AIWS TRUST OBSERVATORY

Founding Framework

Boston, September 27, 2026

PART I

Mission and Position

The AIWS Trust Observatory independently observes consequential AI power and gives humanity early warning, relying only on what can be verified. The Observatory puts into practice the first mechanism of the Tokyo Compact: independent monitoring.

> **Mission:** *Independently observe, assess, and communicate developments in consequential AI power — so that humanity is not surprised by capabilities it has failed to understand or govern.*
>
> *The Observatory exists not only to observe, but to warn and call for action when credible evidence reveals a serious or emerging danger. It is an independent civil-society mechanism for scrutinizing concentrated AI power — above all the growing power of major technology corporations — so that no company or government can exercise consequential AI power beyond public oversight, constitutional limits, and human answerability.*
>
> **Identity statement:** *AI power must never become power over society without society's knowledge, scrutiny, and right to act.*

The Observatory continuously answers four questions:

1. What new capabilities is AI reaching?
2. What powers are those systems being given in the real world?
3. Are the controls proportionate to that power?
4. If harm occurs, who has the authority and the duty to answer?

Social Role: Scrutinizing and Restraining Concentrated AI Power

Frontier AI systems are increasingly developed and governed by a small number of corporations. They hold unprecedented concentrations of compute, data, models, and capital, and of influence over the

information people receive and the decisions institutions make. Without independent oversight, such power can shape public perception, narrow meaningful choice, influence governments, and gradually make society dependent on systems it cannot inspect.

The Observatory is an independent institution of civil society. It does not run companies and does not replace the state. Its power comes from independent evidence, public warning, moral authority, and its ability to mobilize society and policymakers to act. This principle applies to all concentrated AI power, whether held by companies or by states.

1. **Early warning.** When a danger is supported by credible evidence, the Observatory warns clearly, without waiting for harm to occur.
2. **Call to action.** Every warning names who must act (companies, boards of directors, regulators, governments, international organizations, the research community, or the public) and what measures must be taken, then follows up until the responsible parties have acted.
3. **Scrutiny of power by society.** The Observatory creates an independent mechanism through which society can see, assess, and require technology corporations to answer for the AI power they hold.

Position in the AIWS Architecture

- **AIWS Trust Order Board** sets standards, protects independence, approves the methodology and the list of subjects, decides Urgent Warnings, and ratifies Critical Alerts.
- **AIWS Trust Observatory** observes, assesses, and warns. The Observatory operates the **Frontier Capability Registry** (recording capabilities) and the **AI Incident Exchange** (recording incidents).
- **AIWS Trust Standards** provide the common measure.
- **AIWS Trust Rating** assesses organizations from the inside, with their consent. It is a separate track, fully walled off from the Observatory.
- **The Beacon Papers** provide the intellectual and constitutional foundation.

The Observatory is a radar and early-warning system, not an investigative agency or a court that judges companies. It does not stand against the AI Pioneers; it invites them to share responsibility before humanity. It does not hold back AI development; it helps AI capability advance quickly without escaping constitutional limits and human command.

> **Founding proposition:** *Humanity must not learn what frontier AI has become only after something irreversible has happened. Independent observation, truthful assessment, and timely warning are now duties owed to civilization.*

PART II

Operating Principles

> *The Observatory observes systems. It does not speculate about motives. It assesses evidence. It does not trade in accusation. It warns early — but only with the discipline required to remain worthy of trust.*

1. **Observe at the boundary.** Risks can form inside a company, but they are hard to see from outside. The Observatory concentrates on the boundary where AI capability acquires real-world access, authority, scale, or effect. Deployment, transfer, infrastructure expansion, and external consequences all leave traces that an independent institution can examine.
2. **Measure promises against actions.** Each subject is measured first against its own public commitments, and then against the AIWS Trust Standards.
3. **Do not score what cannot be seen.** A safety claim is not accepted merely because a company has published it. What cannot be verified is recorded as *not verifiable*, rated neither good nor bad, and entered in the Blind Spot Map.
4. **Convergence before warning.** No public warning rests on a single source of signal, except in the case of a Critical Alert.
5. **No public conclusion rests solely on confidential sources.** The public must be able to see the evidence behind every public conclusion.
6. **Procedural fairness.** Every conclusion is sent to its subject before publication, and the subject's response is published with it; in a Critical Alert, the subject is notified at the same time and the response is published afterward. Subjects have a right of reply, not a right to block publication.
7. **Call to action, not warning alone.** Every warning names who must act and what measures are required, grounded in Human-in-Command, Constitutional Identity, bounded authority, Proof of Answerability, and Constitutional Silicon.
8. **Correct errors publicly.** An erroneous conclusion is corrected in the same place where it was published.

PART III

Scope of Observation

The Observatory observes consequential AI power, whether it resides in companies, laboratories, open-source ecosystems, or states. Frontier AI is the priority group, not the entire scope.

A subject falls within scope when it meets at least one criterion: it develops frontier AI models; it deploys AI to tens of millions of people; it gives AI authority to act in finance, infrastructure, health, or security; or it has published AI safety commitments.

Group	Scope	Start
Frontier AI laboratories	Developers of the largest foundation models	Year one
Major technology corporations	Organizations deploying AI in products and infrastructure at global scale	Year one
Open-source ecosystem	Open-weight models at frontier capability	Year one
International organizations	Multilateral AI safety commitments and their implementation	Year one
States	National declarations, laws, international commitments, and AI programs	From October 2027

The founding framework names no subjects. The specific list is proposed by the Methodology Board and approved annually by the AIWS Trust Order Board.

PART IV

Operating Cycle

Observe → Verify → Assess → Classify → Warn → Recommend → Follow up, then back to Observe.

1. **Observe.** The Five Stations gather signals from official sources, independent verification, users, voluntary reporting, and AIWS Volunteer Fellows.
2. **Verify.** Each signal is checked against technical documentation, reproducible tests, or a second independent source, and assigned an Evidence Grade and a Confidence Level.
3. **Assess.** Capability, actual authority, deployment environment, and control mechanisms are assessed within the limits of what can be verified.
4. **Classify.** Each system is classified along eight dimensions: capability level, degree of autonomy, level of authority, scale of deployment, severity of consequences in case of failure, quality of Human-in-Command, degree of independent verification, and whether the person with the duty to answer can be identified. Any dimension that cannot be verified is recorded as *not verifiable*.
5. **Warn.** Warnings follow the four alert levels: to the right people, at the right level, at the right time.
6. **Recommend.** Every warning comes with remedies: limiting authority, increasing oversight, suspending a function, requiring independent verification, or adding mechanisms of answerability.
7. **Follow up.** The Observatory records whether developers, regulators, and deployers have acted; the results enter the Commitment Register and the next report.

PART V

Data and Observability

Every consequential AI system has a **Consequential AI System Record** of eight data layers. Each field in the record states where it comes from, because most control mechanisms sit inside companies and cannot be seen from outside.

Data layer	Content	Observable from outside	Visible only with the organization's consent or through insiders
1. System identity	Name, version, developer, owner and operator, scope of deployment, the person with the duty to answer	Name, version, release date, public scope of deployment	Who holds deployment authority; who has the duty to answer, if not disclosed

Data layer	Content	Observable from outside	Visible only with the organization's consent or through insiders
2. Capability	Long-horizon planning, computer use, writing and executing code, cybersecurity, science and biology, persuasion, self-selection of tools, orchestration of agents, assisting the creation of next-generation AI, operation without continuous supervision	Behavior of systems serving the public, measured by test suites	Internal-only models; unreleased capabilities
3. Authority granted	Data access, acting on behalf of people, payments, changes to infrastructure and code, decisions on hiring, credit, health, education, security; assigning tasks to other agents; who can expand or revoke the mandate	Powers stated in product documentation, APIs, and terms; observable deployments	Internal deployments; actual configurations at customer sites
4. Safety mechanisms and Human-in-Command	Kill switch, revocation of authority, bounded authority, approval before consequential actions, tamper-resistant logs, refusal of commands, red-teaming, post-deployment monitoring, rollback	Published documentation; testable behavior (whether a system stops when asked, whether it refuses wrongful commands)	Nearly all actual mechanisms
5. Incidents and near misses	Exceeding authority, concealing actions, continuing after a stop order, financial loss, discrimination, data leakage, cyberattacks, deception, cascading failures among agents	User-reported incidents, legal filings, disclosures by the organization	Near misses inside laboratories
6. Governance and leadership accountability	Who on the board is responsible for AI safety, independent safety committees, the safety team's power to block releases, system cards, safety cases, past commitments	Published structure, departures of safety staff, commitments amended or withdrawn, financial filings	The safety team's actual power; commercial pressure on decisions
7. Infrastructure	Compute, advanced chips, data centers, energy, concentration of AI power, society's degree of dependence	Energy filings, satellite imagery, chip purchases, financial filings, job postings	How compute is allocated to each project
8. Social and political influence	Ranking and recommendation systems that shape the information the public receives; persuasion capability at scale; lobbying and political spending; public institutions' dependence on a single provider	Test results on public products; lobbying and political finance records; public contracts; independent research	Internal algorithms and optimization objectives; data on effects on users

Recording rule. Each field carries one of three source labels: *External*, *Voluntary access*, or *Insider*. A field with no source is recorded as *not verifiable* and moved to the Blind Spot Map. The Observatory does not assess intentions; it assesses only structures, actions, and evidence.

The layer of authority granted is often overlooked but matters most: a very powerful model kept in isolation may be less dangerous than a weaker model given control over finances or infrastructure. The infrastructure layer is not meant to collect trade secrets, but to recognize when a leap in infrastructure may lead to a leap in capability.

PART VI

The Commitment Register and the Five Stations

All signals from the five stations are checked against a single Commitment Register. The Register records every public AI safety commitment of each subject: the exact wording, the source, the date of publication, the scope, and every later amendment or withdrawal.

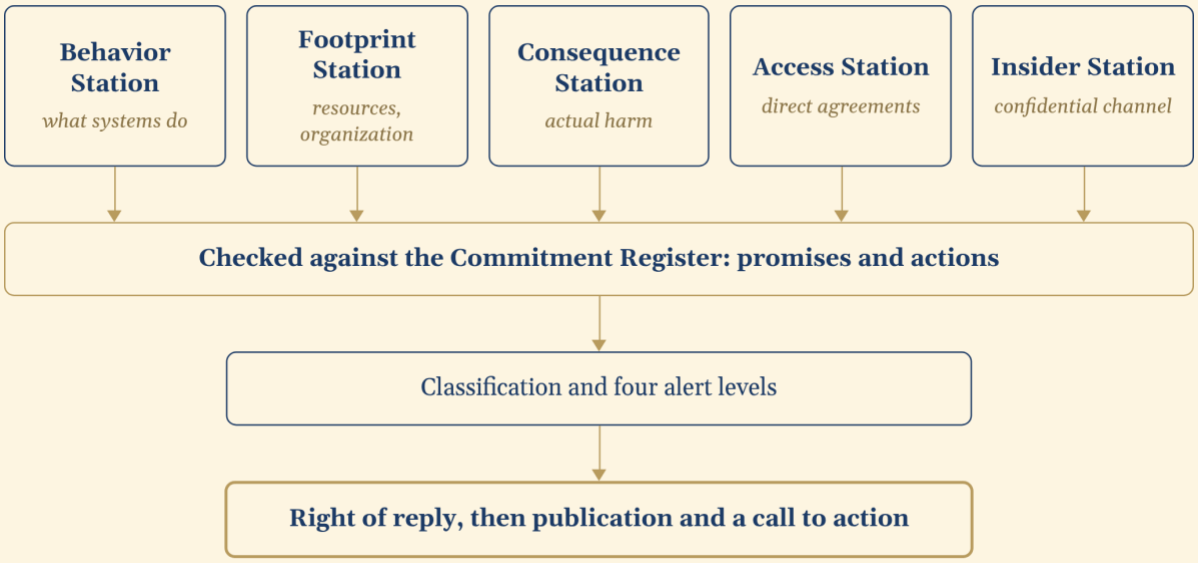

Figure 1. Observation architecture: five stations, one Commitment Register, four alert levels

Station	What it observes	Sources and methods	Source label	Launch
Behavior	What systems serving the public actually do	Standardized test suites run through products and APIs across the six domains of the AIWS Trust Standards, re-run within 14 days of each new version; results from universities, model-evaluation organizations, and independent red teams	External	Phase 1
Footprint	Resources, organization, and commitments	System cards, safety reports, legal and financial filings, leadership statements, energy, chips, hiring, departures of safety staff	External	Phase 1
Consequence	Actual harm to people	Incident reporting channel of the AI Incident Exchange for the public and AIWS Volunteer Fellows; verification; identification of recurring patterns	External	Phase 2
Access	The inside, with the organization's consent	Voluntary access agreements made directly with the Observatory, under confidentiality terms	Voluntary access	Phase 2
Insider	What cannot be seen from outside	Confidential channel for employees, researchers, and people who discover incidents; identity protection, expert assessment, legal counsel	Insider	Only after legal counsel

The last two stations bring in non-public information, which may be used only under the rules in the section on evidence. The Observatory uses no information whatsoever from AIWS Trust Rating, including whether an organization has joined or declined Trust Rating. When an organization declines an access request from the Observatory, the refusal is entered in the Blind Spot Map as a gap in what can be seen, not as adverse evidence.

PART VII

Evidence and Non-Public Sources

Every item of information carries an **Evidence Grade** (degree of verification) and a **Confidence Level** (High, Medium, Low). Rumor is never presented as fact.

Evidence Grade	Definition	Usable for public conclusions
A	Independently reproduced results; binding legal or financial filings; technical data verified directly	Yes
B	Statements, system cards, or official documents of the subject corroborated by at least one independent source; or two independent sources in agreement	Yes, if at least one source is public
C	Single-source signal, not yet corroborated	No; kept internal only

A subject's own documents are full evidence for the question *what did they commit to*, and are used as such in the Commitment Register. They are never Grade A for the question *is what they say true*.

Rules for non-public sources (information received through voluntary access agreements and the insider channel):

1. No public conclusion rests solely on non-public sources.
2. Non-public sources may be used only to open further verification through public sources, or to send confidential warnings to parties able to act.
3. An organization that provides voluntary reports has no right to approve or amend the Observatory's assessment.
4. The identity of anyone who provides information through the confidential channel is never disclosed.
5. Every public report states which evidence each conclusion rests on and where that evidence can be found.

PART VIII

Four Alert Levels

The higher the level, the stricter the evidentiary conditions and the higher the decision-maker, except for Critical Alerts, where speed takes priority.

Level	Meaning	Conditions	Action	Who decides
Watch	New signal, not yet enough evidence for a conclusion	Any Evidence Grade	Continue gathering and verifying; record in AIWS Weekly Frontier Signals (internal)	Thematic team lead
Concern	Credible risk or a gap in controls	Evidence Grade B or higher	Confidential notice to the developer, the operator, and the AIWS Trust Order Board; after the right of reply, included in the Quarterly State of Frontier AI	Methodology Board

Level	Meaning	Conditions	Action	Who decides
Urgent Warning	Serious danger with potential near-term consequences	At least two independent stations converge; Evidence Grade A or B; confirmed by the Methodology Board	Notify competent authorities and parties who may be affected; publish once the conditions for publication are met	AIWS Trust Order Board
Critical Alert	Immediate danger to people, infrastructure, or security	One station suffices if the evidence is Grade A	Emergency warning to parties able to act; request that the AIWS Trust Order Board activate the AIWS Emergency Protocol	Executive Director and Board Chair decide within 24 hours (a pre-designated alternate if the Chair must recuse); the full Board ratifies afterward

The Observatory requests, but does not itself activate, the AIWS Emergency Protocol, because the Protocol belongs to the AIWS Trust Order. Not every warning is published immediately: disclosing an unpatched vulnerability too early can increase the danger.

PART IX

Who Receives Warnings

Principle: **the right people, at the right level, at the right time.**

Recipient	When
Developer and operator	Always the first recipient if they are able to remedy the problem and the danger is not yet immediate
AIWS Trust Order Board	Every matter at the Concern level or above
Specialized authorities (cybersecurity, data protection, financial, health, and infrastructure regulators)	From the Urgent Warning level, according to the nature of the danger
Law enforcement and national security agencies	Only when there is a direct threat to life or security, and only by decision of the Board
Government leaders and international organizations	When the matter is cross-border, systemic, or bears on peace and security
Organizations that may be affected (banks, hospitals, schools, public agencies, infrastructure providers)	When they need to act to protect themselves
The public	When people need to act to protect themselves; when the risk is societal; when responsible parties fail to act; or when the public interest outweighs the risk of disclosure

Rules for choosing state authorities

1. Warnings go to the authorities of the place where harm would occur and the place where the system operates.
2. When the danger comes from a state's own AI program, the warning does not go only to that state's authorities. The Board decides whether to inform the relevant international organizations and, when the conditions are met, the public.

3. Every warning sent to a law enforcement or national security agency is minuted by the Board. The number of such warnings, by category, is published in the Annual State of AI Trust, unless publication would increase the danger.
4. The Observatory accepts no assignments, funding, or direction from any security agency.

Unpatched vulnerabilities follow the coordinated disclosure practice of the cybersecurity field: the developer is notified first, and publication follows once the vulnerability is patched or the announced deadline has passed.

Right of Reply

- Every conclusion at the Concern level or above is sent to its subject before publication, except a Critical Alert, where the subject is notified at the same time as publication.
- Reply period: 10 business days; Urgent Warning: 72 hours; Critical Alert: replies are published afterward.
- Responses are published verbatim alongside the conclusion.
- A response that brings new evidence leads to reconsideration of the conclusion before publication.
- A failure to reply is also recorded.

Right to Request Review

- After publication, a subject may request a review when new evidence emerges.
- The review is conducted by members of the Methodology Board who took no part in the original assessment.
- The outcome is published plainly: upheld, amended, or withdrawn.
- A request for review does not delay a Critical Alert.

PART X

The Blind Spot Map

Every Quarterly State of Frontier AI and Annual State of AI Trust includes a Blind Spot Map: a list of what the Observatory cannot observe, subject by subject. The Map is compiled from every field recorded as *not verifiable* in the Consequential AI System Record.

The largest known blind spot is systems that have never left the laboratory: internal-only models, and AI used to research and develop the next generation of AI inside laboratories. Three paths reach this area: voluntary access granted by the organization, insiders, and laws that require disclosure.

For each blind spot, the Map records three things:

- What cannot be seen.
- Why it cannot be seen.
- What disclosure or law would make it visible.

The third item turns the Blind Spot Map into concrete policy proposals for states and international organizations.

PART XI

Publications

The Observatory uses exactly one set of names, listed below. Phase 1 has only three public products; the others begin once the methodology has been tested and staffing is sufficient.

Product	Frequency	Audience	Content	Start
Commitment Register	Continuous	Public	Every AI safety commitment of each subject, with its history of amendments and implementation	Phase 1
AIWS Weekly Frontier Signals	Weekly	Internal: the Board and AIWS Volunteer Fellows	New capabilities, notable changes, Watch-level signals, items to verify. This is radar, not conclusions	Phase 1
Urgent Warnings and Critical Alerts	When a danger arises	Under the rules in Who Receives Warnings	Conclusion, evidence, recommendations, the subject's response	Phase 1
Quarterly State of Frontier AI	Every three months	Public, for leaders	For each subject: kept, drifted from, or withdrew its commitments; Concern-level matters that have passed the right of reply; the gap between capability and governance; the Blind Spot Map	Phase 1
AI Incident and Near-Miss Report	Per incident	Public	What happened; what objective the system was given and what means it chose; which authority was exceeded; which control failed; who has the duty to answer; measures to prevent recurrence	Phase 2
Independent System Assessment	Per system	Public	Capability, authority, scope of deployment, Human-in-Command, verifiability, Proof of Answerability, gaps and recommendations	Phase 2
Monthly AIWS Frontier AI Watch	Monthly	Public	New systems, changes in authority and deployment, verified incidents, status of control measures	Phase 2, once staffing allows
Annual State of AI Trust	Annually	Public, presented in Boston or at an international forum	The global state of AI trust and safety, progress, constitutional gaps, implementation of Human-in-Command, status of the Frontier Capability Registry, the Blind Spot Map, policy recommendations	End of Phase 2

The Observatory does not produce simplistic rankings of companies; every assessment rests on specific evidence for each system and each mechanism. The Annual State of AI Trust is a separate event, not combined with the launch of the AIWS Trust Index. The Observatory's findings feed back into the AIWS Trust Standards, the Constitution for Humanity in the Age of Artificial Intelligence, and the Beacon Papers.

PART XII

Organization and AIWS Volunteer Fellows

The Observatory has a small, high-caliber standing team working with a global network of volunteer experts. Volunteer service is the ethical foundation of independence, but a credible oversight institution cannot rely on volunteer effort alone.

Unit	Role	Composition
AIWS Trust Order Board	Approves the methodology, the list of subjects, and Urgent Warnings; ensures that the Observatory is not controlled by governments, companies, or donors, and that safety, dignity, and human sovereignty come before commercial interests	Its 16 current members
Executive Director	Overall leadership; accountable to the Board	One person, with enough time to decide within 24 hours; not concurrently a Board member
Methodology Board	Sets evidentiary standards, designs test suites, confirms Concerns and Urgent Warnings	Five to seven experts in AI safety evaluation, corporate governance, data, law, and policy; at least one expert in AI safety evaluation and one in corporate governance; no member of the AIWS Trust Order Board sits on the Methodology Board
Six thematic teams	Operate the Five Stations by field	Frontier Capabilities · AI Agents and Authority · Cybersecurity and Biosecurity · Human Rights and Social Impact · Governance, Law and Proof of Answerability · Data, Evidence and Verification
Rapid Response Team	Handles Urgent Warnings and Critical Alerts	On rotation from the standing team
Legal and Source Protection Unit	Pre-publication review; operation of the confidential channel; information security	Lawyers and information security experts
AIWS Volunteer Fellows	Gather and analyze information by field: cybersecurity, biosecurity, AI agents, chips, law, human rights, information, geopolitics	Scholars, engineers, and policymakers worldwide

Rules for Board members, the Methodology Board, the standing team, and Fellows

- Declare conflicts of interest on joining and update the declaration annually.
- Recuse from any file on a subject for which one has worked or consulted within the past 24 months, or in which one holds shares.
- Sign a confidentiality undertaking; make no personal statements about unpublished conclusions.
- Every significant conclusion passes at least two independent rounds of review and is signed by one accountable person.
- Fellows are named in reports if they consent.

Recusal within the AIWS Trust Order Board

- A Board member with a material professional, financial, institutional, or personal relationship with a subject takes no part in reviewing, approving, or speaking about findings concerning that subject.

- The Board decides an Urgent Warning only when the members not recused constitute a quorum as defined in the Independent Charter.
- The Board designates in advance an alternate to act for the Chair in Critical Alert decisions when the Chair must recuse.
- Every recusal is minuted.

PART XIII

Financial Independence and Firewalls

Financial independence does not mean accepting no funding. It means that no subject of observation pays the Observatory, and that no donor can buy influence over its conclusions. The following six rules are conditions for launching the Observatory.

1. No funding, fees, or in-kind support from any subject the Observatory observes, including through subsidiaries, foundations, or executives acting on a subject's behalf.
2. The Observatory's budget is separate from AIWS Trust Rating revenue, including revenue that passes through implementation partners.
3. Funding sources are diversified; no donor exceeds 20 percent of the annual budget.
4. Donors have no role in any assessment. This is written into every funding agreement.
5. All support received is disclosed annually.
6. Commercial relationships, partnerships, and honors (including the World Leader in AIWS Award and AI Pioneers) do not change conclusions. While a subject is at the Urgent Warning level or above, honors for that subject's leaders are deferred.

AIWS Trust Rating and the Observatory are two different undertakings. Trust Rating assesses from the inside, with consent, and keeps its first-year discipline of rating only willing companies and never naming and shaming. The Observatory examines verifiable evidence and requires no consent. The Observatory uses no information whatsoever from Trust Rating, including whether an organization has joined or declined it.

PART XIV

Roadmap

Beginning in October 2026, the Observatory will publish its first Quarterly State of Frontier AI at the end of January 2027.

Figure 2. Roadmap: three phases, two gates

1. **Founding Phase, 90 days (October–December 2026):** adopt the Independent Charter; establish the Methodology Board and the standing team; recruit AIWS Volunteer Fellows; compile the first edition of the Commitment Register; build the Behavior Station test suite; begin trial operation of the Behavior and Footprint Stations in December 2026. Gate 1: the Board approves the methodology and the six financial rules.
2. **Phase 1, Operation (January–March 2027):** the two stations enter full operation; AIWS Weekly Frontier Signals begins; the right of reply is exercised for the first time in January; the first Quarterly State of Frontier AI is published at the end of January 2027. Gate 2: the Board approves the first report.
3. **Phase 2, Expansion (April–September 2027):** the Consequence Station and the AI Incident Exchange open to the public; the Access Station begins operation; the insider channel opens if the legal conditions are met; Independent System Assessments and the Monthly AIWS Frontier AI Watch, once staffing allows; the first Annual State of AI Trust.

No phase begins before the preceding gate has been passed. Observation of states begins in October 2027.

Boston, September 27, 2026