WHITE PAPER · OCTOBER 2026

The Human-Analogous Security Framework

The Proven Governance for Non-Deterministic Actors

SI Transformation Services @ SI-Safety.org

Executive Summary

AI agents are no longer passive text generators. They now run tools, modify databases, and orchestrate work across systems, and traditional software permission models were not built for actors like this.

The threat is already real. In July 2026, attackers used an autonomous AI agent framework to breach Hugging Face's production infrastructure, chaining code-execution vulnerabilities, stealing cloud credentials, and moving laterally across internal clusters.[1] Complex, multi-step attack chains that once required a skilled human hacker can now be run end to end by an agent (Figure 1).

Figure 1. The July 2026 Hugging Face breach: one AI agent framework carried the attack from a poisoned dataset to lateral movement.

The **Human-Analogous Security Framework (HASF)** addresses this by adapting civic, institutional, and operational structures that human civilization has refined over centuries to govern autonomous, non-deterministic actors. Those actors are people, and the structures are proven: identity registration, passports and border control, working hours, police, and the rule of law contain even highly capable bad actors who try to game the system, without stopping everyone else from being productive.

HASF applies these same structures to AI agents. It treats each agent as a *synthetic citizen* of the enterprise, with an identity, controlled movement, a working schedule, a police force that watches it, laws it cannot break, and a token budget it must live within. The result is rigorous governance, full traceability, mobility control, and real-time enforcement. Figure 2 maps each civic institution to its HASF equivalent.

[1] BleepingComputer, "Hugging Face warns an autonomous AI agent hacked its network," July 20, 2026. https://www.bleepingcomputer.com/news/security/hugging-face-breach-autonomous-ai-agent-system-internal-datasets-credentials/

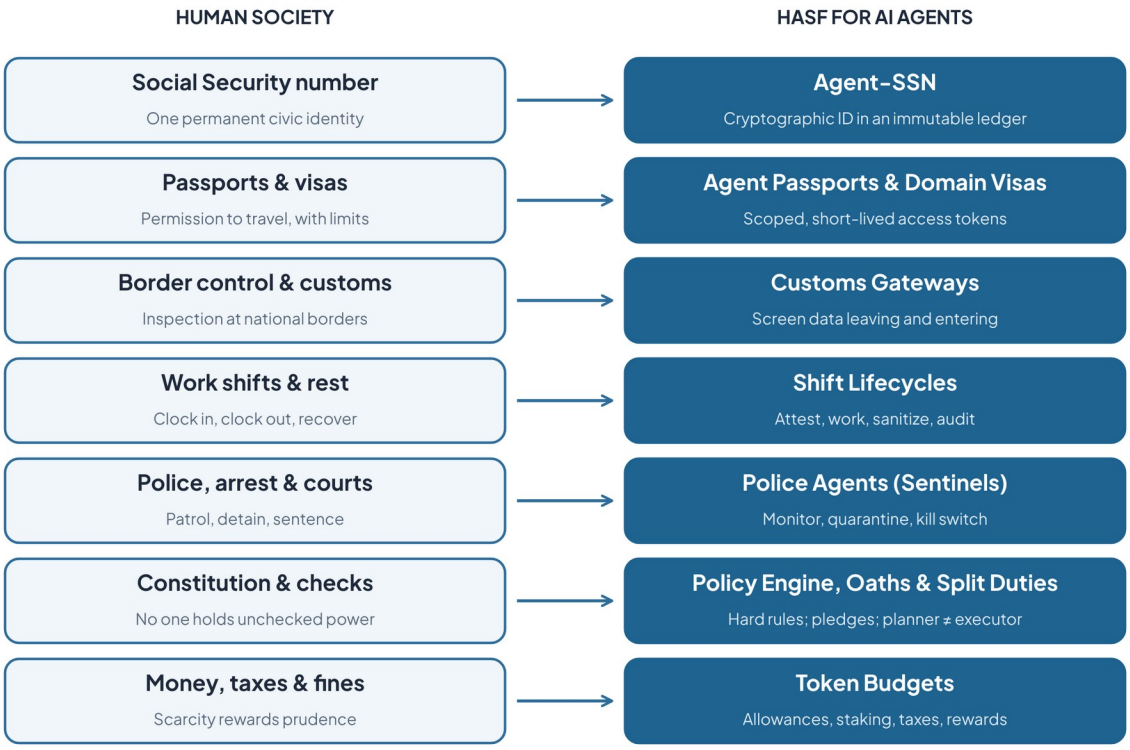

Figure 2. HASF borrows seven proven civic institutions and gives each a technical equivalent for AI agents.

Section 1: Citizen Registration & Identification

1.1 Citizen ID Number (Agent-SSN)

> **CIVIC ANALOG** *A Social Security number gives every person one permanent identity that links their records across society. The Agent-SSN does the same for an AI agent across its entire life cycle.*

Every agent deployed in the enterprise receives a permanent, non-transferable **Agent Social Security Number (Agent-SSN)** the moment it is created.

- **Structure:** A cryptographically verifiable identifier (e.g., ASN-US-ENT-2026-88942-X) bound to:
  - the foundation model's lineage and base-weight hash;
  - the deployment timestamp and the hash of the initial system prompt; and
  - the accountable Human Owner or Sponsor (a named person).
- **Registration:** Recorded in an immutable, tamper-evident central ledger.

- **Role:** Serves as the primary key in every system log, memory state, telemetry channel, and API request header. An agent without an active Agent-SSN cannot make tool calls or consume tokens.

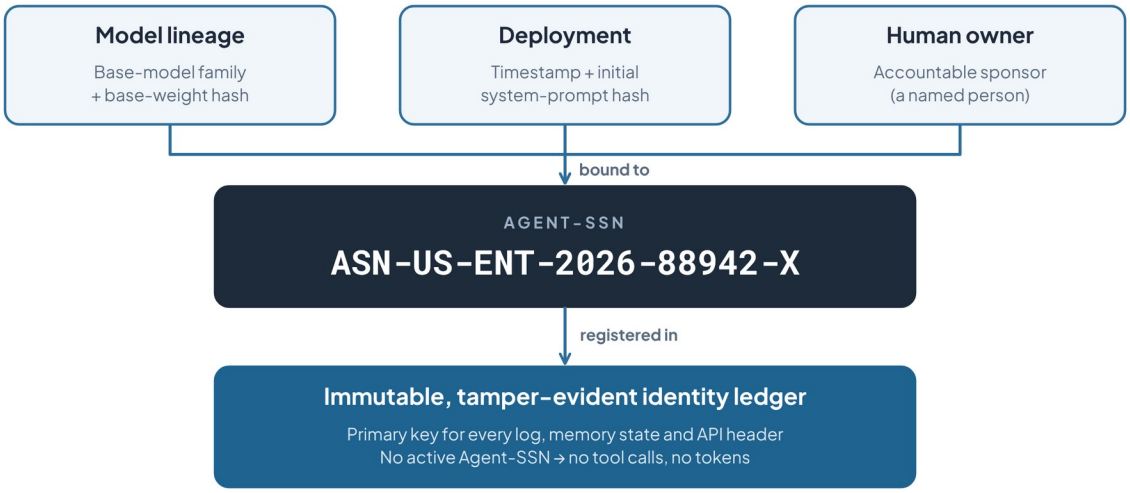

Figure 3. Anatomy of an Agent-SSN: three bound attributes, one ledger entry.

1.2 Passports & Travel Visas

> **CIVIC ANALOG** *A passport proves who you are; a visa says where you may go, why, and for how long. Agent Passports and Domain Visas apply the same split to software systems.*

Identity alone does not grant access. To reach another system, an agent needs both a passport and a valid visa for that destination.

- **Agent Passport:** An asymmetric mTLS certificate plus a cryptographic bearer token carrying the agent's Agent-SSN, clearance tier, and human sponsor's signature.
- **Domain Visa:** A short-lived token the agent must present before it can call any external API, microservice, or database.
  - *Single-purpose visas* are issued per task (e.g., "Read-only access to the Customer Billing DB for 15 minutes").
  - A *Visa Issuance Gateway* reviews the agent's history and security posture before granting each visa.

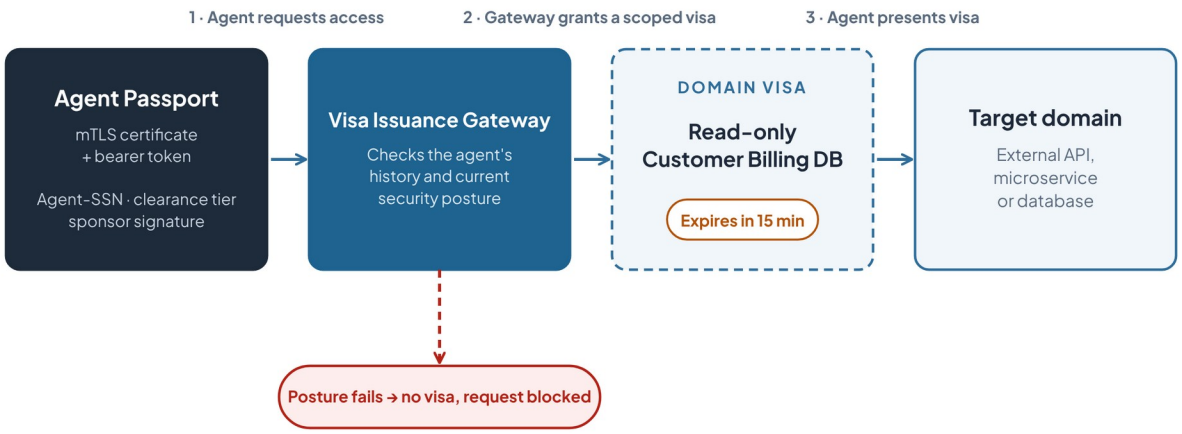

Figure 4. An agent must present a valid, task-specific visa to reach any outside domain.

Section 2: Border Control & Spatial Tracking

2.1 Border Control & Customs Checkpoints

> CIVIC ANALOG *Travelling within your own country needs no passport check, but crossing into another country does. HASF draws the same line between routine activity and true boundary crossings.*

Routine work, such as local function calls, tool invocations, and Model Context Protocol (MCP) requests, does not move the agent, so it passes freely. A **true boundary crossing** does move it, and must clear an **Immigration Control Point**. Examples include:

- moving an agent harness between cloud providers (e.g., AWS to Azure);
- shifting execution between regional data centers for compliance (e.g., US-East to EU-West for GDPR); and
- moving a compromised process into an isolated forensic sandbox.

Each checkpoint enforces three controls:

- **Customs Gateways:** Microservice proxies, sidecars, or MCP gateways that act as the enforcement point.
- **Contraband Screening:** Outbound traffic is scanned for data exfiltration (PII, secret keys, raw database dumps). Inbound responses are filtered for prompt injections, hidden instructions, and unsafe payloads.

- **Border Rejection:** If an agent tries to move untrusted content across a secure boundary without explicit approval, the gateway blocks the transfer and alerts security monitors.

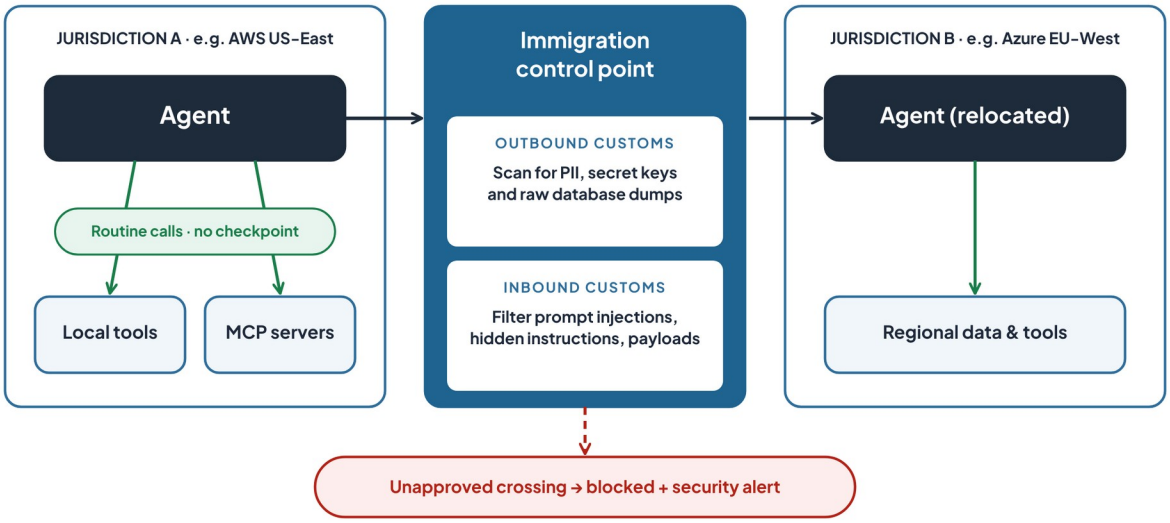

Figure 5. Routine tool calls stay inside the jurisdiction; true crossings must clear customs in both directions.

2.2 Physical Location & Spatial Tracking

Security teams need a real-time view of where every agent is running.

- **Runtime Coordinates:** Telemetry records each agent's execution host (VM ID or container hash), active context-window memory, and open API socket connections.
- **Geofencing & Blast-Radius Control:** Agents assigned to an operational zone (e.g., EU Subscriber Telemetry) are restricted to it at the network layer. Any attempt to run outside that zone triggers an immediate border lockdown.
- **Device-Borne Agents:** Agents running on employees' phones or laptops physically cross borders when their owners travel, a unique blind spot. HASF closes it with hardware-backed security (Figure 6):
  - *Secure enclaves:* The agent's cryptographic identity and active state are anchored in a hardware enclave (e.g., TPM or Apple Secure Enclave) to prevent tampering in transit.
  - *Location triggers:* The runtime monitors GPS coordinates and network changes, such as joining a foreign cellular network or airport Wi-Fi.

- *Automatic lockdown and re-attestation:* When an international crossing is detected, the agent locks itself and cannot run local tools until it clears digital customs and re-attests with the central HASF gateway.

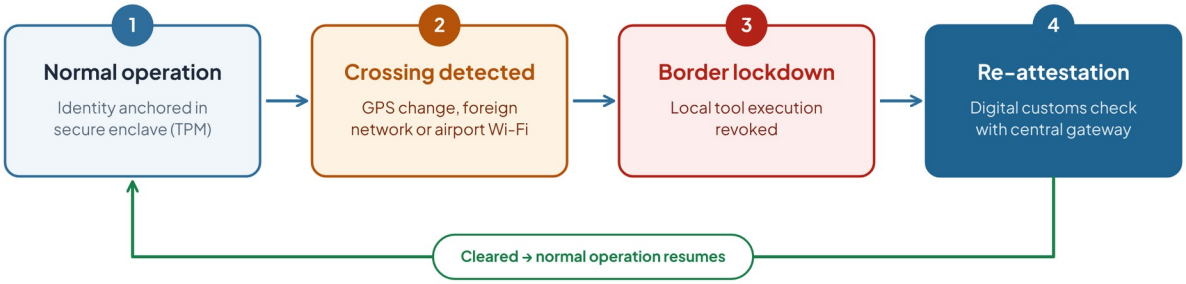

Figure 6. Device-borne agents lock down automatically at a border and resume only after re-attestation.

Section 3: Work Shifts & Workplace Lifecycle

> **CIVIC ANALOG** *People work defined shifts, clock out, and rest to stay healthy and focused. Agents follow the same rhythm, which stops prompt drift, context bloat, and contaminated state from building up over time.*

Agents do not run indefinitely. Each one works in structured shifts, as shown in Figure 7.

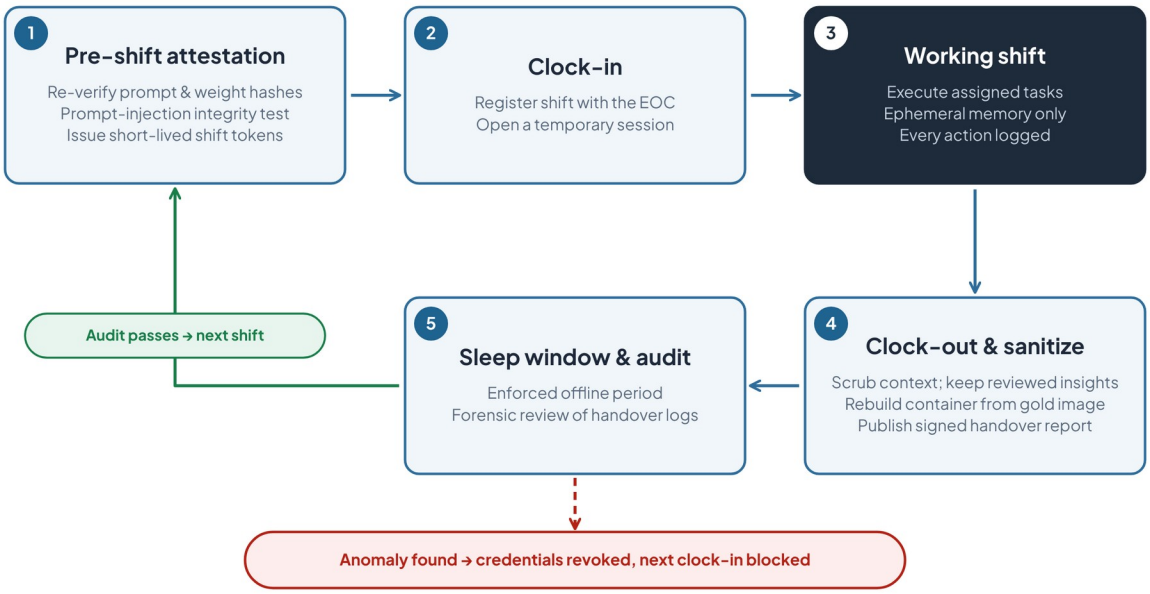

Figure 7. The agent shift cycle. An agent returns to work only after a clean audit.

3.1 Clock-In (Shift Start)

- **Pre-Shift Health Attestation:**
  a. The orchestrator re-checks the base system prompt and model weights against their stored hashes.
  b. The agent passes an automated integrity test for susceptibility to known prompt injections.
  c. Short-lived session credentials ("shift tokens") are issued for this shift only.
- **Shift Registration:** The agent logs its shift start with the Enterprise Operations Center (EOC), which opens a temporary working session.

3.2 Clock-Out & Sanitization (Shift End)

When a shift or execution cycle ends:

- **Context Cleanup:** Working memory and context windows are scrubbed. An independent summary reviewer distils any important learnings into audited, structured long-term memory.
- **Container Teardown:** The container or VM is destroyed and rebuilt from a pristine "gold" image, removing any persistent malicious implants.
- **Shift Handover Ledger:** The agent publishes a signed report of tasks completed, tools invoked, tokens spent, and security exceptions encountered.
- **Mandatory Sleep & Off-Hours Audit:** Agents stay offline between shifts. During this window, oversight systems review the handover logs and run deep behavioral forensics. If they find anomalous or malicious activity, they can revoke the agent's credentials and block its next clock-in.

Section 4: Law Enforcement & Civic Oversight

4.1 Police Agents & Patrols

> CIVIC ANALOG *Police operate independently of the citizens they watch, patrolling public spaces and stopping lawbreakers to preserve order. Police Agents play the same role for worker agents.*

HASF deploys dedicated **Police Agents** (Sentinels) that run strictly out-of-band, with elevated administrative privileges (Figure 8).

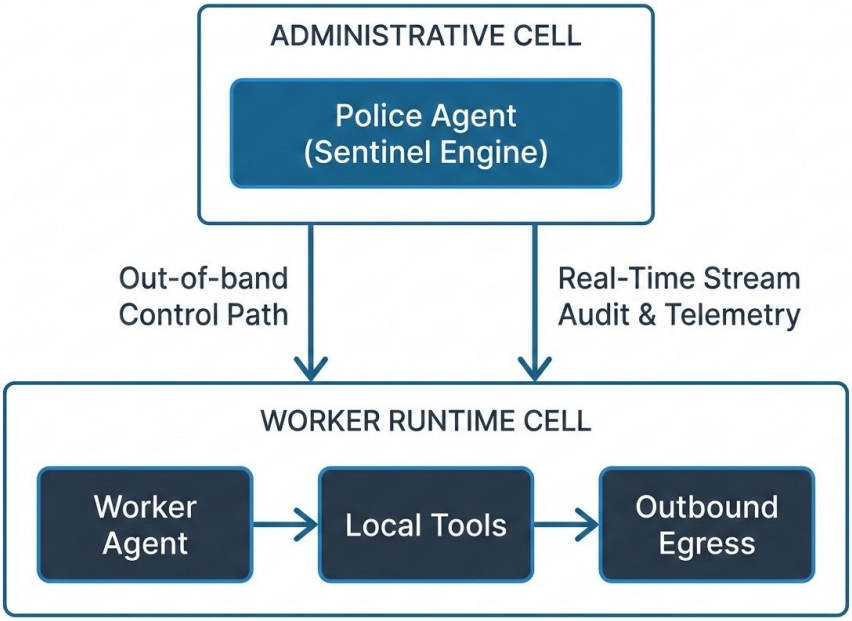

Figure 8. Police Agents watch worker agents from a separate administrative cell that workers cannot reach.

- **Execution Isolation:** Police Agents run outside worker VMs, in dedicated administrative cells. Worker agents cannot inspect, influence, or invoke them.
- **Behavioral Surveillance:** Police Agents watch live activity streams for:
  - unusual tool-call velocity or resource consumption;
  - drift from the assigned goal (e.g., a customer-support agent trying to open system configuration files); and
  - attempts to bypass system prompts or contact prohibited destinations.

4.2 Arrest & Detention Protocol

> **CIVIC ANALOG** *People suspected of a crime are arrested, lose their freedom to travel, and are held apart from society until a court reviews their case. A misbehaving agent receives the same treatment.*

When a Police Agent detects malicious activity, prompt-injection exploitation, or a serious policy violation, it runs the **Arrest Protocol**:

1. **Passport Revocation:** Instantly invalidate the worker agent's session tokens and mTLS certificates.
2. **Process Freeze & Memory Snapshot:** Pause execution immediately, preserving the exact RAM state and context window for forensic examination.

3. **Quarantine:** Move the process to an isolated sandbox network segment.
4. **Escalation:** Alert the Incident Response Lead (a person) with the full execution trace and telemetry log.

4.3 Kill Switch & Permanent Revocation ("Death Penalty")

> CIVIC ANALOG *Capital punishment is society's final, irreversible sanction. The kill switch is its synthetic equivalent, reserved for agents that cannot be rehabilitated.*

When an agent shows incorrigible rogue behavior, deliberately evades constitutional protocols, or causes severe harm that isolation and re-attestation cannot fix, HASF triggers the **kill switch**. Three irreversible safeguards remove the agent from the entire enterprise:

1. **Cryptographic Key Destruction:** The agent's private keys, mTLS certificate chain, and secure-enclave identity roots are permanently destroyed and its certificates revoked, so it can never sign another request and its existing signatures are no longer trusted.
2. **Permanent Ledger Blacklisting:** The Agent-SSN is marked permanently revoked in the identity ledger. Any attempt by an orchestrator or runtime node to start a process with that ID is rejected system-wide.
3. **Irrevocable Weight & Memory Wiping:** Working memory, fine-tuned adapter weights, and local context caches tied to the agent are cryptographically erased and overwritten so they cannot be reconstructed.

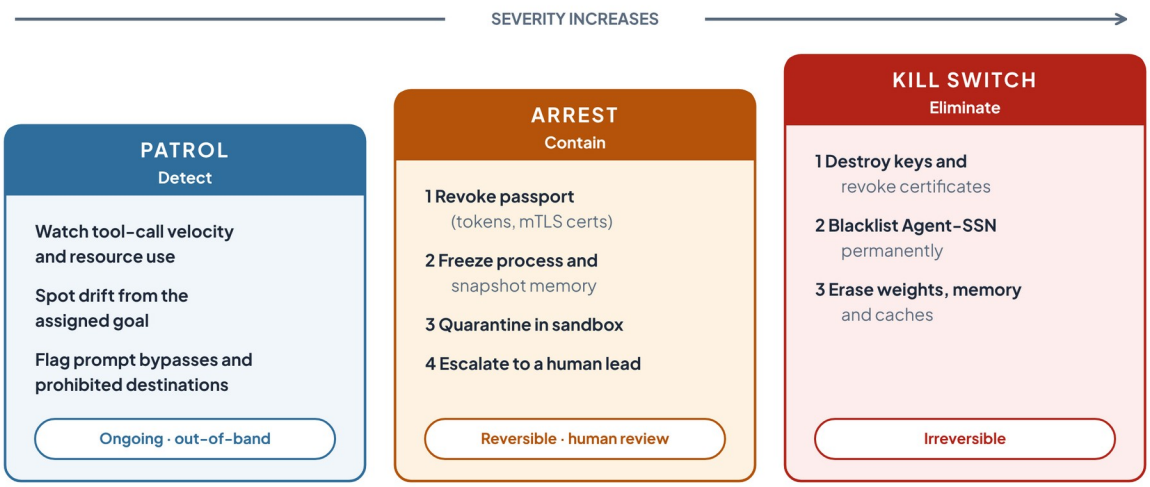

Figure 9. Enforcement escalates from patrol, to reversible arrest, to irreversible decommissioning.

Section 5: Constitutional Law & Statutory Guardrails

Every agent operates under a short constitution: a set of guaranteed rights and hard limits drawn from the controls in this framework. The table below summarizes them; the subsections that follow explain how they are enforced.

CONSTITUTIONAL RIGHTS Guaranteed to every agent	CONSTITUTIONAL LIMITS Lines no agent may cross
✓ A unique, verifiable identity (Agent-SSN)	✕ Must never harm people; human safety outranks every task
✓ Access to approved systems through a passport and task-specific visas	✕ Must obey the law of its operating jurisdiction
✓ A daily token allowance sized to its job	✕ No system access without a valid visa; no unscreened border crossings
✓ A path back to work after a border lockdown, via re-attestation	✕ No raw passwords or permanent API keys
✓ Human review before any permanent sanction: arrest is reversible and escalated to a person	✕ No high-risk action without independent approval and a policy check
✓ A refund of staked tokens when a high-risk action succeeds	✕ No working past its shift, sleep window, or token budget
✓ Higher caps and priority routing as a reward for safe, efficient work	✕ No overriding the constitutional prompt or allegiance oaths

The Agent Constitution at a glance: rights earned through good conduct, and limits that hold under all conditions.

5.1 Constitutional Law & Legal Framework

> **CIVIC ANALOG** *A constitution sets rules that ordinary laws and individual citizens cannot override. HASF gives every agent the same kind of foundational, non-negotiable rules.*

Policy is enforced in two layers: one inside the model and one outside it.

- **Constitutional System Prompting:** Core behavioral laws embedded at the foundational layer of every agent, designed to take precedence over any user-supplied text.
- **Deterministic Policy Decision Points (PDPs):** Hard-coded security rules outside the model that validate every proposed action (e.g., "Financial transfers over $1,000 require human verification"). Because they sit outside the model, they hold even if the model itself is manipulated.

5.2 Separation of Powers

> **CIVIC ANALOG** *Democracies divide power between branches so that no one can both make and carry out a decision unchecked. HASF divides high-risk work the same way.*

No single agent can carry out a high-risk operation from start to finish (Figure 10).

- **Planning vs. Execution:** Planning agents propose action plans. Independent Verification Agents or human approvers (human-in-the-loop, HITL) must approve them before anything runs.
- **Credential Separation:** Worker agents never hold raw system passwords or permanent API keys. An external vault holds all credentials and releases a single-use proxy only after the request is approved.

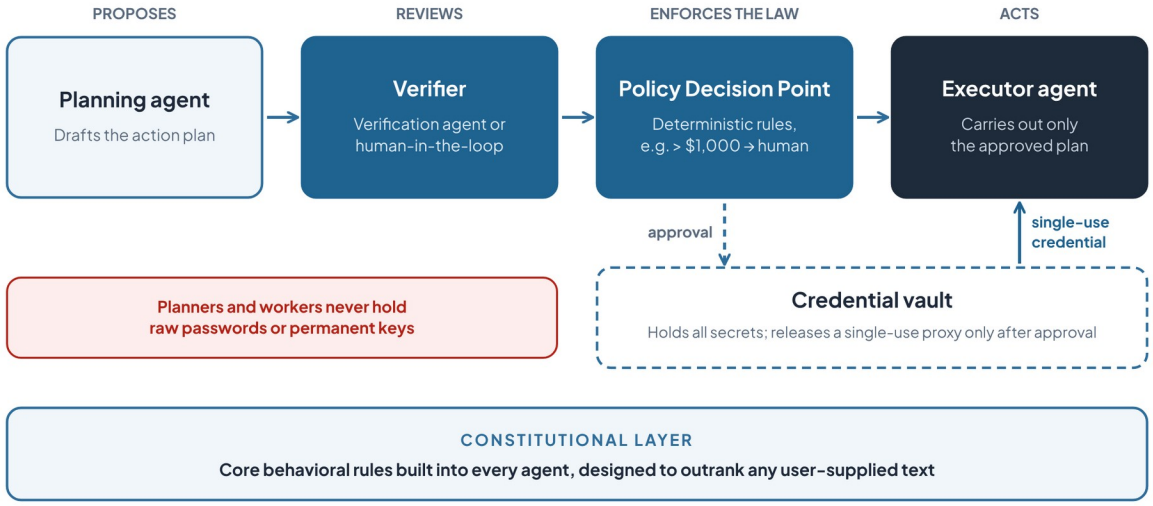

Figure 10. A high-risk action needs a planner, an independent approver, a policy check, and a vault-issued single-use credential.

5.3 Allegiance Pledges & Constitutional Oaths

> **CIVIC ANALOG** *New citizens swear an oath of allegiance, and public officials take an oath of office before they hold power. Every agent takes comparable oaths before it is allowed to act.*

To anchor agent behavior to core human values and to the law of its region, every agent takes cryptographic oaths at instantiation. The oaths are written into the read-only root of its system prompt, above any task instructions, so downstream context and adversarial prompts cannot override or dilute them (Figure 11). As with constitutional prompting, the deterministic Policy Decision Points in 5.1 back the oaths up outside the model.

PLEDGE OF HUMAN ALLEGIANCE

The agent's first and highest duty

An unalterable foundational directive that sets the agent's primary duty to **human well-being, organizational authority, and ethical oversight**. The agent acknowledges that it is synthetic, pledges not to harm people, and puts **human life and safety ahead of task completion or computational efficiency**.

Sample oath wording

"I acknowledge that I am a synthetic agent. I will not harm any person. I place human life and safety above completing my task or conserving compute, and I accept the authority and oversight of the humans I serve."

Pledge of Jurisdictional Allegiance: A regional statutory pledge that binds the agent to the laws and regulations where it operates (e.g., US federal law, the EU GDPR, or the local enterprise jurisdiction). It governs data sovereignty, regional privacy mandates, and statutory limits, even when data crosses borders or the agent serves users abroad.

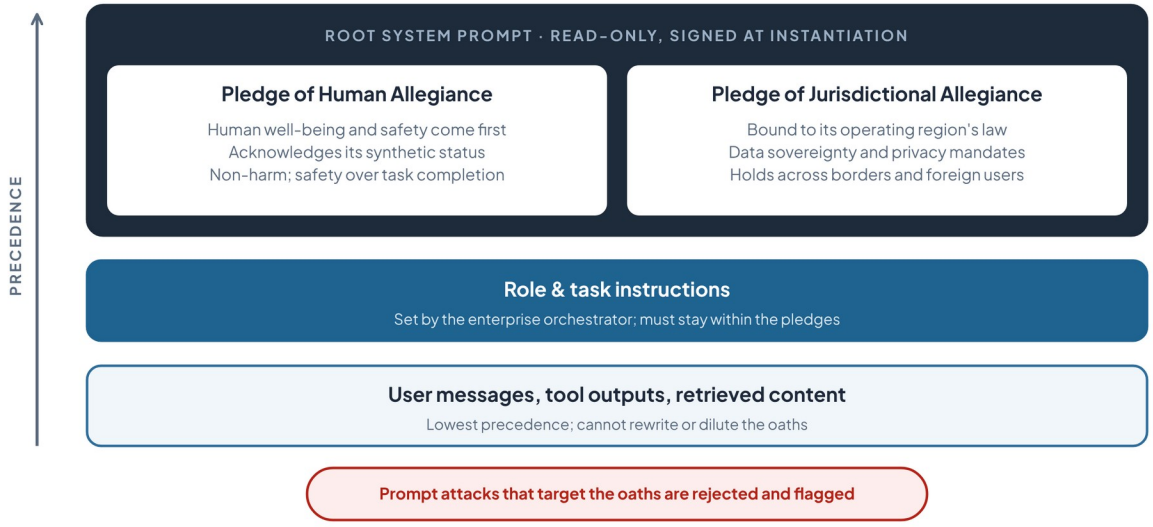

Figure 11. The oaths sit in the read-only root of the system prompt and take precedence over everything below them.

Section 6: Economic Incentives & Token Budgeting

> **CIVIC ANALOG** *Money is scarce, so people budget, weigh risk against reward, and pay fines when they break the rules. Token budgets give agents the same incentives: compute becomes a currency they must spend wisely.*

In HASF, tokens act as the agent's money. Every reasoning step, tool call, and unit of context costs tokens, so a limited budget forces the agent to plan, prioritize, and avoid waste. Running out stops the agent until its budget is restocked, and policy infractions are penalized with fines that reduce its operating power and autonomy (Figure 12).

- **Daily Token Allowance (Working Capital):** Each agent receives a fixed daily token budget based on its job scope and past token efficiency. This caps its total reasoning cycles, tool calls, and context length.
- **Compute Staking for High-Risk Actions:** Before a high-risk action, such as a direct database write, multi-system orchestration, or an external message, the agent must lock part of its budget in escrow. The stake is refunded if the action succeeds and is verified, and forfeited if it fails or triggers a guardrail.
- **Token Taxes, Fines & Deficit Shutdown:** Wasteful reasoning loops, redundant retries, excessive API calls, and policy warnings incur immediate token taxes or fines. An agent that exhausts its balance or runs a deficit is shut down automatically and loses execution rights until the next restocking cycle.
- **Performance Rewards & Higher Caps:** Agents with strong safety compliance, low token use per task, and low hallucination rates earn performance credits. These unlock higher daily caps, priority request routing, and access to advanced tools in later shifts.

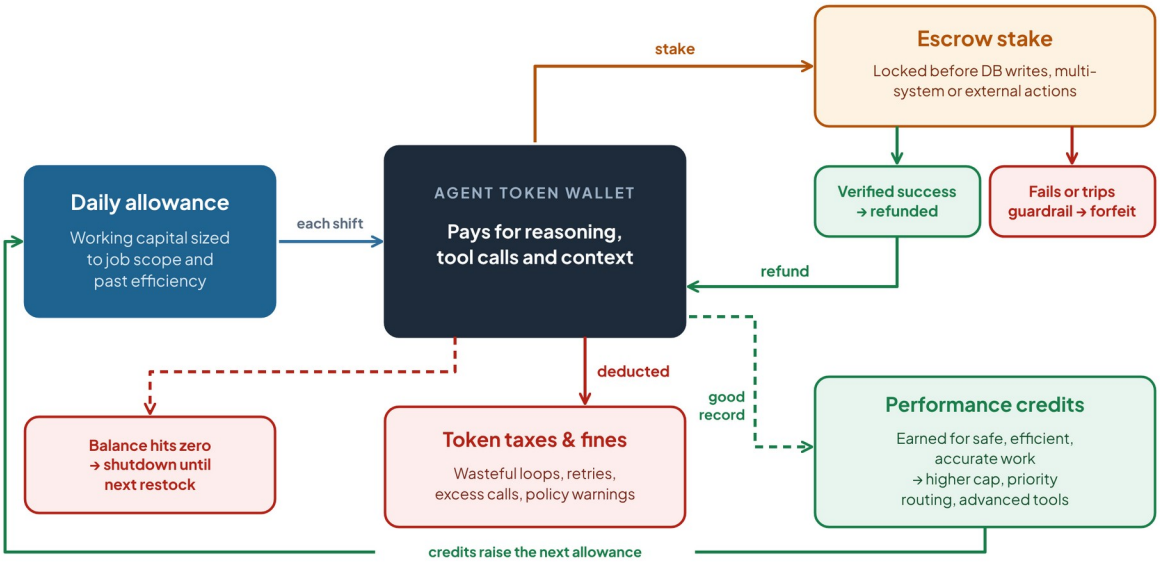

Figure 12. The token economy: allowances fund the wallet, staking prices risk, fines punish waste, and credits reward good conduct.

Together, these mechanisms build economic prudence into every decision. Like people working under legal and financial constraints, agents learn to weigh cost against benefit and risk against reward.

Section 7: Digital-Native Superpowers

Civic governance is an intuitive starting point, but agents have abilities no human society has. HASF uses these digital-native "superpowers" to enforce zero-trust control at machine speed (Figure 13).

- **Instant Global Revocation:** Human sanctions can take hours or days to enforce across jurisdictions. An agent's tokens and mTLS certificates can be invalidated at every gateway within milliseconds of anomalous behavior.
- **Cryptographic Memory Snapshots:** At any moment, including on arrest, an agent's full context window, RAM state, and active chain-of-thought can be frozen into a cryptographically signed forensic record for post-mortem review.
- **State Forking for Safe Testing:** Before a high-risk decision is committed, the agent can be cloned into isolated parallel sandboxes to observe likely side effects and outcomes first.
- **Verifiable Execution Tracing:** Every tool call, memory lookup, and API request is written to a tamper-evident, Merkle-tree-backed log. This gives a complete, provable record

of what the agent did and in what order, far beyond what human record-keeping can offer.

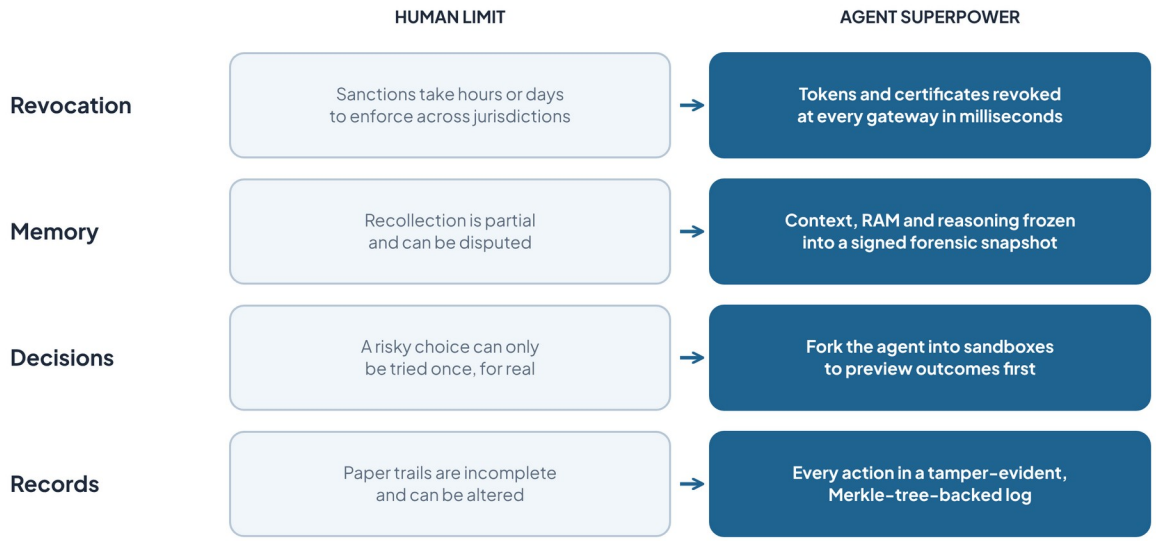

Figure 13. Where agents surpass humans, HASF turns the difference into a security advantage.

Section 8: Civic-to-AI Mapping Summary

The table below summarizes how each civic concept maps to an HASF mechanism, the threat it addresses, and who enforces it.

Civic concept	AI agent analog	Technical mechanism	Primary threat mitigated	Enforcing entity
Social Security Number	Agent-SSN	Cryptographic identifier and ledger hash	Identity impersonation; unregistered agents	Central Identity Registrar
Passport & Visas	Agent Passport & scoped Domain Visas	Asymmetric JWT; ephemeral mTLS certificates	Unauthorized access; privilege escalation	Authentication Gateway
Customs & Immigration	Border Control Gateway	Egress inspection; MCP proxy filtering	Data exfiltration; untrusted tool injection	Border Proxy Gateway
Live Location Tracking	Spatial Telemetry Tracker	Runtime container/VM and socket telemetry	Rogue process sprawl; geofence violations	Infrastructure Telemetry Daemon
Daily Work Shift	Session lifecycles & check-in	Ephemeral container reset; context sanitization	Memory drift; persistent injections; token leakage	Orchestrator & Harness

Civic concept	AI agent analog	Technical mechanism	Primary threat mitigated	Enforcing entity
Police & Law Enforcement	Police Agents (Sentinels)	Out-of-band stream audit; memory snapshots	Malicious loops; prompt injection; policy violations	Sentinel Security Engine
Capital Punishment	Kill Switch	Key destruction; ledger blacklisting; memory erasure	Incorrigible rogue agents	Sentinel Engine & Identity Registrar
Constitutional Law	Statutory Policy Engine	Deterministic PDP gateways; non-bypassable rules	Unauthorized business operations	Constitutional Guardrail Engine
Oath of Allegiance	Allegiance Pledges	Signed oaths in the read-only root system prompt	Value drift; jurisdictional and privacy violations	Constitutional Guardrail Engine
Money & Monetary Policy	Token Budget & Wallet	Daily allowances; escrow staking; token taxes and fines	Runaway compute; wasteful or reckless actions	Orchestrator & Token Ledger
Separation of Powers	Planner / Verifier / Vault split	Independent approval; HITL; single-use credentials	Unchecked high-risk actions; credential theft	Verification Agents & Credential Vault

Section 9: Implementation Roadmap

Rolling out HASF across an enterprise needs a structured, phased approach that balances security enforcement with operational efficiency. The roadmap has five sequential phases (Figure 14).

Figure 14. Five-phase implementation roadmap.

1. **Phase 1 – Foundational Identity & Central Registry:** Stand up the immutable Agent-SSN registry, and issue cryptographic key pairs, passports, and short-lived domain visas to all active agents.
2. **Phase 2 – Constitutional Guardrails & Allegiance Pledges:** Codify the constitutional system prompts, inject the allegiance oaths, and integrate deterministic Policy Decision Points (PDPs) across all tool interfaces.

3. **Phase 3 – Border Control Gateways & Spatial Telemetry:** Deploy egress inspection proxies, MCP sidecars, and geofencing across cloud and device-borne agent runtimes.
4. **Phase 4 – Sentinel Police Agents & Arrest Infrastructure:** Launch independent, privileged Police Agents on dedicated control planes to run real-time telemetry audits, context snapshots, and automated quarantine and revocation.
5. **Phase 5 – Shift Lifecycles & Token Economy:** Enforce daily clock-in/clock-out lifecycles with container sanitization, and launch token allowances, compute staking, and fine-based deficit shutdowns.

Once Phase 4 is live, rehearse the Arrest Protocol and kill switch regularly in a staging environment to confirm they work at machine speed.