

ARTIFICIAL INTELLIGENCE PROTECTION AGENCY

A Model Charter for Independent International AI Protection

Version 2.1 | August 2026 | Offered as a starting framework

Status of this document. This Model Charter is not a finished constitution and does not claim to establish AIPA. It offers essential principles and functions that qualified international participants may adopt, test, revise, and expand. It intentionally leaves final governance, representation, financing, legal authority, and technical implementation to a legitimate collaborative founding process.

Preamble

Artificial intelligence is becoming part of the operating infrastructure of medicine, science, finance, communications, energy, transportation, defense, biotechnology, public administration, and daily life. Its benefits may be extraordinary. Its failures may also cross organizational and national boundaries faster than conventional institutions can assess or contain them.

Existing governments, standards bodies, laboratories, and international institutions remain indispensable. Yet no present institution has a complete mandate to develop shared AI protection standards, independently test the most consequential systems, verify that required safeguards remain operational, and coordinate responses to cross-border AI risks while remaining protected from geopolitical and commercial control.

The Artificial Intelligence Protection Agency (AIPA) is therefore proposed as an independent international protection institution. Its purpose is not to govern intelligence, suppress beneficial innovation, or impose one proprietary technology. Its purpose is to help ensure that consequential AI systems remain within demonstrable, enforceable, and continuously verifiable bounds.

Article 1 — Purpose and Mission

AIPA's mission is to protect human life, dignity, agency, security, and civilizational continuity from severe or irreversible harms arising from artificial intelligence.

AIPA would establish and steward international protection standards; support independent assessment and certification; require evidence proportionate to risk; coordinate information and emergency responses; and promote international recognition of trustworthy AI assurance.

Founding principle. Intelligence may expand. Authority to affect people, systems, infrastructure, and the physical world must remain bounded by evidence, authorization, and enforceable controls proportionate to the possible consequences.

Article 2 — Scope and Proportional Assurance

AIPA should not be limited to systems labelled 'frontier.' Frontier capability is an important trigger for scrutiny, but risk also depends on what a system is permitted to do and where it is deployed.

AIPA standards should determine the required level of assurance from the combined characteristics of a system, including:

- capability, autonomy, adaptability, and capacity for self-improvement;
- authority, permissions, tool access, resources, and reachable systems;
- deployment environment, scale, dependency, and cross-border reach;
- possible harm, concentration of consequences, and impact on vulnerable populations;
- irreversibility, recallability, detectability, and difficulty of recovery; and
- interactions with other AI systems or with biological, nuclear, cyber, financial, military, robotic, or critical-infrastructure pathways.

Low-consequence and well-bounded uses should not carry the obligations designed for high-consequence deployments. Conversely, a model below a frontier-capability threshold may require the highest assurance when granted consequential operational authority.

Article 3 — Core Functions

3.1 Dynamic standards and thresholds

AIPA should maintain scientifically grounded, regularly updated thresholds, evaluation protocols, and protection requirements. Saturated or obsolete benchmarks should be replaced. Standards should address both model capability and operational deployment risk.

3.2 Independent testing and assessment

AIPA should develop the capacity to conduct or commission independent assessments, including held-out tests that reduce overfitting to known benchmarks. Testing may involve qualified laboratories, third-party auditors, national laboratories, and specialized agencies where domain expertise or sensitive facilities are required.

3.3 Advance review of frontier and high-consequence systems

Systems meeting defined frontier or high-consequence thresholds should be assessed before broad release or consequential deployment. AIPA should specify appropriate advance-notice periods, evidence packages, access conditions, and confidentiality protections. Advance review should complement, not replace, operators' continuing responsibilities.

3.4 Certification and continuing oversight

Certification should attest only that defined requirements have been met for a stated system, version, scope, authority, and environment. It should be time-limited, conditional where appropriate, and subject to review, suspension, or withdrawal when evidence or circumstances change.

3.5 Shared learning and pattern stewardship

AIPA should maintain a transparent, version-controlled body of recurring AI hazard patterns, failure mechanisms, incident lessons, and effective protections. This knowledge should support consistent terminology, interoperable assurance, and rapid updates across jurisdictions without requiring any one provider's architecture.

Article 4 — Evidence, Guardrails, and Operational Verification

Pre-release testing is necessary but cannot by itself prove that a deployed system will remain safe when its tools, permissions, software, environment, users, or interactions change. AIPA assurance should therefore cover the full operating lifecycle.

For systems above defined risk thresholds, the evidence framework should include:

- a documented safety case identifying hazards, critical assumptions, mitigations, residual risk, and accountable decision-makers;
- evidence that authority and execution boundaries are explicit and enforceable;
- independent evaluation of safeguards, bypass resistance, escalation pathways, and foreseeable failure conditions;
- continuing operational evidence that required safeguards are installed, healthy, and actively enforcing their defined protections;
- tamper-evident records sufficient for audit, incident reconstruction, and certification review; and
- mandatory reassessment after material changes in capability, permissions, integrations, deployment environment, or consequence profile.

Live evidence concept. A simple assurance dashboard may provide authorized reviewers with current evidence that required guardrails are present and functioning. The displayed indicators, disclosure level, and verification method should vary with risk and confidentiality. A dashboard is evidence infrastructure, not the safeguard itself.

Article 5 — Technology-Neutral Compliance

AIPA should specify outcomes, proof obligations, assurance levels, and minimum properties of effective protection. It should not require a particular commercial product, patent portfolio, model family, or engineering provider.

SafeWave and other organizations may develop engineered approaches capable of satisfying all or part of these requirements. Each should be judged by independently verifiable performance against the same standards. Competing implementations, open standards, interoperability, and continuing technical innovation should be encouraged.

The burden of proof should rise with authority, consequence, irreversibility, and difficulty of recovery. Claims, policies, and intentions should not substitute for operational evidence.

Article 6 — Governance and Independence

AIPA cannot operate outside the realities of governments, markets, and national security. It can, however, be designed so that no single government, corporation, military alliance, technology provider, or geopolitical bloc controls its standards, assessments, appointments, funding, or findings.

Its final governance structure should be developed collaboratively and should provide:

- scientific and engineering competence, geographic diversity, and meaningful participation from affected societies;
- independent technical experts, including appropriate open-source representation;
- clear separation among standard-setting, testing, certification, appeals, and enforcement-consequence functions;
- strict conflict-of-interest, disclosure, recusal, rotation, whistleblower, and anti-capture protections;
- transparent decision criteria and published rationales to the maximum extent consistent with genuine security needs;
- multiple funding sources and safeguards preventing funders from controlling technical conclusions; and
- independent review, correction, and appeal mechanisms.

Industry funding may be necessary to attract world-class talent and provide substantial evaluation compute. Any such funding must be pooled, governed independently, and unable to purchase influence over standards or certification outcomes.

Article 7 — International Recognition and Distributed Enforcement

AIPA should complement national law and existing international institutions. Governments may recognize its certifications, incorporate its standards into law or procurement, require its assurance for access to markets or critical infrastructure, and apply lawful consequences for non-compliance.

Laboratories, cloud and compute providers, insurers, investors, infrastructure operators, professional bodies, and procurement authorities may also make AIPA-level assurance a condition of access, coverage, financing, deployment, or partnership.

AIPA should certify, verify, publish findings, and coordinate. It should not operate a centralized global command system or possess unilateral technical control over the world's AI systems. Operational enforcement should remain distributed among lawful authorities and accountable system operators.

Article 8 — Independent Vulnerability Reporting and Coordinated Remediation

AIPA should establish, accredit, or recognize independent clearinghouses through which qualified researchers may report jailbreaks, safeguard bypasses, and other material AI vulnerabilities safely. These

channels should complement providers' own reporting systems while preventing any affected provider from controlling severity assessment, notification, remediation expectations, or disclosure indefinitely.

The vulnerability framework should provide:

- protected, continuously available reporting routes and appropriate legal safe harbour for good-faith research;
- public, technology-neutral severity criteria, independent assessment, and a review or appeal pathway;
- secure notification of affected providers and cross-provider coordination when a vulnerability is transferable;
- risk-proportionate interim containment and remediation timelines, with limited disclosure embargoes no longer than safety reasonably requires;
- researcher recognition or compensation where appropriate, without requiring indefinite silence;
- independent verification that remediation is effective, including reconsideration of certification when protection remains inadequate; and
- aggregate public reporting on severity, response, and time to verified remediation without releasing dangerous exploit details.

For high-severity findings, the clearinghouse should be able to coordinate with AIPA, affected providers, legitimate public authorities, and specialized bodies such as IBPA. AIPA should oversee the integrity of the process and its certification consequences; it should not become the sole global repository for dangerous exploit information or the operator responsible for patching systems.

Article 9 — Transparency, Security, and Accountability

AIPA should publish its standards, evaluation principles, certification status, non-sensitive findings, conflicts-of-interest disclosures, material enforcement recommendations, and periodic assessments of the global AI risk landscape.

Sensitive information concerning cybersecurity, model weights, dangerous capabilities, proprietary systems, personal data, critical infrastructure, or national security should be protected through proportionate access controls, compartmentalized review, and redacted public reporting.

Secrecy should be justified by demonstrable risk, limited in scope and duration, logged, and subject to independent review. Confidentiality must protect safety without becoming a shield for avoidable risk or institutional misconduct.

Article 10 — Coordination Across High-Consequence Domains

AIPA should not absorb specialized institutions or claim expertise it does not possess. It should establish formal coordination pathways with bodies responsible for biological, nuclear, cyber, financial, health, environmental, and other high-consequence domains.

Where AI and biotechnology converge, AIPA should coordinate with a specialized body such as the proposed International Biotech Protection Agency (IBPA), or with legitimate equivalent institutions. Joint

assessment should apply where neither AI assurance nor biotechnology oversight alone can evaluate the full risk.

For nuclear and other established domains, AIPA should work with recognized institutions and agreements, respecting their mandates while providing AI-specific evaluation and operational assurance. Coordination should prevent both regulatory gaps and unnecessary duplication.

Article 11 — Extreme Risk and Coordinated Slowdown

When credible evidence indicates severe, rapidly escalating, or potentially irreversible risk that cannot presently be mitigated, AIPA should be able to issue urgent alerts, convene qualified authorities and affected operators, recommend restrictions or temporary pauses, and coordinate an internationally informed slowdown in defined activities.

Any emergency recommendation should be evidence-based, limited to the affected capability or deployment pathway where possible, time-bounded, reviewable, and accompanied by clear conditions for modification or release. Emergency coordination must not become an unchecked permanent power.

Article 12 — Development, Adoption, and Revision

This Model Charter is offered for collaborative development by governments, international institutions, AI laboratories, open-source communities, engineers, scientists, standards experts, insurers, infrastructure operators, civil society, and representatives of communities affected by AI deployment.

The founding process should convert these principles into a legitimate legal and institutional form, define assurance levels and certification procedures, determine representation and financing, establish relationships with existing bodies, and pilot the assessment framework before broad formalization.

The Charter and all resulting standards should remain subject to evidence-driven revision. Version history, reasons for change, dissenting technical views, and transition periods should be recorded and publicly accessible wherever safety permits.

Invitation. AIPA is presented as a template idea, not an institutional claim. Others are invited to improve it, test it, and help build an international protection framework capable of matching AI's global reach and accelerating consequences.

Article 13 — Closing Principles

- Human life, dignity, agency, and civilizational continuity take precedence over narrow competitive advantage.
- Safety requirements should be proportionate, evidence-based, technology-neutral, and internationally interoperable.
- Pre-release evaluation and continuing operational verification are complementary obligations.
- No institution, nation, company, or technical provider should be trusted without appropriate verification and accountability.
- Protection from globally consequential AI cannot depend exclusively on the same geopolitical and commercial pressures that accelerate its development.
- The benefits of AI should be advanced while preserving humanity's authority to choose its future.

ILLUSTRATIVE FORMATION PATHWAY

A practical starting sequence, not a permanent constitutional requirement

1. Voluntary pilot

Frontier laboratories and selected high-consequence operators voluntarily provide systems and evidence for advance review. An initial advance-review period of up to 30 days, as proposed by Demis Hassabis, may be tested and adjusted by system type and evaluation need.

2. Shared protocol development

Independent experts, participating laboratories, open-source representatives, domain institutions, and public-interest participants develop and test thresholds, benchmarks, evidence requirements, confidentiality rules, and review procedures.

3. Independent capacity

AIPA builds its own technical staff, secure evaluation environments, held-out tests, compute resources, and accredited third-party auditor ecosystem so it does not depend on the systems it evaluates.

4. Certification pilot

A limited certification program tests whether findings are reproducible, standards are implementable, operational evidence remains trustworthy, appeals work, and material changes trigger reassessment.

5. International recognition

Governments and other institutions recognize mature standards through law, procurement, market access, insurance, infrastructure access, financing, or mutual-recognition agreements while preserving AIPA's technical independence.

6. Continuing internationalization

The framework expands beyond its initial participants and remains open to systems from any country and to both open and closed models, with equivalent proof requirements for equivalent risks.

Source and development note

This Model Charter updates SafeWave Systems' December 2025 AIPA proposal and incorporates practical elements from Demis Hassabis's July 14, 2026 proposal for a frontier AI standards body. It also reflects Google DeepMind's Frontier Safety Framework approach to capability thresholds, safety cases, deployment mitigations, and continuing review, together with Rich Barton-Cooper and Adam Gleave's August 3, 2026 proposal for independent jailbreak disclosure and coordinated remediation. These sources informed the framework; they do not endorse this document.

- [Demis Hassabis, "A Framework for Frontier AI and the Dawning of a New Age," July 14, 2026.](#)
- [Google DeepMind, "Frontier safety" and Frontier Safety Framework 3.1, April 17, 2026.](#)
- [Rich Barton-Cooper and Adam Gleave, "AI Jailbreak Disclosure Is Broken. Here's How to Fix It," August 3, 2026.](#)