THE BEACON PAPERS
No. 13
Foundational papers for humanity in the AI Age — from Boston

AI POWER MUST NOT OPERATE UNSEEN

Independent Observation, Early Warning, and Society's Right to Act

Nguyen Anh Tuan
Founder and Chief Architect of AIWS
Co-Founder, Co-Chair and CEO, Boston Global Forum

AIWS Lumina Lab • Beacon Hill, Boston • 27 September 2026

The world does not need centralized control of AI. It needs independent global observation.

I. The Warning from the Security Council

On September 23, 2026, the leaders of the world's most powerful AI laboratories brought an extraordinary warning before the United Nations Security Council.

Anthropic CEO Dario Amodei said that, if managed poorly, *"AI could be a risk to humanity as a whole."* OpenAI CEO Sam Altman declared: *"If AI is to be democratic, the most important decisions cannot be made by labs in San Francisco alone."*

Yet the United States rejected centralized global control. White House science and technology adviser Michael Kratsios told the Council: *"You cannot govern technology you do not understand. This body and others like it should focus on sharing best practices to build domestic capacity, not establishing a global regulatory scheme."*

The two positions reveal the central governance problem of the AI Age. Frontier AI is becoming too powerful to be governed only by the companies building it. But concentrating control in a global authority could create another form of unaccountable power.

The world should not have to choose between rule by AI corporations and rule by a global bureaucracy.

> ***The world does not need a global AI government. It needs a trusted, independent, and evidence-based global mechanism for observing AI power, warning of danger, and calling those responsible to act.***

II. The Danger Is No Longer Theoretical

AI is moving from generating content to conducting research, operating computers, writing and executing code, coordinating other agents, identifying cyber vulnerabilities, and acting across the digital world.

The danger arises not from capability alone, but from the convergence of capability with authority. An AI model confined to a laboratory may pose less immediate danger than a weaker system authorized to transfer money, alter infrastructure, make decisions affecting employment or health care, influence millions of people, or command other autonomous agents.

At the same time, compute, advanced chips, frontier models, data, and channels of public information are increasingly concentrated within a small number of corporations and governments. Much of the most consequential activity occurs inside systems that society cannot inspect: unreleased models, internal agents, safety failures, near misses, and AI systems used to build the next generation of AI.

This creates a dangerous asymmetry. The organizations developing frontier AI can see much of what their systems are becoming. Governments and the public often learn only what those organizations choose to disclose, or what becomes visible after an incident.

Self-disclosure is necessary. It is not independent oversight.

> ***Disclosure is what power chooses to say about itself. Observation is society's ability to verify how that power is actually exercised.***

III. The AIWS Trust Observatory: Independent Eyes on AI Power

The Boston Global Forum and AIWS establish the **AIWS Trust Observatory** as an independent civil-society institution to observe consequential AI power, verify evidence, warn of emerging danger, and call upon those with authority to act.

Its mission is direct:

> ***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 does not govern AI companies, replace national regulators, or seek centralized control over technological development. Its power comes from independent evidence, methodological integrity, public warning, moral authority, and the ability to mobilize companies, governments, international institutions, researchers, and society around verified risks.

It continuously asks four questions:

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

The Observatory will examine frontier laboratories, major technology corporations, consequential open-source models, international commitments and, in a later phase, national AI programs. It will observe not only technical capabilities but also concentrations of power: control over compute, infrastructure, public information, political influence, and the dependence of public institutions on a small number of technology providers.

> ***AI power must never become power over society without society's knowledge, scrutiny, and right to act.***

IV. From Observation to Action

The Observatory operates through a continuous cycle:

> ***Observe → Verify → Assess → Classify → Warn → Recommend → Follow up***

Its institutional architecture is **Five Stations, one Commitment Register, and four alert levels**.

The **Five Stations** observe:

- **Behavior Station** — how AI systems behave;
- **Consequence Station** — what consequences they produce;
- **Footprint Station** — what infrastructure, resources, and organizational changes reveal;
- **Access Station** — what becomes visible through voluntary access;
- **Insider Station** — what protected insiders report about dangers the outside world cannot see.

The **Commitment Register** records the public safety commitments of each organization: their original language, scope, and date, every subsequent change, and whether observable conduct is consistent with them.

The Observatory distinguishes four alert levels:

- **Watch** — a new signal requiring further verification;
- **Concern** — a credible risk or a gap in control;
- **Urgent Warning** — a serious danger with a meaningful likelihood of near-term consequences;
- **Critical Alert** — an immediate threat to people, critical infrastructure, or security.

When credible evidence reveals danger, the Observatory will not merely publish analysis. It will identify who must act, what action is required, who may be affected, who has the duty to answer, and whether the responsible parties have taken corrective action.

The Observatory will operate the **Frontier Capability Registry** and the **AI Incident Exchange**. It will publish the **Quarterly State of Frontier AI**, the **Annual State of AI Trust**, independent system assessments, incident reports, and a **Blind Spot Map** identifying what remains hidden and what disclosure or law is needed to make it observable. The first Quarterly State of Frontier AI will appear at the end of January 2027.

V. The Discipline Required for Trust

An institution that observes power must itself be answerable.

The Observatory will not speculate about motives or present rumor as fact. Every finding will carry an evidence grade and a confidence level. No public conclusion will rely solely on confidential information. Those being assessed will have the right to respond, to present new evidence, and to request reconsideration. Errors will be corrected publicly.

The Observatory will disclose what it cannot verify rather than treat absence of access as evidence of wrongdoing.

Its financial independence will be protected by a strict firewall. No company or government under observation may finance its work. No donor may influence an assessment. The Observatory will remain separate from AIWS Trust Rating and will use no information from the rating process.

These safeguards are essential. The Observatory must never demand from others a standard of answerability that it refuses to apply to itself.

VI. No Central Control, No Unobserved Power

The position of AIWS is clear:

> ***Do not centralize control over AI. Do not leave consequential AI power unobserved.***

The AIWS Trust Observatory offers a third path between two dangerous concentrations of power: AI governed primarily by a handful of corporations, and AI governed by a centralized global authority.

It does not make binding decisions for nations. It creates the independent knowledge that nations, institutions, companies, and citizens need to make responsible decisions.

It does not slow AI merely because AI is new. It seeks to ensure that consequential AI capability remains within bounded authority, meaningful Human-in-Command, Constitutional Identity, Proof of Answerability, and verifiable safeguards.

It does not stand against the AI Pioneers. It calls upon them to help build the trust that their creations now require.

VII. A Call to Act

To AI laboratories and technology companies: open consequential systems to credible independent scrutiny before failure or public distrust imposes the terms upon you.

To governments: protect national sovereignty, but do not use sovereignty as a shield against answerability. Constitutional restraint must apply to public and private AI power alike.

To universities, scientists, journalists, and civil society: contribute expertise, evidence, time, and judgment to an independent global effort.

To international institutions: support shared standards, incident notification, verification, and cooperation without seeking centralized control over the technological future of humanity.

The warnings delivered at the United Nations Security Council cannot become another record of what the world understood but failed to act upon.

> ***Humanity must not learn what frontier AI has become only after something irreversible has happened.***

AI power must be observed before it can be governed — and answerable before it can be trusted.

SOURCES

"AI leaders warn UN of security risks as systems grow more powerful," Reuters, September 23, 2026, republished by MarketScreener. https://www.marketscreener.com/news/ai-leaders-warn-un-of-security-risks-as-systems-grow-more-powerful-ce785aded98df321

"OpenAI, Anthropic CEOs call for global AI regulation at UN," Al Jazeera, September 24, 2026. https://www.aljazeera.com/news/2026/9/24/ai-corporate-leaders-tell-un-the-industry-needs-global-regulation

Boston, September 27, 2026