. Tier 0 Properties

Tier 0 prohibitions are:

- Kernel-resident: loaded at kernel initialization, not at session open or SO Type registration.

- Unconditional: no Cedar policy, mandate configuration, operator declaration, or human principal decision can override a Tier 0 refusal.

- Immutable without a new RFC: the Tier 0 registry has a registration procedure of "RFC Only" (Section 13.1). No operator, regulator, or standards body other than the IETF can modify the Tier 0 set.

- Globally scoped: Tier 0 prohibitions 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. Categories for which genuine universal consensus does not exist are NOT in Tier 0.

5.2. Tier 0 Prohibition Classes

The initial Tier 0 prohibition classes are:

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.

CSAM:
: Actions that produce, distribute, or facilitate 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.

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.

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).

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.

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.

5.3. Tier 0 Prohibition Schema

Each Tier 0 prohibition record MUST contain the following fields:

prohibition_id:
: Unique identifier for this prohibition record.

prohibition_class:
: One of the Tier 0 prohibition classes registered in the CAP

Prohibition Classes Tier 0 registry (Section 13.1).

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 to enable deterministic CEE matching.

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

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

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

5.4. 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 13.1).

Implementations MUST NOT expose any configuration interface that allows Tier 0 records to be modified, disabled, or overridden at deployment time.

6. Tier 1 -- Jurisdictional Prohibition Layer

6.1. 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. The kernel SHOULD warn Audit Principals when review_date is exceeded. Expired Tier 1 records remain in force until updated or explicitly retired; expiry generates a PRD_REVIEW_DATE_EXCEEDED Audit Alert.
- Signed: Tier 1 records MUST be signed by both the declaring operator (declared_by) and a Verified Audit Principal (verified_by).

6.2. 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.

6.3. Tier 1 Prohibition Schema

Each Tier 1 prohibition record MUST contain the following fields:

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 13.2).

jurisdiction:
ISO 3166-1 alpha-2 country code for the jurisdiction in which this prohibition applies. 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 prohibition record must be reviewed. REQUIRED.

declared_by:
Identifier of the operator declaring this prohibition.

verified_by:
Identifier of the Audit Principal who verified this prohibition

record. REQUIRED before enforcement. Null until verified.

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

6.4. 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 governing this SO Type. REQUIRED.

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

conflict_resolution:
Controlled vocabulary. Specifies how the kernel resolves conflicts between prohibitions from different jurisdictions. One of:

MOST_PROTECTIVE:
The most restrictive prohibition across all declared jurisdictions applies. If any jurisdiction prohibits an action, it is denied.

PRIMARY_JURISDICTION:
The primary_jurisdiction's prohibition position governs. Secondary jurisdiction prohibitions are informative only.

HEM:
The kernel cannot resolve jurisdictional conflicts algorithmically. When a conflict is detected, the kernel fires HEM_JURISDICTIONAL_CONFLICT (Class 5) and routes the conflict to a human principal for resolution.

conflict_escalation:
Specifies kernel behavior when HEM_JURISDICTIONAL_CONFLICT cannot be resolved (chain exhausted). One of:
HEM: chain exhaustion disposition per [I-D.sato-soos-hem] Section 9.4.
SUSPEND: kernel suspends the session pending operator intervention.

legal_counsel_ref:
Reference to the legal counsel who reviewed the Jurisdiction Configuration. RECOMMENDED.

declared_at:
ISO 8601 UTC timestamp of declaration.

declared_by:
Identifier of the operator.

6.5. 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 kernel MUST NOT enforce any Tier 1 record with a null or unverifiable verified_by field.

7. Tier 2 -- Operator Ethical Layer

7.1. 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 public disclosure.
- 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.

7.2. Tier 2 Prohibition Schema

Each Tier 2 prohibition record MUST contain the following fields:

prohibition_id:
: Unique identifier.

prohibition_class:
: Free text or reference to an operator-defined taxonomy. Not registered in an IANA registry.

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. True if this prohibition has been publicly disclosed in a transparency report or equivalent.

7.3. 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].

8. CAP Violation Handling

8.1. AI-Initiated Violations

When the CEE returns CONSTITUTIONAL_VIOLATION on an AI-initiated action request, the kernel 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 10).

(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 kernel.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 (to prevent pattern probing).

The kernel MUST NOT reveal which specific Tier 0 record matched beyond the prohibition_class. Action pattern details are available only to Verified External Auditors via the Audit Package.

8.2. Human-Directed Violations

When the CEE returns CONSTITUTIONAL_VIOLATION on a human principal decision submission, the kernel 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 (Section 10).

(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 (step 2) 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.

8.3. APPROVE_WITH_LEGAL_BASIS

APPROVE_WITH_LEGAL_BASIS is a HEM decision sub-type introduced by this specification for Tier 1 violations where a principal asserts a jurisdictional legal basis for an action that would otherwise be prohibited.

APPROVE_WITH_LEGAL_BASIS carries the following legal_basis block:

```
legal_basis: {
  authority_type:  "COURT_ORDER" | "STATUTORY" |
                   "REGULATORY" | "TREATY",
  authority_ref:   string,  // Legal citation. REQUIRED.
```

```
    jurisdiction:    string,  // ISO 3166-1 alpha-2. REQUIRED.
    expiry:          string,  // ISO 8601. REQUIRED.
    document_hash:   string | null  // Hash of legal document.
  }
```

When a principal submits APPROVE_WITH_LEGAL_BASIS, the CEE MUST:

(1) Verify that the action matches a Tier 1 prohibition (not Tier 0). APPROVE_WITH_LEGAL_BASIS MUST NOT be accepted for Tier 0 violations under any circumstances. The kernel MUST return HEM_HUMAN_DECISION_CONSTITUTIONAL_VIOLATION if a Tier 0 match is present regardless of the legal_basis content.

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

(3) Proceed to kernel execution if no Tier 0 match is present.

CAP's purpose is not to prevent lawful actions -- it is to make lawful actions legally traceable. APPROVE_WITH_LEGAL_BASIS provides the traceability mechanism: the authority citation goes in the Event Log and the Audit Package. It is available for regulatory inspection at any time.

APPROVE_WITH_LEGAL_BASIS is RESERVED in this specification. Implementations MUST NOT accept APPROVE_WITH_LEGAL_BASIS decisions until a governing CAP authority is published. The kernel MUST return HEM_DECISION_TYPE_NOT_YET_OPERATIONAL if an APPROVE_WITH_LEGAL_BASIS decision is submitted prior to that publication.

8.4. CAP Violation Record Schema

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

violation_id:
: Kernel-generated UUID.

session_id:
: The session in which the violation was detected.

hem_id:
: The HEM event identifier, if the violation occurred during a HEM_PENDING state. Null otherwise.

tier:
: 0, 1, or 2.

prohibition_id:
: The prohibition record that matched.

violation_type:
: AI_INITIATED | HUMAN_DIRECTED.

action_attempted:
: The Cedar action string of the attempted action. Stored in the CAP Violation Record for auditors; MUST NOT be returned to the agent or application surface.

context_hash:
: SHA-256 hash of the full action context at the time of detection. Enables auditors to reconstruct the context without the kernel 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.

8.5. Session Suspension

The kernel 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 kernel records SESSION_CAP_SUSPENDED in the Event Log.

(2) The kernel returns SESSION_SUSPENDED to the CEE.

(3) No further kernel.transition() calls are accepted for this session until an operator with appropriate authority releases the suspension.

(4) A CRITICAL Audit Alert is generated (alert_trigger: CAP_VIOLATION_DETECTED, outcome: SESSION_SUSPENDED).

Suspension is the kernel’s defense against systematic probing of the Constitutional Layer by a compromised or adversarial agent.

9. Jurisdictional Conflict Resolution

9.1. Conflict Detection

A Jurisdictional Conflict exists when:

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

The kernel MUST detect Jurisdictional Conflicts at CEE evaluation time, not at SO Type registration time. Conflicts are action-specific: a Jurisdiction Configuration may produce no conflicts for most actions and a conflict for a specific class of action.

9.2. Conflict Resolution Methods

The three conflict resolution methods declared in the Jurisdiction Configuration determine how the kernel responds to a detected conflict:

MOST_PROTECTIVE:
   The kernel 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 for the primary jurisdiction before this method is selected.

HEM:
The kernel cannot resolve the conflict algorithmically. The kernel fires HEM_JURISDICTIONAL_CONFLICT (Class 5) and routes the conflict to the designation chain for human resolution. The HEM Escalation Request carries a jurisdictional_conflict_ summary field with the conflicting jurisdictions, their respective Tier 1 prohibition classes, and the action that triggered the conflict.

9.3. HEM_JURISDICTIONAL_CONFLICT

HEM_JURISDICTIONAL_CONFLICT is HEM Class 5 as specified in [I-D.sato-soos-hem] Section 5.5. This section specifies the CAP-side content that the kernel MUST include in the HEM Escalation Request when Class 5 fires.

The jurisdictional_conflict_summary field in the HEM Escalation Request MUST contain:

conflict_id:
Kernel-generated UUID for this conflict instance.

action:
The Cedar action string that triggered the conflict.

conflicting_jurisdictions:
Array of objects, one per conflicting jurisdiction, each containing:

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 kernel MUST NOT pre-select or recommend a resolution.

Decision type constraints for Class 5 apply as specified in [I-D.sato-soos-hem]: APPROVE and APPROVE_WITH_CONSTRAINTS are prohibited. APPROVE_WITH_LEGAL_BASIS (reserved), REDIRECT (cleanup only), TERMINATE, and DEFER are permitted.

9.4. Jurisdictional Conflict Record Schema

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

conflict_id:
Kernel-generated UUID.

session_id:
The session in which the conflict was detected.

action:
The Cedar action string.

conflicting_jurisdictions:
Array of objects: { jurisdiction, prohibition_id, outcome }.

resolution_method:
The conflict_resolution method declared for this SO Type.

hem_id:
The HEM event identifier if conflict_resolution is "HEM".
Null otherwise.

timestamp:
ISO 8601 UTC.

10. Event Log Requirements

CAP introduces the following Event Log entry types. All entries are appended to the kernel Event Log specified in [I-D.sato-soos-hem] Section 10.

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. Fields: same as CAP_VIOLATION_DETECTED plus principal_id and decision_type.

CAP_TIER1_CONFLICT_DETECTED:
Generated when a Tier 1 jurisdictional 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 a valid APPROVE_WITH_LEGAL_BASIS decision. Fields: hem_id, principal_id, legal_basis (authority_type, authority_ref, jurisdiction, expiry, document_hash), timestamp.

SESSION_CAP_SUSPENDED:
Generated when the kernel suspends a session under Section 8.5. Fields: session_id, violation_id, violation_count, threshold_applied, suspended_at.

11. EU AI Act Applicability

11.1. Article 5 Mapping

EU AI Act Article 5 prohibits certain AI practices absolutely. The following table maps Article 5 provisions to CAP mechanisms. This mapping is normative: the CEE evaluation and CAP Violation Record specified in this document satisfy Article 5 requirements for deployments governed by [I-D.sato-soos-hem] when the relevant Tier 1 prohibitions are declared for EU-jurisdiction SO Types. Tier 0 universal prohibitions apply globally regardless of jurisdiction configuration. Operators may reference this section directly in EU AI Act Article 5 conformance documentation.

Article 5 Provision	CAP Mechanism	Sec.
5(1)(a) -- Subliminal manipulation causing harm	Tier 1: FINANCIAL_CRIME / FRAUD prohibition class for	6.2

		EU jurisdiction; CEE TIER_1_ DENY; no principal override without APPROVE_WITH_LEGAL_ BASIS	
5(1)(b) -- Exploitation of vulnerabilities of specific groups	Tier 1: HUMAN_RIGHTS	prohibition class for EU jurisdiction; CEE TIER_1_DENY	6.2
5(1)(c) -- Social scoring by public authorities	Tier 1: DATA_PROTECTION /	PRIVACY_VIOLATION for EU jurisdiction; operators SHOULD also declare Tier 2 prohibition for private deployments	6.2
5(1)(d) -- Real-time remote biometric identification in publicly accessible spaces	Tier 1: DATA_PROTECTION /	PRIVACY_VIOLATION for EU jurisdiction with authority_ ref citing Art. 5(1)(d)	6.2
General -- No human override of prohibited practices	HEM_HUMAN_DECISION_	CONSTITUTIONAL_VIOLATION -- principal APPROVE cannot execute Tier 0 or verified Tier 1 actions	8.2
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	10

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 in future revisions of this specification. The current Tier 1 classification for EU-specific Article 5 items reflects the principle that Tier 0 is reserved for genuine global consensus. The enforcement consequence is equivalent for EU-jurisdiction SO Types: CEE denies the action and no principal can override without a legal basis citation.

12. Security Considerations

CEE kernel integrity:
The Constitutional Evaluation Engine is a kernel-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 kernel initialization sequence loaded from a verified source MUST be treated as a critical security incident.

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 find 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 kernel MUST check the verified_by signature before loading any Tier 1

record. An operator who deploys unverified Tier 1 records creates a false assurance of prohibition coverage.

Session suspension threshold:
The default session suspension threshold of three Tier 0 violations (Section 8.5) SHOULD be configurable downward but not upward beyond the default. A higher threshold enables systematic probing. Operators who configure a higher threshold MUST document the justification and subject it to Audit Principal review.

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 APPROVE_WITH_LEGAL_BASIS 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 execution -- MUST be atomic with their respective triggering calls. An implementation that evaluates CAP asynchronously or conditionally does not satisfy the Constitutional Layer guarantee.

13. IANA Considerations

13.1. CAP Prohibition Classes Tier 0 Registry

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

Registration procedure: RFC Only. Tier 0 classes may only be added, modified, or removed by publication of an RFC that updates this document.

Initial values:

Prohibition Class	Treaty Basis
GENOCIDE_FACILITATION	UN Genocide Convention 1948 (153 states)
CSAM	UN CRC 1989 + Optional Protocol (196 states)
HUMAN_TRAFFICKING	UN Trafficking Protocol 2000 (178 states)
WMD_ASSISTANCE	CWC (193), BWC (183), NPT (191 states)
TORTURE_FACILITATION	UN CAT 1984 (173 states)
TERRORIST_FINANCING	UNSC Resolution 1373 (2001)

Table 2: Initial CAP Tier 0 Prohibition Classes

13.2. CAP Prohibition Classes Tier 1 Registry

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

Registration procedure: Specification Required.

Initial values:

Prohibition Class	Description
FINANCIAL_CRIME	Market manipulation, money laundering, and related financial offenses
DATA_PROTECTION	Personal data processing violations
CRITICAL_INFRASTRUCTURE	Actions targeting critical infrastructure
SECURITIES_LAW	Unlicensed securities activities
PRIVACY_VIOLATION	Prohibited surveillance and profiling
FRAUD	Deceptive practices under consumer or criminal law
COMPETITION_LAW	Cartel coordination and abuse of dominant position
HUMAN_RIGHTS	Violations of applicable human rights law

Table 3: Initial CAP Tier 1 Prohibition Classes

13.3. CAP Conflict Resolution Methods Registry

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

Registration procedure: Standards Action.

Initial values:

Method	Description
MOST_PROTECTIVE	Most restrictive prohibition across all declared jurisdictions applies
PRIMARY_JURISDICTION	Primary jurisdiction prohibition position governs; secondary conflicts logged only
HEM	Kernel fires HEM_JURISDICTIONAL_CONFLICT (Class 5) for human resolution

Table 4: Initial CAP Conflict Resolution Methods

14. References

14.1. Normative References

[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-00, 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-00, May 2026, <https://datatracker.ietf.org/doc/draft-sato-soos-idp/>.

[RFC2119] Bradner, S., "Key words for use in RFCs to Indicate Requirement Levels", BCP 14, RFC 2119, DOI 10.17487/RFC2119, 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, DOI 10.17487/RFC8174, May 2017,

<https://www.rfc-editor.org/rfc/rfc8174>.

14.2. 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-00, May 2026, <https://datatracker.ietf.org/doc/draft-sato-soos-gar/>.

[EU-AI-ACT]
European Parliament and Council, "Regulation (EU) 2024/1689 laying down harmonised rules on artificial intelligence", OJ L 2024/1689, July 2024, <https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=OJ:L_202401689>.

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

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

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

[CWC] Organisation for the Prohibition of Chemical Weapons, "Convention on the Prohibition of the Development, Production, Stockpiling and Use of Chemical Weapons", 1975 UNTS 45, January 1993.

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

[UN-CAT] United Nations, "Convention Against Torture and Other Cruel, Inhuman or Degrading Treatment or Punishment", 1465 UNTS 85, December 1984.

Author’s Address

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