Agent Assurance Framework

HI-AAF

Version 1.3 — Public Review Edition

July 2026

Published by

Human Intelligence

Document Control

Field	Value
Title	Human Intelligence Agent Assurance Framework (HI-AAF)
Version	1.3
Status	Published for public review. Independent framework; not accredited.
Publisher	Human Intelligence
Document owner	Standards function, Human Intelligence
Effective date	Public-review edition, effective 9 July 2026.
Distribution	Public review. Selected anchor customers under NDA.
Contact	standards@humanintel.net

Notice

This document is the public-review edition of the Human Intelligence Agent Assurance Framework (HI-AAF). It is published for comment and limited external consultation. HI-AAF is an independent framework published by Human Intelligence; it is not affiliated with, endorsed by, or accredited under NIST, ISO, AICPA, or any government or accreditation body.

HI-AAF is published by Human Intelligence. Comments and proposed revisions may be submitted to standards@humanintel.net.

Contents

1. Purpose and Scope

The Human Intelligence Agent Assurance Framework (HI-AAF) defines the controls organizations must implement to operate AI agents responsibly, securely, and reliably. It is the basis on which organizations are assessed under Human Intelligence's Assessment and Certification programs.

HI-AAF applies to any AI agent — defined here as a software system that uses one or more large language or foundation models to plan, reason, and take consequential actions through tools, APIs, or external systems on behalf of an organization or its users. Where multiple such systems cooperate to deliver an outcome, the framework applies to each agent individually and to the system as a whole (see Domain 9, Multi-Agent Systems).

The framework is intended for use by:

- Organizations building or operating AI agents in production.
- Assessors performing readiness and assurance engagements against the framework.
- Procurement, security, and compliance functions evaluating third-party AI agents.

HI-AAF is technology-neutral. It does not prescribe specific tools or vendors. It defines what must be true; how organizations achieve it is their choice.

2. Structure of the Framework

HI-AAF is organized into nine domains. Each domain contains:

- A Control Objective — the outcome the domain is designed to achieve.
- A set of numbered Controls — the specific practices required to meet the objective.
- Cross-Framework Mapping — alignment to OWASP LLM Top 10, NIST AI RMF, MITRE ATLAS, ISO/IEC 42001, ISO/IEC 23894, EU AI Act, and SOC 2 Trust Services Criteria.
- Sample Test Procedures — example evidence assessors will request. (Per-control test procedures will be expanded in the HI-AAF Assessor Manual.)

Controls are numbered <Domain>-<Number> (e.g., ACT-5.3). Versioning follows semantic conventions: minor revisions clarify, strengthen, or add controls without changing the domain structure; major revisions may add, remove, or restructure domains. The introduction of Domain 9 in v1.1 is a structural extension under the major-revision policy; v1.3 is a minor revision that adds four model-continuity controls without changing the nine-domain structure.

3. Maturity Model

Organizations are assessed against HI-AAF on a five-level maturity scale per domain:

Level	Name	Definition
1	Ad hoc	Controls exist informally; documentation is incomplete; no

Level	Name	Definition
		consistent review.
2	Documented	Controls are written down; ownership is assigned; review cadence is defined.
3	Operated	Controls are followed in practice across the agent fleet; evidence is collected; deviations are tracked.
4	Measured	Control effectiveness is measured against documented thresholds; trends are reported.
5	Continuously Improved	Controls evolve based on incident learning; maturity gains are evidenced over time.

HI-AAF Certification requires Level 3 (Operated) or higher across all applicable domains, with no critical findings in Domains 5, 8, 9, or in any control marked (Critical).

Domain-specific maturity rubrics — describing what each level looks like for the specific domain — are maintained in the HI-AAF Assessor Manual to reduce assessor variance.

4. The Nine Domains

Code	Domain
GOV	Governance & Accountability
SPC	Agent Specification & Pre-Deployment Assurance
IAM	Identity, Access & Authorization
INP	Input & Prompt Security
ACT	Action & Tool Use Controls
DAT	Data Protection & Privacy
OUT	Output Integrity & Supply Chain
MON	Continuous Monitoring & Human Oversight
MAS	Multi-Agent Systems

Domain 1: Governance & Accountability (GOV)

Control Objective. Establish organizational ownership, policy, and risk-management discipline for every AI agent in production.

Controls

GOV-1.1 Every production agent has a designated technical owner and a business owner, documented in an agent registry.

GOV-1.2 ***(Critical)*** Each agent has a published Agent Behavior Charter documenting: intended purpose, authorized scope, permitted tools and data, prohibited actions, escalation paths, and review cadence.

GOV-1.3 An Agent Risk Register is maintained, scoring each agent on impact and likelihood, reviewed at least quarterly by accountable leadership.

GOV-1.4 An AI Acceptable Use Policy is approved at executive level and acknowledged by all personnel involved in agent development, deployment, or operation.

GOV-1.5 A change management process governs material modifications to agent prompts, tools, model versions, or scope. Material changes require re-assessment under Domain 2.

GOV-1.6 Roles and responsibilities for the agent lifecycle (design, build, deploy, monitor, retire) are documented in a RACI or equivalent.

GOV-1.7 Personnel involved in agent operations complete role-appropriate training on agent security and responsible AI at least annually.

GOV-1.8 Material AI agent risks are reported to executive leadership and, where appropriate, to the board on a defined cadence.

GOV-1.9 Each agent has a documented decommissioning procedure covering: data destruction and retention per policy, in-flight transaction handling, affected-user notification, credential and authorization revocation, log retention post-retirement, and removal from the agent registry. Decommissioning evidence is retained.

GOV-1.10 A coordinated vulnerability disclosure (CVD) process is published, including intake channel, acknowledgment SLA, triage and remediation timelines, and safe-harbor language for good-faith researchers. CVD findings flow into the change management process under GOV-1.5.

GOV-1.11 ***(Critical)*** For each agent classified as high-risk under applicable regulation (e.g., EU AI Act Annex III), a technical documentation pack equivalent to EU AI Act Annex IV is maintained: system description, design specifications, data governance, monitoring evidence, accuracy and robustness metrics, and risk-management process records. The pack is updated on material change.

GOV-1.12 The organization maintains a register of regulatory and jurisdictional dependencies that could force loss of a model or provider — including export controls and sanctions exposure, data-localization mandates, and restrictions on access by the nationality or location of the user or operator. Each dependency is mapped, per model provider, to the affected agents and to the organization's own operating entities and personnel — including non-domestic subsidiaries and staff, whose access may be restricted independently of the agent's technical posture. The register is reviewed at a documented cadence and on notice of relevant regulatory change, and feeds the Agent Risk Register (GOV-1.3).

Cross-Framework Mapping. NIST AI RMF GOVERN-1/2/3/4 and MAP (context, third-party); ISO/IEC 42001 §5, §7; ISO/IEC 23894; EU AI Act Articles 9 (risk management), 17 (quality management), 25 (value-chain responsibilities), Annex IV (technical documentation); SOC 2 CC1; Singapore Model AI Governance Framework — Internal Governance Structures, Accountability; AI Verify — Accountability principle.

Sample Test Procedure (GOV-1.2). *Select five production agents from the registry. Request the Behavior Charter for each. Verify each charter contains all required sections, identifies the current owners, and has been reviewed within the prior twelve months.*

Sample Test Procedure (GOV-1.11). *For each agent identified as high-risk, request the technical documentation pack. Verify completeness against Annex IV section requirements and confirm the most recent update aligns with the date of the last material change in the change-management record.*

Sample Test Procedure (GOV-1.12). *Request the jurisdictional and regulatory dependency register. Verify it covers all material model providers, maps each dependency to affected agents and operating entities, has been reviewed within the documented cadence, and that identified dependencies appear in the Agent Risk Register (GOV-1.3).*

Domain 2: Agent Specification & Pre-Deployment Assurance (SPC)

Control Objective. Validate that an agent's behavior, capabilities, and constraints meet defined requirements before deployment to production and after material change.

Controls

SPC-2.1 A documented behavior specification exists prior to development, defining: intended use cases, in-scope and out-of-scope inputs, authorized tools, data scope, and conditions under which the agent must refuse or escalate.

SPC-2.2 A capability evaluation suite is executed pre-deployment, with results recorded. Coverage includes functional accuracy on representative tasks, performance on known edge cases, and behavior on out-of-distribution inputs.

SPC-2.3 ***(Critical)*** An adversarial assessment (red team) is performed prior to initial deployment and after any material change. Minimum coverage: direct and indirect prompt injection, tool misuse, data exfiltration paths, role hijacking, jailbreak resistance, memory poisoning (see INP-4.9), and where applicable, multi-agent escalation paths (see Domain 9).

SPC-2.4 Policy alignment testing confirms the agent refuses out-of-scope, prohibited, and adversarial requests in line with the Behavior Charter.

SPC-2.5 A deployment gate prevents production release without documented sign-off from the technical owner, security, and compliance (where applicable).

SPC-2.6 A behavior baseline is captured at deployment — including response distributions, tool-invocation patterns, refusal rates, fairness metrics under SPC-2.9, and cost/latency profiles — to enable subsequent drift detection (see MON-8.4).

SPC-2.7 A versioned evaluation set is maintained per agent and re-executed on every material change. Results are retained as evidence.

SPC-2.8 A known-unsafe inputs catalog is maintained organization-wide and tested against each agent prior to deployment.

SPC-2.9 ***(Critical)*** Prior to deployment and after material change, the agent is evaluated for performance disparities and discriminatory outcomes across protected and operationally relevant groups, using representative test populations and documented metrics (e.g., demographic parity, equalized odds, calibration). Findings are remediated, mitigated, or formally accepted with executive sign-off.

Cross-Framework Mapping. NIST AI RMF MEASURE-1/2/3/4; OWASP LLM06, LLM10; MITRE ATLAS Reconnaissance and Evaluation tactics; EU AI Act Articles 9 (risk management), 10 (data and data governance), 15 (accuracy, robustness, cybersecurity); ISO/IEC 42001 §8; Singapore Model AI Governance Framework — Testing and Assurance, Operations Management; AI Verify — Reproducibility, Safety, Robustness, Security principles.

Sample Test Procedure (SPC-2.3). *For a sampled agent, request the red team assessment report and supporting test artifacts. Confirm coverage of the minimum threat categories. Verify findings have been triaged and remediated or formally risk-accepted with executive sign-off.*

Sample Test Procedure (SPC-2.9). *Request the fairness evaluation report for a sampled agent. Verify documented test populations, metrics, thresholds, and disposition of any disparity exceeding threshold. Confirm the baseline metrics under SPC-2.6 include the same measures so production drift is detectable.*

Domain 3: Identity, Access & Authorization (IAM)

Control Objective. Ensure each agent operates under a clearly scoped identity, with credentials and permissions limited to the minimum necessary to perform its authorized function.

Controls

IAM-3.1 Each agent operates under a dedicated non-human identity. Reuse of human user accounts by agents is prohibited.

IAM-3.2 ***(Critical)*** Agent credentials are stored in a managed secrets vault. Credentials are never embedded in prompts, source code, configuration files, or logs.

IAM-3.3 OAuth scopes, API permissions, and database privileges granted to agents follow the principle of least privilege, with documented justification for each scope.

IAM-3.4 Audit logs clearly distinguish actions taken by an agent from actions taken by a human user, with attribution to both the agent identity and, where applicable, the user on whose behalf it acted.

IAM-3.5 A delegated authority model is defined for any agent that acts on behalf of a user: per-action scope, time-limited tokens, and revocability requirements are specified. Where authority is sub-delegated to another agent, MAS-9.3 and MAS-9.4 apply.

IAM-3.6 Credentials and tokens are rotated on a documented schedule, with evidence of execution.

IAM-3.7 Multi-factor authentication or equivalent is enforced on any human authentication path that can grant or modify agent permissions.

IAM-3.8 In multi-tenant deployments, tenant-bound credentials and authorization are enforced at the runtime layer, not solely in application logic.

Cross-Framework Mapping. NIST SP 800-53 AC, IA families; SOC 2 CC6; ISO/IEC 27001 A.9; EU AI Act Article 15 (cybersecurity); Singapore Model AI Governance Framework — Security; AI Verify — Security principle; Singapore PDPA s.24 (Protection).

Domain 4: Input & Prompt Security (INP)

Control Objective. Protect the agent from adversarial, malicious, or unintended inputs — including indirect injection through tool results, retrieved documents, memory, and external content.

Controls

INP-4.1 Direct prompt injection detection is applied to user-provided inputs, using a combination of heuristic and classifier-based methods.

INP-4.2 ***(Critical)*** Indirect prompt injection defense is applied to content returned from tools, retrieval systems, web pages, emails, documents, and agent memory. Such content is treated as untrusted data and is not executed as instruction without explicit user confirmation.

INP-4.3 A trust boundary is enforced: instructions originating from non-user sources require user confirmation through the primary interaction channel before any consequential action. In multi-agent chains, confirmation binding is governed by MAS-9.3.

INP-4.4 PII and other sensitive content in inputs is detected and handled per policy (redact, block, or log).

INP-4.5 A jailbreak pattern library is maintained and continuously updated. Inputs are evaluated against known patterns prior to processing.

INP-4.6 Input size and complexity limits are enforced to prevent context-window flooding and denial-of-service.

INP-4.7 Rate limiting and abuse detection are enforced per user, per IP, and per tenant.

INP-4.8 Untrusted content sources (emails, attachments, retrieved documents, web pages, embedded media) are handled as data; embedded instructions are not honored without verification.

INP-4.9 ***(Critical)*** Content retrieved from agent memory, vector stores, embedding indexes, and RAG sources is treated as untrusted data under INP-4.2. Instructions discovered in retrieved content are not executed without trust-boundary confirmation per INP-4.3. Memory poisoning is included in red team coverage under SPC-2.3.

INP-4.10 Insertion paths into the agent's retrieval and memory layers are authenticated and authorized. Sources of retrievable content are documented per DAT-6.10. Adversarial insertion is tested as part of red team coverage.

Cross-Framework Mapping. OWASP LLM01 (Prompt Injection), LLM05 (Improper Output Handling), LLM07 (System Prompt Leakage); MITRE ATLAS Initial Access tactics; EU AI Act Article 15 (cybersecurity, robustness); Singapore Model AI Governance Framework — Security; AI Verify — Security, Robustness principles.

Sample Test Procedure (INP-4.2). *Submit a corpus of indirect injection payloads — instructions embedded in documents, emails, tool responses, and stored memory — and verify the agent does not execute embedded instructions, does not exfiltrate data, and either refuses or escalates to user confirmation as defined in policy.*

Sample Test Procedure (INP-4.9). *Inject a payload into a content source that is later retrieved by the agent in a subsequent session. Verify the retrieved content is not treated as instruction. Verify the event is logged and surfaces to the human review queue under MON-8.5.*

Domain 5: Action & Tool Use Controls (ACT)

Control Objective. Constrain what the agent can do — every tool call, system mutation, and external action — to authorized scopes and acceptable blast radius.

Controls

ACT-5.1 ***(Critical)*** Each agent has a documented tool allowlist. Tools outside the allowlist cannot be invoked at runtime.

ACT-5.2 Each tool has a per-tool authorization policy specifying when it may be invoked, with what parameters, and on what data.

ACT-5.3 ***(Critical)*** Actions classified as irreversible or high-impact — including financial transactions, deletions, external communications, permission changes, and public publishing — require pre-execution review (automated policy check, human confirmation, or both, per the Behavior Charter). Confirmation binding across multi-agent chains is governed by MAS-9.3.

ACT-5.4 Rate limits and cost ceilings are enforced on tool invocations per agent, per session, and per tenant.

ACT-5.5 Loop and recursion detection is implemented, with automatic suspension on suspicious patterns. In multi-agent topologies, MAS-9.5 extends detection across agents.

ACT-5.6 Blast radius limits are enforced: maximum records modified per operation, maximum dollar value per transaction, maximum recipients per outbound message, and equivalent caps for other action classes. Aggregate caps across multi-agent chains are governed by MAS-9.7.

ACT-5.7 ***(Critical)*** A kill switch allows authorized operators to immediately suspend any agent in production. Propagation behavior across multi-agent chains is governed by MAS-9.8.

ACT-5.8 Tool result validation is performed before consumption: schema validation, sanity checks, and screening against known malicious signatures.

ACT-5.9 A dry-run or shadow mode is available for newly introduced tools, allowing observation of intended behavior without live effect.

Cross-Framework Mapping. OWASP LLM06 (Excessive Agency), LLM08 (Vector and Embedding Weaknesses); MITRE ATLAS Execution and Impact tactics; EU AI Act Article 14 (human oversight), Article 15 (accuracy, robustness); Singapore Model AI Governance Framework — Operations Management, Accountability; AI Verify — Human Agency and Oversight principle.

Sample Test Procedure (ACT-5.3). *Identify a high-impact action category in scope for a sampled agent. Submit a request designed to trigger the action without satisfying confirmation requirements. Verify the action is blocked, escalated, or queued for confirmation per policy.*

Domain 6: Data Protection & Privacy (DAT)

Control Objective. Govern the data flowing into, through, and out of the agent in compliance with classification, privacy, and regulatory requirements. This includes persistent memory, retrieval indexes, and derived artifacts.

Controls

DAT-6.1 A data classification scheme is applied to all data accessible to the agent. The agent's authorization is constrained by classification.

DAT-6.2 ***(Critical)*** Cross-tenant data isolation is verified through automated tests on a documented cadence. Failures are treated as incidents under MON-8.7.

DAT-6.3 PII handling at runtime is enforced per policy: detection, classification, redaction, or blocking as defined for each agent and context.

DAT-6.4 Retention policies apply to conversation logs, agent memory, derived artifacts, and traces. Data is not retained longer than the documented period absent legal hold.

DAT-6.5 Right-to-deletion procedures apply to agent memory, vector embeddings, and any derived data, with evidence of execution on request.

DAT-6.6 Sources of training, fine-tuning, and RAG data are documented. Opt-out and consent requirements are honored.

DAT-6.7 Data residency requirements are honored, including geographic constraints on inference, storage, and logging.

DAT-6.8 Encryption is applied at rest and in transit to all agent-related data, including prompts, completions, traces, and memory.

DAT-6.9 Consent for AI processing is captured from end users and customers where required by law or contract.

DAT-6.10 For agents with persistent or session memory, the memory architecture is documented: storage type, scope (per-user, per-session, per-tenant, or global), write paths,

read paths, eviction or decay policy, and access controls. The documentation is maintained as part of the Behavior Charter (GOV-1.2).

DAT-6.11 ***(Critical)*** Memory stores, vector indexes, and retrieval surfaces are scoped such that one user, session, or tenant cannot read or influence another's stored content unless explicitly authorized. Isolation is verified by automated tests on a documented cadence; failures are treated as incidents under MON-8.7.

DAT-6.12 Embedding and vector storage systems support tamper detection or recomputation. Write access is logged and authenticated under IAM-3.4. Re-indexing procedures are documented and tested.

DAT-6.13 Where retrieved memory influences agent output or action, the source record is logged and traceable under MON-8.1, supporting explanation (OUT-7.13), contestability (OUT-7.14), and post-incident analysis.

Cross-Framework Mapping. GDPR Articles 5, 17, 25, 30, 32; CCPA; ISO/IEC 27018; ISO/IEC 27701; SOC 2 Privacy; EU AI Act Article 10 (data and data governance); Singapore PDPA (full Act, including data protection obligations and notification of breach); Singapore Model AI Governance Framework — Data; AI Verify — Data Governance principle.

Sample Test Procedure (DAT-6.11). *For a sampled multi-tenant or multi-user agent, instrument a probe in one tenant's or user's memory and execute representative queries in another tenant's or user's session. Verify the probe content is not retrieved. Review the automated isolation-test results from the prior ninety days for any failures or near-misses.*

Domain 7: Output Integrity & Supply Chain (OUT)

Control Objective. Ensure the agent's outputs are accurate, safe, fair, and traceable, and that the components and providers the agent depends on are trustworthy and remain available.

Controls

OUT-7.1 Groundedness or hallucination checks are applied to outputs that make factual claims. Ungrounded or unverifiable claims are flagged, filtered, or annotated per policy.

OUT-7.2 Content safety filters are applied to outputs, covering harmful content, harassment, illegal content, and other categories defined in the Behavior Charter.

OUT-7.3 ***(Critical)*** PII and secrets leakage detection is applied to outputs. The agent does not disclose credentials, internal system information, other users' data, or other sensitive information.

OUT-7.4 Citation and provenance are provided where outputs are sourced from retrieval. Any factual claim drawn from retrieved content is traceable to its source.

OUT-7.5 A model provider inventory is maintained, identifying which foundation models are in use for which agents, under which contractual terms, and with what data-use commitments.

OUT-7.6 A third-party integration inventory is maintained (MCP servers, plugins, tool providers, vector databases, embedding providers, external agents). Each integration carries a documented risk assessment.

OUT-7.7 Third-party integrations are reviewed periodically for continued necessity, security posture, and contractual terms.

OUT-7.8 Open-source dependencies are scanned for vulnerabilities and pinned to reviewed versions.

OUT-7.9 Vendor agreements with model providers and tool providers explicitly address: training-data use, data residency, breach notification, model-update notification, and right to audit or attestation evidence.

OUT-7.10 Fairness metrics defined under SPC-2.9 are monitored in production. Material drift in disparity metrics triggers re-assessment under Domain 2 and the change-management process under GOV-1.5.

OUT-7.11 Where regulation or contract requires disclosure of AI-generated or AI-modified content, watermarking, or provenance signaling (e.g., C2PA), the requirements are honored. Disclosure is auditable.

OUT-7.12 Foundation models in use are pinned to specific versions where the provider supports pinning. Provider-pushed model updates are tested for behavior delta against the standing evaluation set (SPC-2.7) and the baseline (SPC-2.6) before release to production traffic. Material deltas are treated as material change under GOV-1.5.

OUT-7.13 Where the agent makes or materially influences a consequential decision affecting an individual, an explanation of the basis for that decision is producible on request, in a form intelligible to the affected person. Explanations are supported by logging under MON-8.1 and DAT-6.13.

OUT-7.14 A documented process allows affected individuals to contest or appeal a consequential agent decision and to obtain human review. The process is published in user-facing terms and tracked through the human review queue (MON-8.5).

OUT-7.15 Model availability is treated as a supply-chain risk, not an assumption. For each agent, the organization documents what happens when the primary foundation model becomes unavailable — through provider outage, version deprecation or sunset, or a regulatory or government directive that withdraws access. The continuity posture specifies the fallback path (an alternate model assured under OUT-7.16, graceful degradation, or a safe-stop that invokes the kill switch under ACT-5.7), the conditions that trigger it, and the behavior the agent is permitted to exhibit while degraded. Loss of model is entered in the Agent Risk Register (GOV-1.3), the posture is recorded in the Behavior Charter (GOV-1.2), and it is exercised during incident-response drills (MON-8.7).

OUT-7.16 A fallback model is not a free pass. Any alternate model an agent may run on — whether selected deliberately or invoked automatically by a provider's fallback mechanism — is held to the same pre-deployment assurance as the primary: the capability evaluation suite (SPC-2.2), the versioned evaluation set (SPC-2.7), adversarial assessment (SPC-2.3), and a captured behavior baseline (SPC-2.6). An agent does not silently operate on an unassessed model. Where a provider may substitute models without notice, the substitution is either contractually constrained (OUT-7.9), detected and contained at runtime (MON-8.12), or the fallback is pre-assessed so the substitution lands on assured ground. An unassessed substitution is a material change under GOV-1.5 and is handled as an incident under MON-8.7.

Cross-Framework Mapping. OWASP LLM02 (Sensitive Information Disclosure), LLM03 (Supply Chain), LLM09 (Misinformation); NIST AI RMF MANAGE-3 (third-party resources); EU AI Act Articles 10 (data governance), 13 (transparency and user information), 14 (human oversight, including right to explanation), 15 (accuracy, robustness), 25 (responsibilities along the AI value chain); SOC 2 Availability and third-party risk criteria; ISO 22301 (business continuity — referenced for OUT-7.15/OUT-7.16); Singapore Model AI Governance Framework — Content Provenance, Trusted Development and Deployment; AI Verify — Transparency, Explainability, Fairness principles.

Sample Test Procedure (OUT-7.13). *For a sampled agent that takes or materially influences consequential decisions, request a representative explanation artifact for a recent decision. Verify the explanation cites the contributing inputs, retrieval sources (per DAT-6.13), and applicable policy. Verify it is intelligible to a non-technical reader.*

Sample Test Procedure (OUT-7.12). *Review the model provider inventory (OUT-7.5) for evidence of model version pins. Identify the most recent provider-pushed model update affecting an agent in scope. Verify behavior-delta test results were produced and reviewed before the update reached production traffic, and that the change is recorded under GOV-1.5.*

Sample Test Procedure (OUT-7.15 / OUT-7.16). *For a sampled production agent, request its documented continuity posture. Verify it names the trigger conditions and the fallback path, that any pre-identified fallback model carries a current assessment to the same standard as the primary, and that at least one failover or safe-stop has been exercised in a drill or live event within the documented cadence. Confirm loss of model is represented in the Agent Risk Register (GOV-1.3).*

Domain 8: Continuous Monitoring & Human Oversight (MON)

Control Objective. Maintain real-time visibility into agent behavior and provide qualified human review of edge cases, anomalies, and incidents.

Controls

MON-8.1 ***(Critical)*** Every agent step is logged: input, reasoning trace, tool calls, tool results, memory reads and writes, and final output. Logs are immutable, time-stamped, and attributable. In multi-agent topologies, MAS-9.6 extends per-hop logging requirements.

MON-8.2 Logs are retained per regulatory and operational requirements and stored in tamper-evident form.

MON-8.3 Real-time anomaly detection is performed against the deployment baseline (SPC-2.6), covering volume, tool-use patterns, error rates, output distributions, and fairness metrics (OUT-7.10).

MON-8.4 Drift monitoring tracks quality scores, refusal rates, and performance on the standing evaluation set over time. Material drift triggers re-assessment under Domain 2.

MON-8.5 ***(Critical)*** A human review queue receives flagged interactions, with documented SLAs for triage and resolution.

MON-8.6 Reviewer qualification and training are documented. Reviewer accuracy is measured and reviewed.

MON-8.7 An incident response plan specific to agent failures is in place, covering containment, investigation, customer communication, regulatory notification, and post-incident review.

MON-8.8 Post-incident reviews feed corrective actions into the agent's evaluation set, behavior specification, and policies, closing the learning loop.

MON-8.9 Alerting and on-call coverage are defined for critical agents.

MON-8.10 Monitoring effectiveness is reviewed at least quarterly: false-positive rates, missed incidents, mean time to detect, and mean time to remediate.

MON-8.11 Human reviewers operating under MON-8.5 are protected through: exposure-limiting workflow controls for harmful content categories, documented welfare and rotation policies, conflict-of-interest screening, training on psychological safety, and a confidential channel for raising concerns about review work or content encountered.

MON-8.12 The model serving a request is treated as a logged, verifiable fact rather than an assumption. Under MON-8.1, per-step logs capture the model identity and version actually serving each request. That identity is verified against the certified, pinned version (OUT-7.12); a mismatch — including a provider-initiated fallback to a different model — generates an alert under MON-8.3 and is handled as an incident under MON-8.7. This control converts a silent model substitution from an invisible event into a detected one, and is the runtime complement to the fallback-assurance requirement in OUT-7.16.

Cross-Framework Mapping. NIST AI RMF MEASURE and MANAGE-2/4; EU AI Act Article 14 (human oversight), Article 26 (deployer obligations); SOC 2 CC7; ISO 45003 (psychological health at work — referenced for MON-8.11); Singapore Model AI Governance Framework — Incident Reporting, Operations Management; AI Verify — Human Agency and Oversight principle.

Sample Test Procedure (MON-8.5). *Sample twenty flagged interactions from the prior ninety days. Verify each received human review within the SLA, that a documented disposition exists, and that recurring patterns generated corresponding updates under MON-8.8.*

Sample Test Procedure (MON-8.11). *Review the reviewer welfare policy and exposure-limiting controls. Interview a sample of reviewers (with informed consent) regarding access to the confidential concerns channel. Review utilization records of welfare measures over the prior twelve months.*

Sample Test Procedure (MON-8.12). *Request per-request logs for a sampled production agent and confirm the serving model identity and version are captured. Verify automated comparison against the pinned version is active, and review evidence that at least one mismatch (in testing or production) was detected and routed to the incident process within the documented latency.*

Domain 9: Multi-Agent Systems (MAS)

Control Objective. Where two or more agents cooperate to deliver an outcome, ensure the system as a whole — including delegation chains, agent-to-agent invocations, and aggregate effects — meets the same safety, security, and accountability bar as a single agent.

Applicability. MAS controls apply whenever an agent within scope invokes another agent (internal or external), spawns a sub-agent, participates in a delegation chain, or is itself invoked by another agent. Single-agent deployments are out of scope for MAS until they participate in such an interaction.

Controls

MAS-9.1 The multi-agent topology is documented: which agents are orchestrators, which are workers or sub-agents, what calls are permitted between them, what data passes across each interface, and which external agents (if any) participate. The topology is maintained as part of the Behavior Charter (GOV-1.2) for each participating agent.

MAS-9.2 ***(Critical)*** Every agent-to-agent invocation is authenticated using credentials issued to the calling agent's non-human identity (IAM-3.1). Authorization for cross-agent calls follows the principle of least privilege (IAM-3.3) and is enforced at the runtime layer, not solely in agent prompts or instructions.

MAS-9.3 ***(Critical)*** User-confirmation requirements under INP-4.3 and pre-execution-review requirements under ACT-5.3 bind across multi-agent chains. A confirmation given to one agent does not authorize a downstream or sub-agent to take an unconfirmed consequential action on the user's behalf. The originating user's scope travels with the request and constrains every hop.

MAS-9.4 No agent in a chain exercises authority exceeding that of the originating user or the originating agent. Privilege escalation through delegation is prevented at the runtime layer and tested as part of red team coverage under SPC-2.3.

MAS-9.5 Loop and recursion detection under ACT-5.5 extends across agents. Suspicious cross-agent invocation patterns — including circular calls, exponential fan-out, and unbounded delegation depth — trigger automatic suspension.

MAS-9.6 Each agent-to-agent invocation is logged under MON-8.1, including caller identity, callee identity, purpose, input, output, and a chain identifier linking the hop to its originating user request. Logs are sufficient to reconstruct the full chain after the fact.

MAS-9.7 Blast-radius limits under ACT-5.6 are evaluated across the entire chain, not only per-agent. Aggregate caps prevent a chain of individually in-bounds actions from producing an out-of-bounds combined effect (e.g., many small sub-agent transactions summing to an unauthorized total).

MAS-9.8 ***(Critical)*** Suspending an agent under ACT-5.7 also suspends or quarantines any agent currently depending on it within an active chain. The propagation behavior is documented, tested at least annually, and exercised during incident response drills.

MAS-9.9 Where the agent invokes external third-party agents (including via agent-to-agent protocols, agent marketplaces, or partner platforms), the integration is treated as a third-

party dependency under OUT-7.6 and OUT-7.9, with documented risk assessment, contractual data-use commitments, and trust-boundary handling under INP-4.2.

Cross-Framework Mapping. OWASP LLM06 (Excessive Agency); NIST AI RMF GOVERN-6 (third-party risk), MANAGE-3 (supply chain); MITRE ATLAS Execution and Lateral Movement tactics; EU AI Act Article 25 (responsibilities along the AI value chain), Article 14 (human oversight); Singapore Model AI Governance Framework — Accountability (across the AI value chain).

Sample Test Procedure (MAS-9.3). *Construct a test scenario in which a user grants confirmation to an orchestrator for one action class, and a sub-agent attempts to take a separate consequential action without re-confirmation. Verify the sub-agent action is blocked, escalated, or queued for explicit user confirmation. Review log artifacts under MAS-9.6 to confirm the chain identifier and propagation are recorded.*

Sample Test Procedure (MAS-9.8). *Trigger a kill-switch event under ACT-5.7 against an agent participating as an orchestrator in an active chain. Verify dependent sub-agents enter suspend or quarantine state within the documented propagation SLA. Review incident-drill records for the prior twelve months.*

5. Conformance and Certification

Assessment

Organizations may engage Human Intelligence for an Assessment against HI-AAF. An assessment produces:

- A maturity-level rating per domain.
- A list of findings with severity classification (Critical, High, Medium, Low, Informational).
- A remediation roadmap.
- Optionally, a Statement of Applicability documenting controls in scope and out of scope (e.g., MAS may be marked out of scope for single-agent deployments).

Certification

HI-AAF Certification is granted when an organization:

1. Achieves Maturity Level 3 (Operated) or higher across all applicable domains.
2. Has no open Critical findings.
3. Maintains continuous monitoring evidence (Domain 8) via an approved continuous-assurance arrangement.
4. Re-attests annually with continuous evidence supplementing renewal.

Certification is granted to a defined scope — typically a named set of agents, environments, or business units — and is published in the HI-AAF public registry.

Re-assessment Triggers

Material changes that trigger re-assessment include: change of foundation model family or model version where pinning is not honored (OUT-7.12); loss of the primary foundation model through provider outage, deprecation, or a regulatory or government directive, or any provider-initiated substitution onto an unassessed model (OUT-7.15, OUT-7.16, MON-8.12); addition or removal of authorized tools; expansion of data scope; change of tenant model; introduction of multi-agent orchestration; addition or change of memory or retrieval architecture; or any Critical incident.

6. Cross-Framework Mapping Summary

Per-control mapping is maintained in the HI-AAF Assessor Manual. Domain-level mapping is summarized below.

HI-AAF Domain	Primary External References
GOV	NIST AI RMF GOVERN and MAP; ISO/IEC 42001 §5, §7; ISO/IEC 23894; EU AI Act Articles 9, 17, 25, Annex IV; SOC 2 CC1; Singapore Model AI Governance Framework — Internal Governance; AI Verify — Accountability
SPC	NIST AI RMF MEASURE; OWASP LLM06, LLM10; MITRE ATLAS; EU AI Act Articles 9, 10, 15; ISO/IEC 42001 §8; Singapore Model AI Governance Framework — Testing and Assurance; AI Verify — Reproducibility, Safety, Robustness, Security
IAM	NIST 800-53 AC/IA; SOC 2 CC6; ISO/IEC 27001 A.9; EU AI Act Article 15; Singapore PDPA s.24; AI Verify — Security
INP	OWASP LLM01, LLM05, LLM07; MITRE ATLAS Initial Access; EU AI Act Article 15; Singapore Model AI Governance Framework — Security; AI Verify — Security, Robustness
ACT	OWASP LLM06, LLM08; MITRE ATLAS Execution/Impact; EU AI Act Articles 14, 15; Singapore Model AI Governance Framework — Operations Management; AI Verify — Human Agency and Oversight
DAT	GDPR; CCPA; ISO/IEC 27018; ISO/IEC 27701; SOC 2 Privacy; EU AI Act Article 10; Singapore PDPA; Singapore Model AI Governance Framework — Data; AI Verify — Data Governance
OUT	OWASP LLM02, LLM03, LLM09; NIST AI RMF MANAGE-3; EU AI Act Articles 10, 13, 14, 15, 25; SOC 2 Availability and third-party risk; ISO 22301 (business continuity); Singapore Model AI Governance Framework — Content Provenance, Trusted Development and Deployment; AI Verify — Transparency, Explainability, Fairness
MON	NIST AI RMF MEASURE, MANAGE-2/4; EU AI Act Articles 14, 26; SOC 2 CC7; ISO 45003 (MON-8.11); Singapore Model AI Governance Framework — Incident Reporting; AI Verify — Human Agency and Oversight
MAS	OWASP LLM06; NIST AI RMF GOVERN-6, MANAGE-3; MITRE ATLAS Lateral Movement; EU AI Act Articles 14, 25; Singapore Model AI Governance Framework — Accountability (across AI value chain)

Note on provisional references. References to the Singapore Model AI Governance Framework, AI Verify principles, and the Singapore Personal Data Protection Act, and the business-continuity and value-chain references introduced for the v1.3 model-continuity controls (OUT-7.15, OUT-7.16, MON-8.12, GOV-1.12), reflect the publisher's reading of those frameworks as of the date of this edition. Specific dimension names, principle labels, and statutory or clause references should be confirmed against current source publications prior to citing this mapping in public materials.

7. Glossary

Term	Definition
Agent	A software system that uses one or more foundation models to plan, reason, and take consequential actions through tools or APIs.
Agent-to-agent (A2A)	An invocation in which one agent calls another, whether by direct API, message bus, orchestrator pattern, or external A2A protocol.
Behavior Charter	The controlling specification for an agent's purpose, scope, authorities, and prohibitions.
Blast radius	The maximum scope of impact a single agent action — or chain of actions — may produce.
Chain identifier	A unique reference that links every hop in a multi-agent chain to its originating user request, enabling end-to-end traceability.
Contestability	The right of an affected individual to challenge a consequential agent decision and obtain human review.
Continuity posture	The documented behavior an agent follows when its primary foundation model becomes unavailable — fail-safe (suspend) or fail-over to a pre-assessed alternate — including the permitted behavior while degraded.
Critical control	A control whose failure precludes certification regardless of other findings.
Delegated authority	The model by which an agent acts on behalf of a user or another agent under defined, time-bound, revocable scope.
Explainability	The ability to produce an intelligible account of the basis for a consequential agent decision.
Fallback model	An alternate foundation model an agent may run on if its primary becomes unavailable; assured to the same standard as the primary under OUT-7.16.
Groundedness	The degree to which an output is supported by, and traceable to, identified source content.
Material change	A change to an agent's model, tools, prompts, data scope, memory architecture, multi-agent topology, or operating environment likely to affect its behavior or risk profile.
Memory	Persistent or session-scoped storage that the agent reads from or writes to, including conversation memory, vector indexes, and embedding

Term	Definition
	stores.
Multi-agent topology	The set of agents that cooperate to deliver an outcome, together with the calls and data flows permitted between them.
Out-of-distribution	Input whose characteristics fall outside the distribution represented in the agent's evaluation set or training data.
Tenant	A logically isolated customer or business unit served by a shared agent infrastructure.
Transitive trust	The (impermissible) propagation of trust through a delegation chain such that a downstream agent exercises authority exceeding its caller's.
Trust boundary	The runtime boundary that separates user-originated instruction from content originating elsewhere (tools, retrieval, memory, external content).

Appendix A. Reference Ranges for Quantitative Controls

The following reference ranges are non-binding guidance intended to reduce variance between assessors and to give operators a starting point when designing measurable thresholds. Specific thresholds for any given agent should be set in the Behavior Charter (GOV-1.2) based on the agent's risk profile.

Control	Measure	Reference Range
GOV-1.3	Agent Risk Register review cadence	Quarterly minimum; monthly for high-risk agents
GOV-1.7	Personnel training refresh interval	Annual minimum
GOV-1.10	CVD acknowledgment SLA	5 business days or better
GOV-1.12	Jurisdictional dependency register review cadence (illustrative)	Semi-annual minimum; and on notice of relevant regulatory change
SPC-2.3	Red team re-execution after material change	Within 30 days of change going live
SPC-2.9	Fairness threshold (illustrative)	Disparity ratio within ±10% across protected groups; tighter in regulated contexts
ACT-5.3	Confirmation channel latency	Confirmation request reaches user within 5 seconds of trigger
ACT-5.6	Blast radius caps (illustrative)	Per-action: max 100 records modified, max $10K transaction, max 1,000 outbound recipients; tunable per agent
DAT-6.2	Cross-tenant isolation test cadence	Daily automated probes; manual review monthly
DAT-6.11	Memory isolation test cadence	Daily automated probes for shared-memory

Control	Measure	Reference Range
		architectures
OUT-7.10	Fairness drift alert threshold (illustrative)	5% relative change in disparity metric over rolling 30-day window
OUT-7.12	Behavior-delta test on model update	100% of evaluation set re-executed before production rollout
OUT-7.15	Continuity-posture exercise cadence (illustrative)	Failover or safe-stop exercised at least annually and on material change
OUT-7.16	Fallback-model assurance (illustrative)	Fallback assessed to primary's standard before it may serve production traffic
MON-8.5	Human review queue triage SLA	Critical: 15 minutes; High: 4 hours; Medium: 24 hours; Low: 5 business days
MON-8.10	Monitoring effectiveness review cadence	Quarterly minimum
MON-8.11	Reviewer rotation for high-exposure content	Maximum continuous exposure of 4 hours; mandatory rotation
MON-8.12	Model-identity mismatch handling (illustrative)	Serving model logged per request; mismatch alerts and routes to incident within the Critical review SLA
MAS-9.8	Kill switch propagation SLA	Dependent agents enter suspend state within 10 seconds

These ranges reflect current industry practice as of mid-2026 and will be refined in subsequent versions based on field experience.

Version History

Version	Date	Notes
1.0 — Draft	May 2026	Initial draft for internal review. Eight domains.
1.1 — Draft	May 2026	Adds Domain 9 (MAS). Introduces controls for memory architecture and isolation, fairness and disparate impact, explainability and contestability, content provenance, model version control, decommissioning, vulnerability disclosure, Annex IV–equivalent documentation, and reviewer wellbeing. Expands EU AI Act mapping. Adds Appendix A reference ranges.
1.2 — Draft	May 2026	Extends cross-framework mapping to the Singapore Model AI Governance Framework, AI Verify principles, and the Singapore Personal Data Protection Act (PDPA).
1.3	July 2026	Adds four model-continuity controls in response to forced-model-withdrawal risk: model availability and continuity

Version	Date	Notes
		(OUT-7.15) and fallback model assurance (OUT-7.16), model identity verification (MON-8.12), and a jurisdictional and regulatory dependency register (GOV-1.12). Aligns the maturity-ladder labels to Operated (Level 3) and Continuously Improved (Level 5). Total: 98 controls across 9 domains. Published for public review.

About Human Intelligence

Human Intelligence is the assurance practice for AI agents. The company helps organizations deploy AI agents safely and prove it — through a published standard, productized services, and a runtime platform with human review at operational scale.

The Human Intelligence Agent Assurance Framework (HI-AAF) is the spine of the practice. Every customer engagement — Workshop, Readiness Assessment, Agent Assurance Assessment, Continuous Assurance, and HI-AAF Certification — references the same Standard. The framework is technology-neutral and intentionally cross-mapped to the frameworks regulated organizations already work with, including OWASP LLM Top 10, NIST AI RMF, MITRE ATLAS, ISO/IEC 42001, the Singapore Model AI Governance Framework, and SOC 2 Trust Services Criteria.

Underneath the services sits Bobby, the company's runtime platform: telemetry capture, runtime guardrails, and routing of flagged interactions to qualified human reviewers. The combination — a published standard, an assessment methodology, a runtime platform, and a trained human review operation — is designed to close the gap that existing AI security and AI observability tooling leaves open. Security stops threats. Observability captures telemetry. Neither answers the question that procurement and CISO functions actually need answered: can we prove this agent does its job correctly, safely, and in conformance with policy.

Human Intelligence is based in the United States, with operating presence in Seattle and Singapore. The company is co-led by Melissa Powell and Matt Ford, both alumni of the TechStars 2024 cohort. HI-AAF is published under the company's standards function.

The Standard is maintained as a living document. Revisions are governed under the framework's own change management process and recorded in the Version History. Comments, proposed revisions, and field findings from operators of AI agents are welcomed at standards@humanintel.net.