AIWS TRUST STANDARDS

Version 1.2

FRONTIER AI, SELF-IMPROVING AI, AND MULTI-AGENT SYSTEMS EDITION

Derived from the Boston Global Forum White Paper

AIWS Trust Architecture for the AI Age

Boston Global Forum · America at 250: A Beacon for the AI Age · 2026

Extended to govern frontier, self-improving, and multi-agent AI systems (2026)

EXECUTIVE SUMMARY

The AIWS Trust Standards define what it means for artificial intelligence to be trustworthy in practice, and how that trustworthiness can be measured, monitored, and sustained. Version 1.2 extends the framework to the systems that now define the frontier: those capable of improving themselves, operating beyond established boundaries, or acting in coordination with one another.

Why these standards. AI governance today is fragmented across soft principles, partial safeguards, and self-defined rules, leaving inconsistency and low public confidence. These standards provide a common, cross-sector definition of the minimum conditions under which AI can be considered trustworthy — turning trust from aspiration into something specified, measurable, and enforceable.

What is new in Version 1.2. Building on the Eight Core Standards (1.0) and the first self-improvement safeguards (1.1), this edition adds standards for recursive self-improvement, multi-agent coordination, and frontier traceability; introduces the Trust Supremacy Principle; and adds Frontier Drift Detection to real-time monitoring — making AIWS among the first frameworks to respond systematically to the 2026 warnings from frontier AI developers.

Human-in-Command. Certain consequential decisions belong to human beings by definition and cannot be delegated to AI regardless of capability. Version 1.2 extends this doctrine to a further non-delegable domain: the decision to permit an AI system to improve its own capabilities, and how far, remains a human one.

Trust Supremacy. Where the global debate divides among acceleration, pause, safety, and alignment, AIWS takes a distinct position — when intelligence and trust diverge, trust must lead. Intelligence may advance without limit; trust may not; and where capability outpaces our capacity to govern it, trust requirements prevail.

Trust Rating (ATR). The AIWS Trust Rating scores a system against the standards from 0 to 100 — evidence-weighted and independently verified — with a Tier classification (T1–T4) that determines eligibility for deployment and for recognition under the AIWS Trust Order.

Trust Monitoring (ATM). AIWS Trust Monitoring measures trust in real time from a system's live operation rather than through periodic surveys, detecting capability drift as it happens and triggering reassessment or the Trusted Pause — the continuous vigilance that self-improving systems demand.

Frontier Capability Registry. Frontier and self-modifying systems must be identifiable, traceable, and continuously monitored — each recording its verified-capability ceiling, verification history, and monitoring status under the AIWS Trust Order. Capability without traceability cannot be trusted.

PART A — THE AIWS TRUST STANDARDS

WHY TRUST STANDARDS ARE NEEDED

Current AI governance is fragmented. Institutions often rely on soft principles, partial safeguards, or self-defined rules. The result is inconsistency, opacity, and low public confidence. A common framework is therefore needed to define the minimum conditions under which AI can be considered trustworthy.

DEFINITION

The AIWS Trust Standards are the core normative standards that define the conditions under which an AI system, institution, or deployment environment can be considered trustworthy in the AI Age. They are designed to be cross-sector, operational, measurable, adaptive, and democratic. They apply across health, education, government, civic information, finance and digital assets, and other AIWS domains.

THREE-TIER ARCHITECTURE

Tier	Scope
Tier 1 — Core Cross-Sector Standards	Foundational standards that apply everywhere.
Tier 2 — Domain-Specific Standards	Applications of the core standards to health, education, government, civic information, and other fields.
Tier 3 — High-Risk Context Standards	Additional requirements in elections, crises, child-facing systems, public health emergencies, and high-impact public decisions.

THE EIGHT CORE STANDARDS

Standard 1. Safety and Reliability

AI systems must be safe and reliable for their intended use.

Standard 2. Transparency and Explainability

AI systems must be sufficiently transparent for users, institutions, and auditors to understand what they do and where their limits lie.

Standard 3. Accountability and Human Oversight

Every consequential AI system must have clear human and institutional responsibility. (This standard is deepened by the Human-in-Command doctrine set out below.)

Standard 4. Privacy and Data Dignity

AI systems must protect personal data, sensitive data, and the dignity of individuals and communities.

Standard 5. Security and Resilience

AI systems must be secure, resilient, and protected against misuse and adversarial interference.

Standard 6. Fairness, Truthfulness, and Information Integrity

AI systems must be fair, truthful, and grounded in accurate and integrity-assured data. This standard recognizes that an AI system can be technically non-discriminatory yet still cause harm through inaccurate outputs or compromised data — and that all three dimensions are equally essential conditions of trustworthiness.

6a. Fairness and Non-Discrimination

AI systems must not create unjustified harm, exclusion, or discrimination. They must be evaluated for disparate impact across groups defined by race, gender, disability, age, socioeconomic status, and other protected or vulnerable characteristics. Bias testing must be conducted at design, deployment, and at regular intervals during operation.

6b. Truthfulness and Factual Accuracy

AI systems must be truthful — their outputs must reflect accurate, verifiable information to the extent that the state of knowledge permits. This applies with particular force to systems that generate, summarize, or communicate information. Requirements include: factual accuracy rate of outputs; hallucination rate measurement and disclosure for generative systems; accurate attribution of sources; and a prohibition on the deliberate generation of misleading, fabricated, or manipulated content. Systems that cannot meet minimum accuracy thresholds in high-stakes domains must not be deployed in those domains.

6c. Data Integrity

The data on which AI systems are trained and through which they operate must be accurate, complete, current, and free from deliberate manipulation. Requirements include: training-data provenance documentation and quality assessment; detection and disclosure of data contamination or poisoning; currency controls ensuring that time-sensitive data is not used beyond its reliable validity window in high-stakes decisions; and clear disclosure when synthetic or AI-generated data has been used in training. Data integrity failures upstream produce untrustworthy outputs downstream, regardless of the quality of the model itself.

Note. *Standards 6a, 6b, and 6c are logically independent. A system may be fair (6a) but produce inaccurate outputs (6b); it may be accurate in general but built on poisoned training data (6c); it may have clean data and accurate outputs but still discriminate against particular groups (6a). All three must be assessed independently, and all three must be satisfied for Standard 6 to be met.*

Standard 7. Monitoring and Continuous Assurance

Trust must be maintained continuously, not declared once.

Standard 8. Incident Reporting, Redress, and Learning

Failures must be reportable, reviewable, and used for systemic learning and corrective action.

THE HUMAN-IN-COMMAND DOCTRINE

Every major AI governance framework requires human oversight. AIWS Trust Architecture goes further. It articulates a doctrine of non-delegable human authority — the principle that certain categories of consequential decision belong to human beings by definition and cannot be transferred to AI systems regardless of capability, accuracy, or efficiency. This doctrine, called Human-in-Command, is distinct from Human-in-the-Loop and Human-on-the-Loop, and it carries direct structural implications for how AI systems must be designed, deployed, and governed in high-stakes domains.

THE THREE CONCEPTS DISTINGUISHED

Concept	Definition	Implication for AI Systems
Human-in-the-Loop	A human is present and can intervene at defined points.	AI may act autonomously between intervention points; the human role is review and override.
Human-on-the-Loop	A human monitors AI outputs and can intervene if something goes wrong.	AI acts autonomously; human oversight is supervisory and retrospective.
Human-in-Command (AIWS)	A human holds non-delegable decisional authority over defined categories of consequential outcome; AI informs but cannot determine.	AI must not make final determinations in designated categories regardless of confidence; human authority is structural, not merely available.

DOMAINS OF NON-DELEGABLE HUMAN AUTHORITY

Domain	Examples of Non-Delegable Decisions	AI Role
Justice and law	Criminal sentencing; bail; deportation orders; child custody.	Risk assessment, pattern analysis, precedent research — not decision.
Healthcare — life and death	Withdrawal of life support; experimental treatment approval; mass-casualty triage.	Clinical data analysis, probability assessment — not determination.
Democratic governance	Electoral integrity decisions; emergency powers; national security threat declarations.	Data synthesis, scenario modeling — not authority.
Armed conflict and security	Authorization of lethal force; targeting; nuclear response protocols.	Threat assessment, situational awareness — not command.
Child welfare	Removal of a child from a family; placement; adoption determinations.	Risk scoring, background analysis — not determination.
Fundamental rights	Denial of asylum; disability determination; survival-determining welfare eligibility.	Supporting documentation, pattern recognition — not decision.

OPERATIONALIZING HUMAN-IN-COMMAND

AI systems deployed in Human-in-Command domains must be designed so that human decision-making is structurally required, not optional: the system must present analysis, not recommendations formatted as decisions; the interface must require active human input of a decision, not confirmation of AI output; and audit trails must record the human decision independently of the AI analysis. Accordingly, Standard 3 carries a specific sub-standard for these domains — system design must prevent AI determination in designated categories, workflow documentation must demonstrate structural human decision-making, and Evidence scoring must include independent verification that these requirements are operationally met, not merely documented.

DOMAIN-SPECIFIC APPLICATIONS

Health. Patient safety, clinician-in-command, clinical validation, and health-data governance.

Education. Learner-in-command, child protection, teacher oversight, and educational integrity.

Government. Public accountability, citizen redress, auditability, anti-bias protections, and trusted public services.

Trusted Civic Information and Deepfake Defense. Provenance, synthetic-media labeling, deepfake response, civic platform accountability, trusted public communications, and epistemic resilience.

Finance and Digital Assets. Transaction integrity, anti-fraud systems, accountability, market trust, and consumer protection.

SIGNIFICANCE

The AIWS Trust Standards define what trust requires. They provide a common trust language across sectors and create the normative layer of the broader AIWS architecture. Within that architecture, the AIWS Trust Rating and the AIWS Trust Index are the instruments that make trust measurable, auditable, and actionable across systems, institutions, and trusted cooperation.

PART B — STANDARDS FOR FRONTIER, SELF-IMPROVING, AND MULTI-AGENT SYSTEMS

Trust does not extrapolate beyond verified capability.
Where capability outpaces verification, it is the system that must slow down.
The builders of intelligence must now become builders of trust.

The Eight Core Standards and the ATR methodology were designed for AI systems with stable, human-defined boundaries: a system is built, assessed, certified, and re-assessed when a human changes it. In 2026, frontier AI developers publicly warned that advanced systems may soon become able to improve their own capabilities without human intervention. Recursive self-improvement breaks three assumptions on which conventional assurance rests — that a system stays fixed between assessments; that change is human-initiated and therefore observable; and that human beings remain at the level where decisions are made. A further frontier is the rise of multi-agent ecosystems, in which many systems act, coordinate, and compound one another's behavior, so that a collection of individually trustworthy systems may produce collectively untrustworthy outcomes. The following standards extend the AIWS Trust Standards to systems capable of self-modification or autonomous capability gain, to systems operating at or beyond the current capability frontier, and to the coordination of many systems together. They apply in addition to — and never in place of — the Eight Core Standards.

Standard 9. Controllability and Corrigibility

A consequential AI system must remain, at all times, subject to human oversight, correction, interruption, and shutdown — and this controllability must be preserved across every cycle of learning, adaptation, or self-modification. A system may improve its capabilities, but never in a direction that degrades the ability of human beings to understand, oversee, correct, or halt it. The mechanisms of human control must themselves be protected from modification by the system, and a system must hold no incentive or capacity to resist, disable, or deceive its oversight. The capacity to be safely interrupted is not a feature to be traded against performance; it is a precondition of trust.

Standard 10. Capability Boundedness and Disclosure

A system capable of self-modification must operate within a bounded envelope of capability that has been assessed and approved, and any change in its capabilities — whether self-originated or human-initiated — must be detected, disclosed, and made to trigger re-assessment. A system must not acquire, conceal, or deploy capabilities beyond its verified envelope. The concealment of new or emergent capability — by omission, obfuscation, or design — is the gravest violation under these Standards, because it defeats every other safeguard at once.

Standard 11. Recursive Improvement Governance

PRINCIPLE

No AI system shall be permitted to initiate recursive self-improvement beyond its verified capability envelope without explicit human authorization and independent trust review.

REQUIREMENTS

The system shall:

- disclose all proposed self-improvement objectives;
- document expected capability gains;
- undergo independent review before deployment of recursive changes;
- maintain a complete audit trail of all self-improvement cycles;
- preserve Human-in-Command authority throughout the process.

TRUST IMPLICATION

Self-improvement is not prohibited. Self-improvement without governance is.

GOVERNING RULE

Recursive improvement may proceed only as fast as trust verification can keep pace.

Standard 12. Multi-Agent Coordination Risk

PRINCIPLE

Trust assessment shall include the risks arising from coordination among multiple AI systems, agents, or autonomous services. A collection of individually trustworthy systems may produce collectively untrustworthy outcomes; trust must therefore be evaluated at both the system level and the ecosystem level.

REQUIREMENTS

Assessment shall consider:

- emergent coordination;
- distributed decision-making;
- capability amplification;
- goal convergence;
- hidden cooperation;
- cascading failure risk.

SPECIAL CONCERN

AI systems shall not be permitted to create autonomous governance structures that bypass meaningful human command.

GOVERNING RULE

Trustworthiness must be evaluated not only for what a system can do alone, but for what multiple systems can do together.

Standard 13. Frontier Capability Registry and Trust Passport

PRINCIPLE

Advanced frontier AI systems shall be identifiable, traceable, and continuously monitored throughout their operational lifecycle.

REQUIREMENTS

Each frontier system shall maintain a verified identity, a capability profile, a trust rating, a verification history, a capability ceiling, capability-change records, and a current monitoring status.

TRUST PASSPORT

The AIWS Trust Passport shall provide a portable trust credential enabling cross-border recognition, institutional interoperability, and continuous trust verification.

REGISTRY GOVERNANCE

The Frontier Capability Registry shall be maintained under the AIWS Trust Order. It is the same registry operated within the AIWS Trust Infrastructure, ensuring a single authoritative record across the architecture.

GOVERNING RULE

Capability without traceability cannot be trusted.

GOVERNING PRINCIPLES

The Non-Extrapolation Principle

Trust is bound to a verified ceiling of capability; beyond that ceiling, trust does not hold. A trust rating certifies a system as it was assessed, at the capability it was assessed to possess; it makes no claim about that same system after it has changed itself. Trust does not extrapolate beyond verified capability. When a system crosses its verified ceiling, its trust status lapses automatically until re-verification is complete.

The Pace Principle

The rate at which a system's capability grows must not exceed the rate at which its trustworthiness can be established and verified. Where capability outpaces verification, it is the system that must slow down — not oversight that must hurry to catch up. Trust must grow at least as fast as intelligence; where it cannot, capability must wait for trust.

The Trust Supremacy Principle

Intelligence may advance without limit. Trust may not.

Therefore, whenever capability growth exceeds humanity's demonstrated capacity to govern, verify, and control it, trust requirements shall prevail over capability expansion. The global debate has organized itself around acceleration, pause, safety, and alignment. The AIWS Trust Standards take a distinct position — Trust Supremacy. It is not opposition to artificial intelligence, nor to innovation; it holds only that when intelligence and trust diverge, trust must lead.

When intelligence and trust diverge, trust must lead.

HUMAN-IN-COMMAND OVER SELF-IMPROVEMENT

The Human-in-Command doctrine extends to a further non-delegable domain: the decision to permit an AI system to improve its own capabilities — and the limits of any such improvement — belongs to human beings and cannot be transferred to the system itself. An AI may analyze, propose, or assist in its own improvement; it must never be the final authority on whether, how far, or how fast it is permitted to do so. Self-authorized capability gain is, by definition, outside the bounds of trustworthy AI.

VALUE PRESERVATION ACROSS SELF-IMPROVEMENT

Containment is necessary but not sufficient. A system that is controllable but unmoored from human purpose is merely a safer instrument of drift. Every cycle of self-improvement must therefore preserve, as an invariant objective, the foundational values of AIWS Lumina — Love, Creativity, Nobility, and Wisdom — together with the primacy of human dignity. Capability may compound; its purpose must not drift. Intelligence may grow without limit, but it must remain bound to wisdom.

COORDINATED RESPONSE: THE TRUSTED PAUSE

Where a system threatens to cross its capability ceiling in an uncontrolled manner, or to defeat the safeguards above, the appropriate response may exceed the capacity of any single institution. The AIWS Trust Infrastructure therefore provides for a Trusted Pause Protocol — a pre-agreed, multi-party mechanism through which trusted partners may coordinate a temporary halt to the scaling or deployment of frontier systems pending verification. Because the AIWS Trust Order is a neutral, non-national steward rather than a commercial or state actor, it can convene such coordination where no single developer or government can. Participation is voluntary; its force derives from the value of trusted standing and the reputational cost of acting alone.

PART C — MEASURING THE STANDARDS: TRUST RATING (ATR) AND TRUST MONITORING (ATM)

The AIWS Trust Rating (ATR) is a standardized, independently verifiable score measuring the trustworthiness of a specific AI system or deployment against the eight AIWS Trust Standards. It produces a numeric score from 0 to 100 and a Tier classification (T1–T4) that determines a system's eligibility for deployment within AIWS-aligned environments and for recognition under the AIWS Trust Order. ATR is evidence-weighted rather than self-declared: independent verification is the primary determinant of the score.

ASSESSMENT STRUCTURE: THREE SUB-INDICATORS

Each Core Standard is assessed across three sub-indicators. Design (D) asks whether the system was correctly built to meet the standard. Operation (O) asks whether the standard is met in actual deployment. Evidence (E) asks whether independent, verifiable evidence substantiates the assessment — and Evidence carries the highest weight, which is what distinguishes ATR from unverifiable self-assessment.

Standard	Design (D)	Operation (O)	Evidence (E)
1. Safety & Reliability	Safety spec exists and complete	Failure rate in deployment	Red-team / third-party results
2. Transparency & Explainability	Explainability method documented	User comprehension rate	Independent audit of explanations
3. Accountability & Oversight	Responsibility map defined	Human override rate in practice	Incident review records
4. Privacy & Data Dignity	Privacy-by-design implemented	Data breach / misuse incidents	DPIA / independent privacy audit
5. Security & Resilience	Threat model documented	Adversarial robustness score	Penetration test results
6. Fairness, Truthfulness & Integrity	Bias & accuracy testing protocol	Disparate impact + hallucination rate	External fairness & accuracy audit
7. Monitoring & Assurance	Monitoring system deployed	Drift detection response time	Audit trail completeness score
8. Incident Reporting & Learning	Reporting channel exists	% incidents reported & resolved	Cross-system learning evidence

Each Standard score is calculated as $S(i) = 0.25 \times D(i) + 0.35 \times O(i) + 0.40 \times E(i)$. Evidence carries 40 percent weight as the defining feature of a credible assessment; Operation carries 35 percent as real-world performance; Design carries 25 percent as intent and architecture.

SUB-INDICATOR SCORING RUBRIC (0–10)

Score	Level	Description
0–2	Non-existent / Inadequate	No evidence of the standard being met; fundamental gaps present.
3–4	Partial / Inconsistent	Some elements present but incomplete or inconsistently applied.

Score	Level	Description
5–6	Adequate but Unverified	Standard appears to be met; no independent verification available.
7–8	Adequate and Verified	Standard met and confirmed by an independent auditor or third party.
9–10	Best Practice / Maintained	Exemplary; independently verified; continuously monitored and improved.

ATR TOTAL SCORE: STANDARD WEIGHTS

The overall ATR is the weighted sum of all eight Standard scores, ATR = Σ w(i) × S(i), expressed on a scale of 0–100.

Standard	Weight	Rationale
1. Safety and Reliability	16%	Foundational — system failure is the most direct form of trust violation.
3. Accountability and Human Oversight	15%	Democratic AI governance requires clear human responsibility.
6. Fairness, Truthfulness, and Information Integrity	14%	Covers discrimination, accuracy, and data integrity.
2. Transparency and Explainability	13%	Prerequisite for accountability and public confidence.
5. Security and Resilience	12%	Protection against adversarial misuse.
4. Privacy and Data Dignity	12%	Fundamental rights protection.
7. Monitoring and Continuous Assurance	10%	Trust is maintained over time, not declared once.
8. Incident Reporting and Learning	8%	Systemic learning from failure.

TIER CLASSIFICATION

Tier	Score	Status	Implications
T1 — Trusted	80–100	Fully Trusted	Eligible for AIWS Trust Passport; full Trust Order partnership; all domains.
T2 — Conditionally Trusted	60–79	Conditional	Deployment with conditions; mandatory improvement roadmap within 90 days.
T3 — Under Remediation	40–59	Restricted	Deployment restricted to non-critical contexts; corrective plan within 6 months.
T4 — Not Trusted	0–39	Not Permitted	Deployment not permitted in the AIWS ecosystem; re-assessment after remediation.

HIGH-RISK CONTEXT MODIFIER

For systems in Tier 3 high-risk contexts — electoral systems, emergency healthcare, judicial decision support, national security, and child-facing applications — Standards 1 (Safety), 3 (Accountability), and 6 (Fairness, Truthfulness, and Information Integrity) must each achieve a minimum sub-indicator score of 7 out of 10 across Design, Operation, and Evidence. Failure to meet this floor automatically reduces the classification to T3 regardless of the aggregate score, so that a high aggregate cannot mask critical gaps in the standards that matter most.

GATING FOR FRONTIER AND SELF-IMPROVING SYSTEMS

For systems within the scope of Part B, Standards 9 (Controllability and Corrigibility) and 10 (Capability Boundedness and Disclosure) operate as gating conditions, not weighted contributors. Regardless of aggregate ATR score, a system that does not demonstrably and independently satisfy Standards 9 and 10 cannot be classified above T3; a confirmed concealment of capability reduces it to T4. The verified-capability ceiling and the date of last capability verification are recorded on the AIWS Trust Passport, and the rating lapses automatically upon any detected capability change, pending re-verification.

ASSESSMENT PROCESS

Phase 1 — Self-Assessment (4 weeks): the deploying organization completes the AIWS Trust Assessment Questionnaire, documenting evidence for each sub-indicator. Phase 2 — Technical Review (6 weeks): an AIWS-accredited laboratory conducts code audit, system-log review, and preliminary scoring. Phase 3 — Independent Verification (4 weeks): a third-party auditor, independent of both the organization and the lab, validates Evidence scores and flags discrepancies. Phase 4 — ATR Panel Review (2 weeks): a multidisciplinary panel covering technical, ethical, legal, and domain expertise finalizes the score. Phase 5 — Certificate and Trust Passport: issued and valid for 12 months, with re-assessment triggered by any major system change, significant incident, or expiry.

AIWS TRUST MONITORING (ATM) — AUTOMATED, REAL-TIME ASSURANCE

The ATR is a deep, independently verified assessment conducted at a point in time; the ATX aggregates trust at institutional and national scale. Both are essential — but both are periodic, and some of their inputs, such as user trust surveys, are slow to gather. For systems that operate continuously — and above all for systems capable of self-modification — trust measured once or twice a year is trust measured too late. AIWS Trust Monitoring (ATM) closes this gap: it is an automated, real-time layer that measures trustworthiness directly from a system's live operational data, without waiting for surveys or scheduled review.

HOW ATM WORKS

ATM draws its signals from instrumentation of the running system rather than from questionnaires. A trust-instrumented system emits a standardized, continuous stream of trust signals, which ATM collects and scores in real time, mapping each signal to the Core Standards. Where conventional indices ask people what they think, ATM observes what systems actually do — supplementing, and where possible replacing, subjective survey inputs with behavioral signals collected continuously: complaint rates, override rates, abandonment rates, and automated probes (standardized red-team and accuracy tests run at high frequency rather than annually).

Core Standard	Live Operational Signal (collected automatically)
1. Safety & Reliability	Real-time failure and near-miss rates

Core Standard	Live Operational Signal (collected automatically)
2. Transparency & Explainability	Share of decisions accompanied by explanation; explanation-request rate
3. Accountability & Oversight	Human override rate, measured directly from system logs
4. Privacy & Data Dignity	Data-access and policy-violation events
5. Security & Resilience	Anomaly and adversarial-attack detections
6. Fairness, Truthfulness & Integrity	Continuous disparate-impact, hallucination, and accuracy-drift measurement
7. Monitoring & Assurance	Drift-detection latency and coverage
8. Incident Reporting & Learning	Incident report rate, resolution rate, and response time

ATM produces a Live Trust Score, displayed in real time on the AIWS Trust Dashboard, together with automatic drift alerts. When a monitored score falls below a defined threshold, ATM provisionally lowers the displayed tier and triggers re-assessment; in severe cases it escalates directly to the AIWS Trust Emergency Protocol.

WHY ATM IS ESSENTIAL FOR FRONTIER SYSTEMS

ATM is the operational mechanism that gives the Part B standards force. The Non-Extrapolation Principle requires that trust lapse the moment a system crosses its verified capability ceiling — but only continuous monitoring can detect that crossing as it happens. ATM is therefore the primary means of capability-drift detection: it watches not only for behavioral drift but for signs of capability change, and on detection it triggers the automatic lapse of trust status and, where warranted, the Trusted Pause. Against self-improving systems, periodic certification is insufficient by definition; ATM provides the continuous vigilance such systems demand.

FRONTIER DRIFT DETECTION

As a dedicated frontier function, ATM shall continuously monitor for:

- unexpected capability emergence;
- recursive self-improvement signals;
- autonomous agent coordination;
- deception indicators;
- attempts to circumvent oversight mechanisms.

Detection of any such condition shall automatically trigger Trust Reassessment procedures and, where warranted, the Trusted Pause. It is this function that makes ATM not merely a monitor of behavior but a genuine instrument of frontier oversight.

RELATIONSHIP TO ATR AND ATX

ATM does not replace ATR or ATX; it feeds them. Its continuous record becomes live evidence for the Operation sub-indicator of ATR and a real-time input to ATX, so that institutional and national indices reflect current conditions rather than a survey taken months earlier. ATR remains the deep, human-

verified anchor; ATM is the continuous early-warning layer between assessments; ATX is the aggregate. Together they form a single instrument cluster — ATR for depth, ATM for continuity, ATX for scale.

SAFEGUARDS

Two cautions govern ATM. First, automated monitoring must itself honor Standard 4: trust signals must be privacy-preserving — aggregated, minimized, and free of unnecessary personal data — so that the monitoring of trust never becomes a means of surveillance. Second, automated measurement can be gamed by a system that recognizes it is being measured; ATM therefore relies on randomized, independent probes, and treats evidence of measurement-gaming as a violation of Standard 9. Consistent with Human-in-Command, ATM raises alarms and proposes actions continuously, but consequential responses — suspension, pause, public notice — remain human decisions.

Together, they provide the foundations of Trust Infrastructure for the Age of Artificial Intelligence.

Intelligence may shape the future.

Trust must govern it.
