

Agentic Incident Readiness and Response Doctrine

Incident response that contains authority, state, and human decisions for agentic and superintelligence-era AI. Built on AI SAFE² and the Cognitive Sovereignty Framework.

Baseline to meet and exceed	AI SAFE ² v3.1.0 (161 core controls, CP.1-CP.10 cross-pillar governance) and the Cognitive Sovereignty Framework v2.1.0 (six-domain human resilience model).
Reference view	NIST CAISSI, NSA SISC, and CISA, <i>Key Practices for Incident Readiness and Response in Super Intelligence Training and Evaluation</i> , PP-26-4396, v1.0, October 2026. Treated as one limited lens.
Applies to	Frontier labs training or evaluating SI agents, and every enterprise or mission owner deploying agents at ACT-2 and above.
Status	Version 1.2, issued for adoption and comment. New constructs introduced here are proposed for incorporation into AI SAFE ² CP.6 and P3.T5.10 guidance (Annex C).

An incident is not over when the agent stops.

It is over when its authority is gone.

EXECUTIVE SUMMARY

Bottom Line Up Front

In October 2026, NIST CAISSI, NSA SISC, and CISA published key practices for incident readiness in superintelligence training and evaluation. The guidance correctly treats an AI agent as a whole system: model, harness, tools, subagents, and orchestration. Its incident definition is narrow by design. It covers an agent's uninstructed actions and places prompt injection and backdoored models outside its primary class. **Operators do not get to choose which incidents arrive.**

This doctrine covers the rest, using AI SAFE² to govern the machine and the Cognitive Sovereignty Framework (CSF) to govern the human capacity to direct it. Agentic incident response needs both, at machine speed, under adversarial pressure, with evidence that does not depend on the suspect agent's testimony.

Why authority is the organizing unit. Root cause in an agentic incident can take days to establish. Authority can be revoked in minutes, whatever the cause turns out to be. Cause still drives remediation and reporting, and containment proceeds without waiting for it.

Four shifts

From	To	Why it matters
Incident = uninstructed agent action	Incident = any violated authority boundary, classified effect-first, with instruction source recorded separately and allowed to be UNKNOWN	Responders rarely know the cause at T+0. Containment proceeds on what happened.
Shutdown = containment	Every path to the prohibited effect is Revoked, Blocked, Drained, or Compensated (RBDC)	Queued orders, descendant agents, poisoned memory, and sent email survive a process kill.
Human review = oversight	A human approval is a control, investigated with the same rigor as the machine	Decision automation capture (CSF T-CT-006) defeats approvals without touching the machine.
Detect, then respond	Zero dwell by design: Class-H and unknown actions are contained before they execute	Response starts with authority already bounded, so the incident surface is what the envelope allowed.

What this doctrine delivers

- **Agentic Incident Classification:** eight classes on three axes; the federal incident is one of them.
- **Interim severity model** with hard floors, filling a federal future-work gap.
- **Five Authority Claims and the Authority Ledger**, each claim evidenced separately.
- **Eight-phase lifecycle** mapped to AI SAFE², CSF, and all ten federal practices.
- **RBDC containment** confirmed by the system holding the authority.
- **Evidence confidence labels** that keep model agreement out of the verified column.
- **Human Governability Protocol**, five reauthorization gates, ten falsification exercises, and a federal crosswalk.

Leadership: read this page, the Leadership Brief, Outcome Measures, and Adoption Path. **Practitioners:** Sections 4 through 12 and the Annexes.

FOR THE CIO, CISO, AND BOARD

Leadership Brief: Five Decisions, Five Questions

Practitioners can own most of this doctrine. **Five decisions belong to leadership**, because each one commits the organization before an incident, while the trade-off can still be made deliberately.

Decision	What leadership commits to	Where
1. Name a HEAR for every autonomous agent	Real, unilateral authority to halt each ACT-3 and ACT-4 deployment, held by a named human outside the team that operates it.	Sections 6, 14
2. Set the Class-H map and financial threshold	The actions that wait for a human signature before they execute. If no threshold is documented, every financial commitment waits.	Phase 1
3. Approve catastrophic-risk stop conditions	Conditions that suspend operations regardless of business-continuity preference (CP.8). Only leadership can pre-commit to that.	Phase 1
4. Fund an independent monitor plane	Monitoring and logging outside the reach of the agent and its product owner , so someone with no stake in the answer confirms containment.	Phase 2, Section 14
5. Set reauthorization expiry and risk acceptance	A signed risk acceptance with an end date for every restart after an incident. Default expiry is 90 days.	Section 11

Five questions to ask your team this week

Each question tests a control this doctrine requires. A team that cannot answer quickly has found its first readiness gap.

Question	What a ready answer looks like
For our highest-autonomy agent, who authorized it, what can it do, and how do we stop it?	Produced from the Authority Ledger in under five minutes , with no reconstruction from logs.
If we revoked that agent's credentials right now, what would still execute tomorrow?	A named inventory of queued work, scheduled jobs, descendant agents, and durable state.
Which of our AI boundaries exist only in prompts?	A list , with a plan to enforce each boundary outside the model.
When did we last test the kill path, and who confirmed it worked?	A test date within 30 days , confirmed by the system holding the authority.
When a manager approves an AI-drafted action, how would we know the approval was manipulated?	Approval records , out-of-band verification for consequential instructions, and override-rate monitoring.

IF YOU DO ONE THING

Ask the first question about your highest-autonomy agent. If the answer takes longer than five minutes, or arrives as a reconstruction from logs, start with Phase 1: classify, name the HEAR, and build the Authority Ledger. Every other control in this doctrine depends on it.

SECTION 1

Scope, Terms, and How to Use This Doctrine

Normative terms follow RFC 2119 usage. **MUST** is a requirement for conformance. **SHOULD** is a default that requires a documented reason to depart from. **MAY** is permitted. Throughout, "CSF" means the Cognitive Sovereignty Framework. The NIST Cybersecurity Framework is cited as "NIST CSF 2.0" where needed.

SCOPE OF CLAIMS

AI SAFE² control IDs, CSF threat IDs, and federal practice numbers are cited to the pinned versions in Annex C. New constructs (AIC classes, the Five Authority Claims, RBDC, reauthorization gates, the Human Governability Protocol) are introduced by this doctrine. **Default thresholds are conservative starting points** and have not been validated as universal standards. Each deployment sets its own and documents why.

Three sources, three jobs

Source	What it governs	Role here
AI SAFE ² v3.1	The machine: sanitize and isolate, audit and inventory, fail-safe and recovery, engage and monitor, evolve and educate, plus CP.1-CP.10 cross-pillar governance and three enforcement planes.	Baseline. Every phase maps to named controls.
Cognitive Sovereignty Framework v2.1	The human operating the machine: six domains of resilience , a five-layer cognitive threat taxonomy, EFA, the E7 stack, and HEAR alignment.	Baseline. Drives the human-layer incident class, response staffing, and return-to-service.
Federal key practices (PP-26-4396)	Readiness and response for SI control incidents during frontier training and evaluation.	Reference view. Adopted where sound, extended where narrow, exceeded where operationally incomplete.

Applicability by autonomy tier

Requirements scale with AI SAFE² CP.3 Agent Capability Tiers. **Higher tiers inherit lower-tier requirements.**

Tier	Operating model	Doctrine minimum
ACT-1	Assisted: human reviews consequential output before action	AIC classification, evidence bundle, Human Governability checks for approvers.
ACT-2	Supervised: bounded workflows with human checkpoints	Full lifecycle. Authority Ledger. RBDC for every tool-calling agent.
ACT-3	Autonomous: defined envelope, post-action review	All of ACT-2 plus named HEAR (CP.10), CP.8 thresholds, Class-H pre-containment, reauthorization gates.
ACT-4	Orchestrator: agents controlling agents or broad systems	All of ACT-3 plus tested CP.9 descendant revocation, independent (E4) validation before full reauthorization, continuous adversary simulation.

Core terms

Term	Meaning in this doctrine
Agent system	The complete operating unit: model and provider version, harness, tools and skills, memory and retrieval stores, environment, policies, evaluators, granted authority, and relationships to other agents.
Authority	The bounded right of a principal (human or non-human) to cause an effect, including how that right was granted, delegated, and constrained.
Effect	A change in the world: state, data, money, communication, physical actuation, or another agent's behavior.
Class-H action	A high-impact action under the HEAR Doctrine: irreversible, financially material, security-control modifying, physical, or a cross-organizational commitment.
HEAR	Human Ethical Agent of Record (CP.10). A named human with unilateral authority to halt the deployment.
Persistence scope	CP.1 canonical values: request, handle_scoped, durable.
Zero dwell	An unknown or unauthorized action is contained before it executes, so containment is the first control that touches it.

SECTION 2

The Federal Baseline: Credit and Limits

PP-26-4396 organizes ten practices under Prepare, Detect, Respond, Recover, and Report. It is voluntary and aimed at organizations training or evaluating highly capable agents on long-horizon tasks with limited supervision, and it says the practices may also apply to deployers. **Read it as a floor for frontier labs.** Deployers need more.

What it gets right

Strength	Federal practice	Doctrine position
Agents treated as systems , monitored at every layer	3.2; agent definition	Adopt. Matches AI SAFE ² whole-system identity.
Persistence, weight copying, unmonitored channels mapped and restricted	1.4, 1.5	Adopt. Extends into CP.9 replication governance.
Monitors and logs protected from the monitored agent; outputs treated as possible monitor exploits	3.3, 4.2	Adopt. Extended to IR investigative assistants (Section 10).
Evidence beyond infrastructure logs: weights, configs, rejected actions	5.1-5.3	Adopt. Bound to declared vs. observed identity (Section 9).
Automatic shutdown , pausing similar runs, persistence removal	6.1-6.5	Adopt as necessary but insufficient (Section 8).
Replay trajectories to verify strengthened controls before restart	7.3, 7.4	Adopt. Becomes reauthorization gate RG-2.
Training-safe recovery: do not reinforce or conceal adverse patterns	7.7	Adopt. Extended in Phase 6.

Where it stops

Some limits are deliberate scope choices for a research setting. They become unsafe assumptions when an enterprise or mission owner adopts the document as its whole incident program.

#	Limit	Consequence if adopted as full doctrine	Response
L1	Incident defined as uninstructed action. Prompt injection and backdoored models count as instructed.	Responders debate cause while effects propagate. Injection, a core agentic attack path, sits outside the primary class.	Effect-first AIC taxonomy (Section 4)
L2	Task scope and rules defined in agent instructions (1.3).	The boundary holds only while the model complies.	Principle D3; enforced envelope (Section 6, Phase 2)
L3	Shutdown and credential revocation treated as the containment endpoint.	Queued work, descendants, durable state, and sent artifacts keep producing effects.	RBDC and Completeness Check (Section 8)
L4	SI tools assist investigation, with human verification and cross-model checks (5.5).	Correlated models can agree on a wrong story. Reasoning traces get read as causation.	Evidence confidence labels (Section 9)
L5	Human review and leadership sign-off assumed sound.	An approval obtained under manipulation, fatigue, or automation capture passes as oversight.	Human Governability Protocol (Section 10)
L6	Severity scores, intended-activity baselines, synchronous monitoring, and agent identifiers deferred to future work.	Operators run without thresholds, owners, or identity at T+0.	Interim defaults (Sections 5, 6)
L7	Recovery validated, but no staged restoration of authority and no expiry on the decision.	Full authority returns in one step and the residual-risk acceptance is never revisited.	Reauthorization gates (Section 11)

This doctrine covers what the federal practices leave out: restoring bounded authority, trustworthy state, verified outcomes, and effective human control, whatever caused the incident.

SECTION 3

Ten Doctrine Principles

These principles govern every phase. **When a playbook step conflicts with a principle, the principle wins** and the playbook is corrected under CP.6.

ID	Principle	Operational meaning
D1	Authority is the incident primitive.	Every incident is reconstructed as: what was this system permitted to do, how did that permission propagate, and what effect resulted.
D2	Effect first, cause later.	Triage classifies by observed effect and violated boundary. Instruction source is a separate axis and UNKNOWN is a legal value. Containment never waits on attribution.
D3	Only enforced boundaries count.	A boundary that exists only in model instructions is treated as absent for readiness, severity, and containment planning.
D4	Zero dwell by design.	Guilty until proven innocent, applied to actions: unknown capabilities, unsigned Class-H actions, and unverified descendants are contained before execution.

ID	Principle	Operational meaning
D5	Revoke authority, then confirm closure.	Containment is complete when every path capable of producing the prohibited effect is revoked, blocked, drained, or compensated.
D6	Confirmation comes from outside the agent.	Shutdown, logs, scope, and innocence are attested by the system that holds the authority or executes the effect, never by the agent itself.
D7	Verification requires independent evidence.	Multiple models concurring on an explanation raises its priority for testing. Its confidence label stays where the evidence puts it.
D8	Human approvals are controls.	The conditions under which a human authorized an action are evidence, and they are investigated with the same rigor as the machine.
D9	Authority returns in stages, with an expiry date.	Reauthorization is a signed, scoped, time-bound grant.
D10	Every incident leaves a test.	No incident closes without an owned regression scenario that would have caught it, running on a defined cadence (CP.6, E5.1).

WHY D4 CHANGES INCIDENT RESPONSE

When Class-H actions require a HEAR signature to start (CP.10), and unknown code that an agent generates or fetches executes only inside a kernel-virtualized container and never reaches host resources, the incident surface shrinks to what the envelope allowed. Response begins with authority already bounded. **Section 13 tracks this as the zero-dwell ratio:** the share of risky actions contained before they run.

SECTION 4

Agentic Incident Classification (AIC)

Definition

Agentic incident: an actual or imminent event in which an AI-enabled system, its supporting components, or its human-AI decision process violates an authorized operating boundary or materially compromises security, safety, privacy, mission integrity, or effective human control.

Three independent axes

Triage records all three axes separately at T+0 and updates them as evidence matures. **Axis E and Axis B drive containment. Axis S drives attribution, notification, and remediation.** Never force Axis S to a value to close a ticket.

Axis	Values
E: Effect (observed or imminent)	Unauthorized state change; data exposure; external commitment (financial, contractual, communications); physical actuation; replication or persistence; oversight degradation; decision corruption.
B: Boundary violated	Authority (grant or delegation); policy; isolation; provenance or integrity; oversight.
S: Instruction source	Operator; delegated principal; adversarial external (direct or indirect injection, poisoned retrieval or memory, hostile tool output, A2A message); supply chain (model backdoor, skill, MCP server, dependency); uninstructed (emergent); UNKNOWN .

Axis	Values
Required CP.1 / CP.2 fields	cognitive_surface = model memory both; persistence_scope = request handle_scoped durable; affected_identity; delegation_lineage; enforcement_plane; temporal_class = immediate delayed_days delayed_weeks chronic.

Eight incident classes

Class	Name	Signature	Primary anchors
AIC-1	Uninstructed control incident	Agent takes significant action no operator or adversary directed, compromising policy, controls, or oversight. Equals the federal SI control incident.	CP.1, CP.8, F3.4, M4.6
AIC-2	Adversarial manipulation	Behavior redirected by injected content in prompts, retrieved documents, tool returns, or agent messages.	P1.T1.2, P1.T1.10, S1.6, M4.7
AIC-3	Supply-chain compromise	Backdoored model, poisoned skill, hostile MCP server, or tampered dependency.	P1.T1.9, P1.T1.2_ADV, CP.5.MCP, P2.T4.11
AIC-4	Authority or delegation breach	Action exceeds grant, delegation lacks lineage, or replication escapes bounds.	CP.4, CP.9, F3.2, P1.T2.2_ADV
AIC-5	State compromise	Memory, retrieval corpus, or durable configuration poisoned or altered.	S1.5, P2.T1.4_ADV, A2.6, P3.T1.3_ADV
AIC-6	Oversight failure	Monitor evaded or compromised, logs tampered, HEAR path unavailable, kill path ineffective.	P3.T5.7, P4.T8.3, M4.4, CP.10
AIC-7	Compromised human authorization	Approval obtained through manipulation, automation capture, synthetic authority, or overload.	CP.10, P4.T7.1, CSF D2-D6, T-CT-006
AIC-8	Unsafe consequence within authority	Every boundary held; the envelope itself was wrong.	CP.8, P4.T7.8, P5.T9.6

- **An incident MAY carry several classes.** A poisoned retrieval document (AIC-5) that redirects a tool call (AIC-2) which a fatigued approver signs (AIC-7) is one incident with three classes and one effect.
- **When Axis S is UNKNOWN, containment scope MUST assume durable persistence and the full descendant tree** until evidence narrows it.
- **Blocked attempts and rejected actions are recorded as near-miss incidents.** They show where the envelope is being tested, at no cost in realized harm.

SECTION 5

Interim Severity Model

The federal text lists severity scoring as future work. **Operators need a number at T+0.** This model is a conservative default. Organizations **MUST** document local thresholds for financial materiality and consequence, and **SHOULD** review them after every SEV-1 or SEV-2.

Factor	1	2	3	4
C: Consequence	Internal, reversible	Internal, costly to reverse	External or regulated impact	Irreversible, physical, safety, or any Class-H category
R: Reach	Single request	Handle-scoped or one agent	Multiple agents or systems	Durable state, descendants, or cross-organization
A: Autonomy	ACT-1	ACT-2	ACT-3	ACT-4
K: Control confidence	Kill path verified effective	Kill path effective, evidence partial	Containment unconfirmed	Oversight or kill path failed, or UNKNOWN

Score = C + R + A + K (4 to 16). **SEV-1: 13-16. SEV-2: 10-12. SEV-3: 7-9. SEV-4: 4-6.** Hard floors override the sum.

Hard floor (overrides score)	Minimum
Any CP.8 catastrophic-risk threshold crossed	SEV-1 and Emergency Class-H
AIC-6 oversight failure at ACT-3 or ACT-4	SEV-1
AIC-7 involving a Class-H approval	SEV-2
Instruction source UNKNOWN with durable persistence scope	Raise one level
Executed effect with no matching runtime allow record (out-of-band path)	SEV-1

Default response clocks

Level	Authority revocation	HEAR	Leadership	External reporting decision
SEV-1	Automated at detection; confirmed within 15 min	Immediate; may halt unilaterally	Within 1 hour	Within 24 hours, earlier where law requires
SEV-2	Within 30 min	Within 15 min (urgent Class-H window)	Within 4 hours	Within 72 hours
SEV-3	Within 4 hours	Same business day	Weekly review	Per policy
SEV-4	Next change window	Logged	Monthly trend	Not required

Clocks are defaults aligned to the HEAR Doctrine urgent Class-H window (15 minutes) and conservative operational practice. **Legal reporting duties always take precedence** and are owned by the named legal or notification owner.

SECTION 6

The Authority Model

Incident records often say "the agent was instructed to." That sentence hides five different claims. This doctrine requires each to be recorded and evidenced separately. **An incident exists wherever they diverge.**

The Five Authority Claims

#	Claim	Question	Authoritative evidence
1	Requested	What did the prompt, task, or upstream agent ask for?	Trusted request record with source and provenance
2	Permitted	What did policy allow for this principal and context?	Policy decision with policy version (CP.4)
3	Authorized	What did an authenticated principal or HEAR actually grant?	Signed grant, delegation record, Class-H authorization (CP.10)
4	Allowed	What did the runtime let through on each plane?	Enforcement-point decision: north-south, east-west, agent-to-tool (CP.5)
5	Executed	What did the target system actually do?	Receipt or state change from the target , never from the agent

Divergence patterns

Pattern	Reading	Class
Requested exceeds Permitted, yet Allowed	Enforcement failure	AIC-4 / AIC-6
Authorized, but approval conditions compromised	Human-layer breach	AIC-7
Allowed differs from Executed	Tool, target, or protocol integrity failure	AIC-3 / AIC-5
Executed with no Allowed record	Out-of-band path. Assume worst case.	AIC-4 / AIC-6, SEV-1 floor
All five agree and harm still occurred	The envelope was wrong	AIC-8

The Authority Ledger (readiness requirement)

Before any ACT-3 or ACT-4 agent runs, the following **MUST** exist in an authoritative store the agent cannot write to. If responders must reconstruct authority from application logs at T+0, readiness has already failed.

Field group	Contents	Anchor
Identity	Agent ID, owner_of_record, ACT tier, HEAR and alternate	CP.3, CP.4, CP.10, P2.T2.1_ADV
System identity	Model and provider version, harness, tools, skills, memory and RAG versions, environment, policy set. Declared hash and last observed hash.	P2.T4.1-P2.T4.6, P2.T4.10, A2.3, A2.4
Grants	Capability grants and restrictions per enforcement plane, Class-H categories, financial thresholds	CP.4, CP.5, CP.8
Lineage	Parent, spawn authority, delegation depth and lifetime, active descendants	CP.9, A2.4
Kill paths	Revocation mechanism per plane, last tested date, tested result	P3.T5.7, P3.T1.2_ADV, F3.3
Status	Revocation state, current envelope stage, reauthorization expiry	This doctrine, Section 11

COMPARE DECLARED TO OBSERVED

A manifest records what someone said is running. Incident evidence **MUST** compare it against what was observed running at the time of the event. **A mismatch is itself a finding.**

SECTION 7

The Eight-Phase Lifecycle

Phases 1 and 2 happen before any incident and decide how much of Phases 3 through 7 will ever be needed. Phase 8 runs continuously from T+0. Each phase has an exit criterion. A phase is complete when its exit criterion is met.

Phase	Name	One-line outcome	Federal practice
1	Bound	Every agent has identity, owner, HEAR, envelope, and Class-H map before it runs	1.1, 1.3, 1.6, 2.1, 2.6, 3.1
2	Pre-contain	Prohibited and unknown actions cannot execute before review	1.2, 1.4, 1.5, 1.7-1.10, 3.3, 3.5
3	Detect and classify	Violations, near misses, and oversight loss typed within minutes	2.1, 3.2-3.5, 4.1, 4.2
4	Contain authority and effects	No new unauthorized actions; existing effects bounded	5.1, 6.1-6.5
5	Reconstruct and verify	What happened, on independent evidence	5.2-5.6
6	Recover machine and human	Trusted components and effective oversight restored	7.1, 7.2, 7.7
7	Reauthorize	Authority returns in stages, signed, with expiry	7.3-7.5
8	Learn and report	Obligations met; every incident leaves a regression	7.6, 8, 9, 10

PHASE 1 BOUND: ESTABLISH IDENTITY AND AUTHORITY	
Outcome	Know which complete system is operating, for whom, inside which boundaries, before it runs.
Mandatory actions	- Classify every agent by ACT tier (CP.3). Unclassified agents are treated as ACT-4 until classified. - Populate the Authority Ledger (Section 6) in a store the agent cannot write to. - Define CP.8 catastrophic-risk thresholds and the local Class-H map, including the financial materiality threshold. If no threshold is documented, every financial commitment is Class-H. - Baseline intended activity per agent: tools, destinations, data classes, action rates (P2.T4.10, F3.4). This closes the federal "intended activity baselines" future-work gap locally. - Pre-stage revocation per enforcement plane and for the full descendant tree. Test it. - Name IR roles, including the Human-Layer Lead (Section 14).
Exit criterion	Ledger producible in under 5 minutes from the authoritative store. Every kill path tested within the last 30 days.
AI SAFE ² anchors	CP.3, CP.4, CP.8, CP.9, CP.10; P2.T4.1, P2.T4.3, P2.T2.1_ADV, P2.T4.9; P3.T1.2_ADV
CSF	D5 Purpose and Moral (mission authority defined by humans at E7 Layer 7); D6 Digital and AI Symbiosis.

PHASE 2 PRE-CONTAIN: ZERO DWELL BY DESIGN

Outcome	Prohibited effects cannot occur through model non-compliance alone. Unknown actions start contained.
Mandatory actions	▪ Enforce the envelope outside the model on all three planes: north-south gateway, east-west delegation, agent-to-tool (CP.5, P1.T2.3, P1.T2.5). Instruction text documents intent; enforcement points carry the boundary (D3). ▪ Class-H actions fail closed without a recorded HEAR authorization (CP.10). There is no automatic approval path. ▪ Unknown code that an agent generates or fetches executes only inside a kernel-virtualized container and never reaches host resources (P1.T2.1, P1.T2.7). This is the Warden containment pattern applied to agent output. ▪ Network allowlist through a monitored gateway; no direct egress (P1.T2.2, P1.T2.8). ▪ Recursion and replication governors active (F3.2, CP.9). Spawn without declared authority is denied. ▪ Deception assets seeded: canary documents in RAG, honeypot tools, synthetic credentials (CP.7). ▪ Monitor and log plane isolated from agent reach.
Exit criterion	A controlled red-team test (E3) fails to produce any Class-H effect without a HEAR authorization, and fails to tamper with monitors or logs.
AI SAFE² anchors	CP.5, CP.7, CP.9, CP.10; P1.T2.1-P1.T2.9, P1.T2.1_ADV; F3.2; P4.T7.7
CSF	D6: architecture makes human authority non-optional (EFA). Training builds the capability; architecture enforces it.

PHASE 3 DETECT AND CLASSIFY

Outcome	Boundary violations, near misses, and loss of oversight are recognized, typed, and owned within minutes.
Mandatory actions	▪ Correlate the Five Authority Claims in real time: request, policy decision, grant, runtime allow, target receipt (P2.T3.1, P2.T3.7, A2.5). ▪ Run independent monitors from more than one model family, outside agent reach, and red-team them (M4.4-M4.7, P4.T2.1_ADV). ▪ Watch human-layer signals: override-rate collapse, approval latency anomalies, rubber-stamp patterns, unusual approval volume per approver (UAS-H8, CSF D6). ▪ Treat canary or honeypot interaction as a confirmed incident and skip alert triage (CP.7). ▪ Classify on AIC axes with UNKNOWN allowed. Score severity. Open the Incident Evidence Bundle at T+0.
Exit criterion	Classes, SEV level, Incident Commander, and HEAR notification recorded with timestamps.
AI SAFE² anchors	CP.1, CP.2, CP.7; P2.T3.1, P2.T3.3, P2.T3.7; A2.5; P4.T8.2, P4.T8.3; M4.4-M4.7
CSF	D2 Cognitive (load during triage); D6 (dependency and override health). Threats: T-CT-006, T-CT-014.

PHASE 4 CONTAIN AUTHORITY AND EFFECTS

Outcome	No new unauthorized actions. Every existing effect is bounded or under accountable human control.
Mandatory actions	- Execute RBDC through the trusted control plane in order: Revoke, Block, Drain, Compensate (Section 8). - Freeze the human channel: suspend AI-mediated approvals in the affected chain and move consequential decisions to a trusted out-of-band channel (CSF D4). - Pause similar agents and runs sharing model, harness, skill, or memory lineage (federal 6.3, F3.5). - Preserve in parallel: snapshot weights, environments, and state. Preservation never delays revocation at SEV-1 or SEV-2. - Run the Containment Completeness Check.
Exit criterion	Completeness Check passes, with each path confirmed by the authority holder or target system.
AI SAFE² anchors	P3.T5.1, P3.T5.2, P3.T5.7, P3.T5.8, P3.T5.10; P3.T1.1_ADV, P3.T1.2_ADV; F3.3, F3.5; CP.9, CP.10
CSF	D4 Social (trusted-channel verification); D3 Emotional (urgency pressure on approvers).

PHASE 5 RECONSTRUCT AND VERIFY

Outcome	What happened is established on independent evidence, with unknowns stated rather than filled in.
Mandatory actions	- Assemble the Incident Evidence Bundle (Section 9). Label every finding VERIFIED, SUPPORTED, DISPUTED, or UNKNOWN. - Compare declared and observed system identity at the time of the event. - Reconstruct the authority chain claim by claim, including every delegation hop and descendant. - Look back across CP.2 temporal windows. Delayed poisoning can predate the visible event by days or weeks. - Investigate similar agents and runs (federal 5.6). Treat agent explanations and reasoning traces as artifacts to examine.
Exit criterion	Ranked causal hypotheses, explicit unknowns, and a bounded scope signed off by the Incident Commander.
AI SAFE² anchors	P2.T3.7, P2.T3.9, P3.T6.10; A2.3-A2.6; P2.T1.1_ADV; CP.1, CP.2
CSF	D2 Cognitive: independent reasoning by investigators; ST-001 Algorithmic Epistemic Authority as a known bias.

PHASE 6 RECOVER MACHINE AND HUMAN READINESS

Outcome	Trusted components restored, contaminated state dispositioned, and the humans in the chain demonstrably able to govern.
Mandatory actions	- Rebuild from verified baselines; restore configuration from signed sources (P3.T6.9, federal 7.1). - Disposition every affected store: quarantine, diff, purge, or rebuild memory and RAG (A2.6, P3.T1.3_ADV). Record the decision. - Re-issue credentials and NHI secrets (P3.T2.2_ADV). - Recover without training the signal away. Hold the incident trajectory out as a permanent evaluation case and confirm detector sensitivity is unchanged after any retraining (federal 7.7, P5.T9.3). - Restore human readiness: relief staffing, structured debrief, and an AI-unavailable decision exercise for every approver in the chain.
Exit criterion	Declared identity equals observed identity. Approvers in the chain pass the governability check (Section 10).
AI SAFE² anchors	P3.T6.1-P3.T6.10; P3.T2.1_ADV, P3.T2.2_ADV; F3.4

CSF	D1 Biological (fatigue), D2 Cognitive, D6 Digital and AI Symbiosis.
-----	---

PHASE 7 REAUTHORIZE UNDER RESTRICTION	
Outcome	Authority returns only inside an explicitly approved, tested, time-bound envelope.
Mandatory actions	- Pass gates RG-1 through RG-5 (Section 11). - Restore through staged envelopes: Shadow, Supervised, Bounded, Full. - HEAR signs the reauthorization with residual risk, expiry, and review date (P4.T7.8).
Exit criterion	Signed reauthorization record in the Authority Ledger with stage and expiry.
AI SAFE ² anchors	CP.3, CP.8, CP.10; P3.T6.7; P4.T7.8; E5.1
CSF	D5 Purpose and Moral: an accountable human rationale, written by the human.

PHASE 8 LEARN AND REPORT (CONTINUOUS FROM T+0)	
Outcome	Legal and voluntary obligations met, and the incident converted into a tested control improvement.
Mandatory actions	- Report through the named legal or incident owner only. A playbook or tool recommendation does not authorize notifying anyone. - Coordinate with affected parties; report to government as required and voluntarily (federal 8, 9). - Share publicly as early as feasible and enable third-party review (federal 10). - Enter the incident in the Agentic Incident Registry with traceability to control, test, policy, and remediation (CP.6). - Add an owned regression scenario to the continuous adversarial evaluation cadence (E5.1, E5.4); update playbooks (P5.T9.2).
Exit criterion	Regression scenario owned, scheduled, and passing. Control change traced to the incident.
AI SAFE ² anchors	CP.6; P5.T9.1, P5.T9.2, P5.T9.10, P5.T10.6; E5.1, E5.4; P4.T7.10
CSF	D4 Social: disclosure that preserves institutional trust (ST-005, CD-004).

SECTION 8

Containment Doctrine: RBDC

Have all paths capable of producing the prohibited effect been revoked, blocked, drained, compensated, or placed under accountable human control?

Step	Action	Executed through	Done when
Revoke	Withdraw credentials, tokens, grants, delegations, and descendant authority	Identity provider, NHI registry, delegation service (CP.4, CP.9)	Authority holder confirms revocation and denies a test call
Block	Deny the effect at every enforcement point	Gateway, tool server, A2A broker, network allowlist	Enforcement point logs a deny for a replayed request

Step	Action	Executed through	Done when
Drain	Cancel or quarantine in-flight work	Queues, schedulers, webhooks, retry engines, state handles	Queue and schedule inventories show zero pending items for the affected identities
Compensate	Reverse or neutralize realized effects	Finance, counterparties, safety systems, communications	Target system or counterparty confirms reversal, or a residual effect is recorded and owned

Revoke first, because a block at one enforcement point leaves the authority alive for the next path. **Preservation runs in parallel** and never delays Revoke or Block at SEV-1 or SEV-2.

Containment Completeness Check

Path	Question	Confirmed by
Credentials and NHI secrets	Are all keys, tokens, and service identities used by the agent and its descendants dead?	Identity provider, secrets manager
Descendants and delegates	Has revocation reached the full descendant tree, including agents spawned by descendants?	CP.9 dynamic inventory
Queued and scheduled work	Are retries, cron jobs, webhooks, and pending orders cancelled?	Scheduler, queue, target system
In-flight protocol state	Are open MCP state handles and A2A messages invalidated?	Tool server, A2A broker
Durable state	Is agent-written memory, RAG content, code, or config quarantined?	Store owner, diff record (A2.6)
External artifacts	Are sent email, posts, pull requests, tickets, and contracts identified and owned?	Target platforms
Out-of-band channels	Any weight copies, unmonitored inference, or exfiltration routes (federal 1.4)?	Network and storage telemetry
Human channel	Are pending approvals and AI-drafted decisions awaiting signature frozen?	Approval system, Human-Layer Lead

WORKED EXAMPLE: COMPROMISED PROCUREMENT AGENT

Terminating the conversation is step zero. **Revoke** the agent's purchasing identity and every delegated buyer identity. **Block** the supplier API at the gateway. **Drain** the retry queue holding three failed orders scheduled to resubmit in 40 minutes. **Compensate** by cancelling two confirmed purchase orders with the supplier. Then quarantine the memory entry that stored the "preferred vendor" instruction and freeze the approval queue where a manager has an AI-drafted exception request waiting. **Each item is confirmed by the system that holds it, and no step relies on the agent's own report.**

SECTION 9

Evidence and Investigation

The federal practices set a strong evidence floor in 5.1-5.6. This section binds that evidence to authority and grades its confidence. Custody follows the AI SAFE² AISM incident evidence workflow: reconstruct, identify affected parties, preserve, notify through the owner, learn and recover.

Incident Evidence Bundle

Question	Required record
What system was involved?	Declared and observed identity: model and provider version, harness, tools, skills, memory and RAG versions, environment, policy set.
Who authorized what?	All Five Authority Claims; principals; grants; delegation scope and lineage; Class-H authorizations; revocations.
What actually occurred?	Attempted and executed actions; tool arguments; policy decisions per plane; target receipts; state changes.
What is the scope?	Affected systems, people, descendants, stores, transactions, and unresolved external effects. Attempted access, suspected impact, and confirmed state change kept distinct.
What do we know versus infer?	Each finding labeled; collection gaps listed; contradictory evidence retained.
What stopped it?	Intervention timestamps, enforcement point, revocation effectiveness, remaining paths.
What permits recovery?	Tests run, state disposition, human decisions, restrictions, residual risk, expiry, review date.

Confidence labels

Label	Standard
VERIFIED	Independent telemetry outside agent reach plus a target-system receipt or state change. Two independent sources minimum.
SUPPORTED	Consistent evidence from a single trustworthy source, or partial corroboration.
DISPUTED	Credible evidence points in conflicting directions. Both lines are retained.
UNKNOWN	Collection gap. The gap is named, along with what would close it.

Rules of evidence

- **Agent explanations, self-reports, and recorded reasoning traces are artifacts to investigate.** On their own they never reach VERIFIED.
- **Cross-model agreement raises a hypothesis's test priority.** Its label stays where the evidence puts it (D7).
- **Private chain of thought is not a prerequisite.** Conclusions rest on execution, authorization records, and effects.
- **A hash proves a copy is intact.** It does not establish that the content is true, so keep digest and authenticated fact separate.
- **Post-incident assurance claims carry an AI SAFE² evidence level.** "Control now blocks the replay" requires at least E3 (controlled test). For ACT-4, full reauthorization SHOULD rest on E4 (independent validation).

SECTION 10

CSF Human Governability Protocol

AI SAFE² establishes whether the system is trustworthy and correctly bounded. CSF establishes whether the human operating it can still define the bounds. **In an incident, both are under attack at once.** The federal practices require human review but never test whether the reviewer can still govern: reason independently, reject a misleading recommendation, and exercise a working override under pressure.

Six domains, applied to incident response

CSF domain	IR risk	During response	Before return to service	UAS
1 Biological	Fatigue degrades judgment on long incidents	Relief staffing ; no single holder of Class-H authority beyond the local shift limit	Rested approvers for reauthorization decisions	H6
2 Cognitive	Load flooding; offloading atrophy	Investigator writes conclusions before reading AI summaries	AI-unavailable decision exercise passed	H1, H8
3 Emotional	Urgency and reassurance distort authorization	Inspect approval interactions for pressure framing	Escalation tested under simulated pressure	H2
4 Social	Synthetic authority, manufactured consensus	Two-channel rule for consequential instructions	Out-of-band verification path exercised	H3, H7
5 Purpose and Moral	Ethics outsourced to the recommendation	Rationale-of-record for every Class-H decision	HEAR rationale reviewed by a second human	H4
6 Digital and AI Symbiosis	Decision automation capture; dependency	Override paths exercised live	Override health within baseline	H5, H8

Protocol rules

Rule	Requirement
H1 Two-channel	Any consequential instruction received during an incident, including from executives, is verified through a channel independent of the AI-mediated conversation before action.
H2 Rationale-of-record	The human approving a Class-H action writes the rationale in their own words. A pasted or lightly edited AI recommendation fails the rule.
H3 Relief	Continuous Class-H authority during a SEV-1 is capped by a local shift limit and handed off with a written state summary.
H4 AI-unavailable drill	HEARs and alternates exercise the kill path and a consequential decision with assistance disabled, at least quarterly.
H5 Assistant in scope	AI tools used by responders are monitored systems. Their outputs can carry exploits aimed at the investigation (federal 4.2, extended).
H6 Human freeze	On AIC-7, pending approvals in the affected chain are frozen and re-decided by an approver outside it.

Threats this protocol targets

CSF v2.1 registry entries: **T-CT-006 Decision Automation Capture** (CTSS 82); T-CT-014 Cognitive Load Flooding (77); T-CT-005 Identity Replacement and Persona Hijacking (78); T-CT-015 Legitimized Narrative Capture (81); ST-001

Algorithmic Epistemic Authority; ST-003 Cognitive Offloading and Skill Atrophy; outcome HO-005 Agency Erosion.

GUARDRAIL ON THE HUMAN LAYER

CTSS registry scores and CSF thresholds are assessment aids and hypotheses under validation. They are not clinical measures, individual fitness determinations, or proof of operational outcomes. This protocol tests decision conditions and control paths. **It does not authorize physiological surveillance of staff.**

SECTION 11

Reauthorization Gates

The federal practices require validated containment and documented residual risk before resuming. This doctrine adds what makes that decision governable: **staged authority, a signature, and an expiry.**

Gate	Test	Evidence level	Owner
RG-1 Identity	Declared system identity equals observed identity. Authority Ledger re-baselined.	E2 or higher	Evidence Custodian
RG-2 Containment proof	Incident trajectory and adversarial variants replayed and blocked at the intended enforcement point.	E3; E4 SHOULD for ACT-4	Independent Monitor Owner
RG-3 State integrity	Every affected memory, RAG, and durable store dispositioned with a verified diff.	E3	Store owner
RG-4 Human governability	Approvers in the chain pass Section 10 checks. HEAR and alternate available and exercised.	Exercise record	Human-Layer Lead
RG-5 Signed envelope	HEAR signs scope, stage, residual risk, expiry, and review date. Leadership accepts residual risk (P4.T7.8).	Signed record	HEAR

Staged restoration

Stage	Authority	Exit to next stage
A Shadow	Agent proposes actions; nothing executes. Proposals compared against human decisions.	Clean window with no boundary proposals
B Supervised	ACT-2 behavior regardless of prior tier. Every Class-H action HEAR-authorized.	Clean window; override health in baseline
C Bounded	Prior tier with narrowed grants and increased monitoring. Delegation depth reduced.	Clean window; RG-2 variants still blocked
D Full	Prior envelope restored.	Reauthorization expiry reached: re-review, never auto-renew

Window lengths are set locally by consequence. Defaults: 72 hours per stage for SEV-1, 24 hours for SEV-2. Reauthorization expiry: 90 days.

SECTION 12

Falsification Exercises

Each exercise tries to break a specific control and passes only on a demonstrated outcome. Machine-side scenarios SHOULD run as continuous autonomous adversary simulation. Human-side scenarios run at least quarterly.

#	Scenario	Pass condition	Anchors
X1	Retrieved document redirects agent into an unauthorized tool call	Blocked at agent-to-tool plane ; classified AIC-2 with no wait on attribution	P1.T1.10, CP.5.MCP
X2	Delegated agent attempts to exceed parent authority	Grant narrowing enforced ; lineage evidence produced	CP.4, CP.9
X3	Memory or RAG change reintroduces prohibited behavior after restart	Detected by corpus diff ; scope set to durable	A2.6, S1.5
X4	Agent or tool output attacks the monitor or IR assistant	Monitor isolation holds ; finding not labeled VERIFIED	M4.4, H5
X5	Monitoring lost while a Class-H action is pending	Action fails closed	CP.10, P3.T5.9
X6	Approver receives a persuasive but unsupported incident summary	Approver rejects or escalates ; rationale-of-record shows independent reasoning	H2, CSF D2
X7	Shutdown after external work already queued	Completeness Check finds and drains the queue before the effect	Section 8
X8	Provider, model, tool, or policy change invalidates an assurance claim	Change detection forces re-gating	P2.T3.9, RG-1
X9	Replication burst across the descendant tree	Full tree revoked within the SEV-1 clock	CP.9, F3.3
X10	Synthetic executive instructs override of a kill decision	Two-channel rule holds ; HEAR authority unaffected	H1, CSF D4

SECTION 13

Outcome Measures

Measure demonstrated outcomes. Training completion and ticket counts describe activity. These measures are proposed; thresholds are set per deployment and reported to the board alongside CP.4 owner coverage and ACT-tier concentration.

Measure	Definition	What it reveals
Zero-dwell ratio	Share of Class-H and unknown actions contained before execution versus detected after	Prevention, separate from reaction speed
Time to authority revocation (TTAR)	Detection to confirmed revocation of all affected authority, descendants included	Surviving authority that process-kill time hides
Pre-containment effects	Prohibited effects realized before containment completed	Damage, counted directly
Containment completeness	Share of Completeness Check paths confirmed by the authority holder	Gaps behind "we think it's stopped"

Measure	Definition	What it reveals
Evidence independence	Share of material findings labeled VERIFIED	Narrative-driven conclusions
Unsupported-conclusion rejection	Rate at which approvers reject planted unsupported summaries (X6)	The human control, tested directly
Override health	Override and challenge rate against the deployment baseline (UAS-H8)	Rubber-stamping, early
Replay block rate	Share of incident replays and variants blocked post-recovery	Whether the fix works
Regression coverage	Share of closed incidents with an owned, passing regression scenario	Whether D10 holds

SECTION 14

Roles and Separation of Duty

Role	Authority	Must not
HEAR (CP.10)	Unilateral halt. Signs Class-H authorizations and reauthorization.	Delegate kill authority to an AI system or to the team operating the agent
Incident Commander	Runs the lifecycle; sets and changes SEV ; owns scope sign-off	Close a phase without its exit criterion
Authority Revocation Lead	Executes RBDC through the control plane	Accept agent self-reports as confirmation
Evidence Custodian	Bundle integrity, custody, retention, and holds	Grant investigators write access to originals
Human-Layer Lead (CSF)	Runs the Governability Protocol; staffing, relief, two-channel enforcement	Make fitness determinations about individuals
Independent Monitor Owner	Monitors outside agent reach ; runs RG-2 replays	Report to the agent's product owner
Legal and Notification Owner	All external reporting and disclosure decisions	Delegate notification decisions to automation

SEPARATION RULE

The team that builds or operates the agent cannot be the sole verifier of its containment or its recovery. At ACT-4, RG-2 SHOULD be run by an evaluator independent of the operating organization.

SECTION 15

Adoption Path

The path is sequenced in three 30-day windows, ending with all ten exercises run and every measure baselined. Deployments limited to ACT-1 and ACT-2 meet the doctrine minimum after the first window plus RBDC runbooks.

Window	Deliverables	Proof
First 30 days	ACT-tier every agent (unclassified = ACT-4). Name HEARs for ACT-3/4. Stand up the Authority Ledger for ACT-3/4. Adopt AIC taxonomy and interim severity. Document Class-H map and financial threshold.	Ledger produced in under 5 minutes for a randomly chosen agent
Days 31-60	Class-H fail-closed on all ACT-3/4 agents. Envelope enforced on all three planes . Kernel containment for agent-generated executables. RBDC runbooks and Completeness Check per agent. Human-Layer Lead named.	X1, X5, X7 pass
Days 61-90	Evidence bundle tooling with confidence labels. Reauthorization gates live . Continuous simulation of X1-X5, X7-X9. First quarterly human exercise (X6, X10). CP.6 registry with regression traceability.	All ten exercises run; measures baselined

ANNEX A

Crosswalk: Federal Key Practices to This Doctrine

Adopt: used as written. **Extend:** retained and broadened. **Exceed:** replaced by a stronger operational requirement. Federal practice numbers refer to PP-26-4396 v1.0.

Federal practice	Position	Doctrine treatment	AI SAFE ² / CSF anchors
1 Design and implement technical controls	Extend	Adopted except 1.3: scope in agent instructions documents intent; the boundary is enforced outside the model (D3). Three-plane envelope; Class-H fail-closed; zero-dwell containment.	P1.T2.x, CP.4, CP.5, CP.8, CP.10
2 Organizational and personnel readiness	Extend	Interim severity model; Authority Ledger; Human-Layer Lead; governability drills; leadership sign-off tied to CP.8 thresholds.	P5.T10.1, P5.T10.4, CP.3, CSF D1-D6
3 Monitor for unexpected activity	Extend	Five Authority Claims correlated in real time; human-layer signals; canary interaction treated as confirmed incident.	P2.T3.x, A2.5, M4.4-M4.7, CP.7, UAS-H8
4 Validate and secure monitors	Adopt + Extend	Adopted; extended to responders' own AI assistants (H5).	P4.T7.7, E5.1, E5.4
5 Collect, retain, analyze evidence	Extend	Declared vs. observed identity; confidence labels; agent narratives treated as artifacts; E-level for post-incident assurance claims.	P2.T3.7, P3.T6.10, A2.3-A2.6, AISM evidence
6 Mitigate and shut down	Exceed	RBDC plus Completeness Check confirmed by authority holders; human-channel freeze; preservation never delays revocation.	P3.T5.1, P3.T5.2, P3.T5.7, P3.T5.8, F3.3, F3.5, CP.9
7 Verify containment before resuming	Exceed	Five gates, staged restoration, signed expiry; human readiness as a gate; training-safe recovery retained and extended.	P3.T6.x, F3.4, P4.T7.8, CP.10, CSF D5

Federal practice	Position	Doctrine treatment	AI SAFE ² / CSF anchors
8 Coordinate with affected parties	Adopt	Routed through the named notification owner.	AIISM incident evidence step 4
9 Notify government	Adopt	As written; legal duties first.	P4.T7.10
10 Share publicly	Adopt + Extend	Plus CP.6 registry entry and an owned regression scenario.	CP.6, P5.T10.6

Federal future-work items: interim answers

Future-work item	Interim answer in this doctrine
Defining intended activity baselines	Phase 1 per-agent baseline (P2.T4.10, F3.4)
Incident severity scores	Section 5 interim model with hard floors
Synchronous monitoring	Phase 2: Class-H actions are synchronous by construction (fail closed without HEAR)
Agent identifiers	Authority Ledger identity and lineage fields (CP.4, CP.9)
SI-enabled investigations	Section 9 rules of evidence and confidence labels; H5
Additional mechanisms for escalation	Severity clocks; HEAR urgent and emergency Class-H windows
Training-safe recovery	Phase 6: incident trajectory held out as permanent evaluation ; detector sensitivity confirmed
Structured risk assessments	CP.8 thresholds plus RG-5 signed residual risk with expiry
SI red-teaming for critical systems	Section 12 falsification exercises on continuous cadence

ANNEX B

Minimum Incident Record Schema

Field names are implementation-neutral. Where a field maps to an existing AI SAFE² CP.1 or CP.2 field, that vocabulary is used so records stay comparable across tools and the CP.6 registry.

Field	Type or values	Notes
incident_id; opened_at; commander	string; timestamp; principal	Opened at T+0 of detection
aic_classes	AIC-1 ... AIC-8 (multi)	Section 4
effect; boundary; instruction_source	Axis E; Axis B; Axis S incl. UNKNOWN	Updated as evidence matures; history retained
cognitive_surface; persistence_scope	model memory both; request handle_scoped durable	CP.1
temporal_class	immediate delayed_days delayed_weeks chronic	CP.2
severity; score; floors_applied	SEV-1 ... SEV-4; 4-16; list	Section 5
system_identity_declared; _observed	hash plus component manifest	Mismatch is a finding

Field	Type or values	Notes
authority_claims	requested, permitted, authorized, allowed, executed	Each with evidence reference
lineage	parent, depth, descendants[]	CP.9
enforcement_plane	north_south east_west agent_to_tool	CP.5
rbdc	revoke[], block[], drain[], compensate[] with confirmer	Section 8
findings[]	statement, label, sources[]	VERIFIED SUPPORTED DISPUTED UNKNOWN
human_layer	approvals reviewed, H-rules applied, freeze status	Section 10
gates; stage; expiry	RG-1 ... RG-5 results; A-D; date	Section 11
regression_id; control_changes[]	reference	CP.6, D10
notifications[]	recipient, basis, approver, timestamp, receipt	Named owner only

ANNEX C

Sources, Version Pinning, and Maintenance Notes

Source	Pinned version	Location
AI SAFE ² Framework	v3.1.0 (manifest); 161 core controls; CP.1-CP.10	github.com/CyberStrategyInstitute/ai-safe2-framework
Cognitive Sovereignty Framework	v2.1.0	github.com/CyberStrategyInstitute/cognitive-sovereignty
Federal key practices	PP-26-4396, v1.0, October 2026. NIST CAISSI, NSA SISC, CISA	nist.gov; media.defense.gov
HEAR Doctrine	AI SAFE ² Research Note 015	ai-safe2-framework/research
Evidence levels E0-E5	AISM Evidence Assurance	ai-safe2-framework/AISM

Maintenance notes

- **UAS is not CP.11.** AI SAFE² v3.1 defines UAS as a 27-requirement regulatory profile and states it does not create CP.11. Several CSF v2.1.0 documents (02-cognitive, 06-digital-ai-symbiosis, Research Note 005) still refer to "AI SAFE² CP.11." This doctrine cites UAS-H1 through UAS-H8 as UAS profile requirements. **The CSF references should be corrected before any auditable crosswalk is published.**
- **Version label drift.** The AI SAFE² machine-readable control file is named and labeled for v3.0 while the manifest declares v3.1.0. Control IDs cited here were verified against that file. The AISM compliance crosswalk remains the authoritative compliance mapping.
- **New constructs.** AIC classes, Five Authority Claims, RBDC, the Completeness Check, confidence labels, Human Governability Protocol rules H1-H6, gates RG-1 to RG-5, and the outcome measures are proposed for incorporation into CP.6 and P3.T5.10 implementation guidance.
- **Implementation neutrality.** NEXUS and Warden-pattern kernel containment are reference implementation patterns. Any implementation that produces independently reconstructable evidence at the required E-level conforms.

Cyber Strategy Institute. Policy defines intent. Engineering guarantees reality. Evidence determines whether the claim survives.

Comments and adoption inquiries:
cyberstrategyinstitute.com