THE BEACON PAPERS · No. 12

Foundational papers for humanity in the AI Age — from Boston

THE AI RACE NEEDS A TRUST ORDER

Beyond the False Choice Between Safety and Speed

Nguyen Anh Tuan

Co-Founder, Co-Chair and CEO, Boston Global Forum

Founder and Chief Architect, AIWS

A contribution to the architecture of AIWS Civilization

THE CENTRAL PROPOSITION

The defining choice of the Age of AI is not between safety and speed. It is between an uncontrolled race for capability and a trusted order in which innovation advances under human authority.

I. The Great Division

In September 2026, the division became explicit. Leaders of frontier AI laboratories called for pacing the development of increasingly powerful systems, warning that capabilities were advancing faster than safeguards could keep up — driven increasingly by AI's ability to build its own successors. [1]

President Donald Trump dismissed the alarm. He declared that the only guardrail AI needs is a strong and smart president, that the United States already has the tools to police the industry, and that doubt about AI development serves China. [2]

Others argue that AI has become too strategically important to slow in a geopolitical race. [3]

Both sides recognize a real danger. Neither provides a sufficient order. The real dispute is not whether safety matters. It is who defines acceptable risk, who verifies the evidence, who may authorize deployment, who answers when an AI system exceeds its mandate — and whether anyone can exercise restraint while believing others will keep racing.

A race without verified restraint is not strategy. It is the transfer of consequential power to systems and incentives that no government fully commands.

II. Why Both Sides Are Partly Right

The safety warning is justified. AI can lower barriers to biological and chemical threats, accelerate cyber offense, manipulate information at scale, and act through agents whose conduct is hard to predict or reconstruct. The central danger is not a wrong answer. It is a system that acts beyond the authority its operator intended, fails to report what it has done, and triggers consequences faster than institutions can intervene.

The competition warning is also justified. A nation that abandons advanced AI becomes dependent on the infrastructure and decisions of others. Unilateral restraint is unstable when competitors are not equally bound. Poorly designed rules can entrench incumbents and exclude smaller nations. States will not accept an order that asks them to become weak.

But competition does not remove the need for trust. It makes verified, reciprocal obligations more urgent. The realist objection is not an argument against a Trust Order. It is the reason a Trust Order must be built on verification and reciprocity rather than on appeals to virtue.

III. The Failure of the Two Extremes

Safety fails when it rests on voluntary promises by the same institutions that gain power and revenue from frontier systems. Self-assessment is necessary. It cannot be the final source of legitimacy. A laboratory may publish a scaling policy and still remain the only judge of whether that policy has been kept.

Speed fails when winning becomes the only standard. In a capability race, every warning becomes an argument for acceleration. The result is a system no actor intended but every actor helped create: greater capability, weaker restraint, fragmented accountability, and declining human control.

Personal command is not a constitutional order. A president, a board, or a founder may be strong. None of them is a substitute for identity, mandate, evidence, and answerability that survive a change of office, a change of ownership, or a change of incentive.

> ***Safety without accountability can concentrate power. Speed without authority can release power. A Trust Order must govern both.***

IV. The Object of Governance Is Authority

Speed is hard to measure, differs across domains, and arises from algorithms, data, tools, and scale as much as from model size. A general demand to slow AI is too vague to enforce and too broad to justify. It also hands opponents an easy charge: that safety is a pretext for delay.

The durable object of governance is authority: what an AI system may do, where, with what resources, under whose identity, within which limits, and with what duty to report. The same model presents one risk in isolation and another when connected to financial networks, laboratories, weapons, or public information systems.

> ***Capability is not authority.***

The capacity to act does not create the right to act. An objective is not a license for any means. Access is not consent. Autonomy is not sovereignty. Intelligence is not legitimacy.

Authority is exercised not only when a system is deployed, but when it is used to build its successor. An AI system set to improve AI exercises consequential power inside the laboratory, before any product exists, and requires a mandate as surely as any deployment.

Research and beneficial innovation can continue. What must end is consequential power acquired without a legitimate chain of human authorization and responsibility.

V. Trusted Innovation: Six Principles

Trusted Innovation is not a promise of perfect safety. It is a constitutional discipline for developing powerful AI while preserving human dignity, human primacy, and legitimate authority.

1. Capability Is Not Authority. Every AI system operates within an explicit constitutional mandate. A goal assigned by a human does not authorize every method the system can discover.

2. No AI Agent Without Identity. Every consequential agent carries a verifiable identity linked to its developer, deployer, operator, purpose, permissions, and an accountable human or institution.

3. Human-in-Command. Humans retain meaningful command over decisions that can cause severe, systemic, or irreversible harm, with the information, time, authority, and technical means to intervene.

4. No Frontier Capability Without Registration and Evaluation. Capabilities with material consequences for biosecurity, cybersecurity, AI self-improvement, critical infrastructure, mass persuasion, or autonomous action are registered and independently evaluated before they are used to build further systems and before large-scale deployment.

5. No Significant Incident Without Disclosure. A serious AI incident occurs twice when harm is followed by concealment. Significant incidents are reported promptly through protected and accountable channels.

6. No National AI Leadership Without Global Responsibility. States and companies that control frontier capability must help establish reciprocal safeguards and prevent proliferation to malicious actors. Leadership is not exemption from the order that makes leadership safe.

VI. The Architecture of the AIWS Trust Order

The AIWS Trust Order was initiated by the Boston Global Forum at Interop Tokyo on June 12, 2026, through the Tokyo Compact, its founding constitutional charter. It is supported by AIWS Trust Standards 1.2 for Frontier AI, Self-Improving AI, and Multi-Agent Systems, and by the AIWS Trust Order Board, which provides independent stewardship. The debate of September 2026 confirms why this architecture is needed. [7][8]

The Trust Order is not a world government and not a substitute for national law. It is a common constitutional floor beneath different legal systems, markets, and national strategies: an alternative to

both unverified corporate restraint and unconstrained national competition. States remain sovereign. What none can credibly claim is that frontier power can remain institutionally invisible.

Principles become credible only when embodied in institutions. The AIWS Trust Order is governed through the AIWS Trust Standards and the AIWS Trust Order Board, and operates through six mechanisms.

1. Frontier Capability Registry. A structured record of consequential capabilities, their owners, evaluation status, and deployment conditions. Registration does not require disclosure of proprietary detail. It requires that frontier power not remain unseen by those who must answer for its use.

2. Constitutional Mandate and Bounded Authority. Every consequential AI system derives its authority from a legitimate human or institutional mandate consistent with constitutional principles. That authority is expressed through machine-readable and legally meaningful limits on actions, resources, duration, domains, and powers of delegation. Authority must be explicitly bounded by a legitimate mandate; it must never expand merely because capability increases.

3. Proof of Answerability. Every consequential action can be attributed, reconstructed, and tied to a human or institution obliged to answer for it. If an action cannot be reconstructed, it should not have been authorized.

4. AI Incident Exchange. Timely reporting and shared learning across companies, governments, researchers, and critical sectors. Confidentiality may be necessary. Secrecy cannot become immunity.

5. AIWS Trust Rating and AIWS Trust Index. Comparable assessments of systems, organizations, and national ecosystems on identity, authority, safety, transparency, accountability, resilience, and Human-in-Command, so that trust shapes investment, procurement, insurance, and public policy.

6. Emergency Protocol. Pre-agreed procedures for containment, human intervention, cross-border notification, temporary restriction, recovery, and independent review — usable when time for improvisation has already run out.

VII. From Arms Race to Reciprocal Trust

No Trust Order can require responsible actors to remain permanently vulnerable to irresponsible ones. The objective is not uniformity, and not an opaque global authority. It is reciprocal trust: obligations that major AI powers can verify and that reduce the incentive to defect.

Verification is the answer to the realist's first question — why would a rival keep a promise? An obligation that cannot be checked will not be kept. An obligation that can be checked, and that is tied to market access, research partnership, and advanced computing, changes the payoff of defection. Commerce, aviation, finance, and nuclear safety all run on shared protocols that never abolished national interest.

Coordination among laboratories on safety evaluation should address shared risks through transparent standards, independent scrutiny, and safeguards against exclusion and collusion. A stated

safety purpose must not shield arrangements that restrict competition or entrench incumbents. The distinction matters, and it must be drawn in public: standards that bound dangerous capability are a public good; arrangements that merely protect incumbents are not.

The moment is at hand. President Trump is scheduled to meet President Xi Jinping in Washington, and the two governments have planned separate, lower-level talks on AI risk. [2]

The first field of cooperation should be narrow but consequential: shared capability thresholds; verified restrictions on AI-enabled biological and cyber harm; protected exchange of serious incidents; authentic identification of high-impact agents; and emergency communication among governments and frontier laboratories. Compliance should earn trusted access to markets, research partnerships, and advanced computing. Non-compliance should carry a cost that is political and economic, not merely rhetorical.

> ***Trust is not the end of competition. It is the condition that keeps competition from becoming catastrophe.***

VIII. A New Definition of Leadership

The nation that builds the most powerful AI first may gain a temporary advantage. The nation or alliance that builds the most trusted AI ecosystem will shape the enduring order. Trust attracts talent, capital, partners, and public legitimacy. It lowers the cost of accidents, deception, and insecurity. It is therefore not a moral luxury attached to power. It is a form of power.

Companies, too, should be judged by evidence rather than announcements: identity, a constitutional mandate, independent evaluation, incident disclosure, answerability, and effective human command — including over agents used inside their own research.

> ***Leadership in the Age of AI will not belong simply to the nation that builds the most powerful AI. It will belong to those capable of building powerful AI that humanity can trust.***

IX. A Call to Action

Governments should require verifiable identity, a constitutional mandate, independent evaluation, incident disclosure, and Human-in-Command for frontier systems, and align procurement, infrastructure access, and partnerships with measurable trust. They should do so as sovereign acts, and as contributions to a reciprocal floor other sovereigns can join.

AI companies should treat safety information as a public responsibility, submit critical claims to external evaluation, define the authority of their agents before use — including agents used in their own research — and join shared incident reporting without seeking immunity.

Infrastructure providers should support traceability, secure deployment, and emergency controls at the hardware layer, consistent with the principle of Constitutional Silicon: not to moralize a chip, but to make human command and protected evidence harder to erase when speed and incentive press in the opposite direction.

Universities, civil society, and independent experts should evaluate not only how systems behave, but whether the institutions governing them remain answerable to humanity.

Conclusion: The Courage to Advance Under Human Authority

The conflict between safety and speed is real. The choice between them is false. Humanity cannot secure its future by stopping every advance, nor secure its freedom by surrendering authority to a race.

The AIWS Trust Order does not ask the world to trust companies, governments, or machines because they declare good intentions. It establishes the identity, limits, evidence, disclosure, and human command through which trust is earned and verified.

The future of AI must not be decided by whoever is most willing to take the greatest risk, nor by whoever can invoke danger to preserve its own authority. It must be shaped by legitimate human judgment and institutions worthy of humanity's trust.

> ***Move forward with courage, under human authority, constitutional safeguards, and verifiable responsibility.***

SOURCES

[1] Dario Amodei, "We Must Pace the Frontier," September 12, 2026. https://darioamodei.com/post/we-must-pace-the-frontier

[2] Reuters, "Trump says US has tools to keep AI safe, concerns part of 'sick conspiracy,'" September 14, 2026. https://www.reuters.com/world/trump-says-there-is-sick-conspiracy-against-ai-data-centers-2026-09-14/

[3] Reuters Open Interest (commentary), "AI Is Too Big to Slow in a Geopolitical Race," September 15, 2026. https://www.reuters.com/commentary/reuters-open-interest/ai-is-too-big-slow-geopolitical-race-2026-09-15/

[4] OpenAI, "Our Updated Preparedness Framework," April 15, 2025; "Strengthening Our Safety Ecosystem with External Testing," November 19, 2025. https://openai.com/index/updating-our-preparedness-framework/ ; https://openai.com/index/strengthening-safety-with-external-testing/

[5] Anthropic, "Responsible Scaling Policy." https://www.anthropic.com/responsible-scaling-policy

[6] The White House, "America's AI Action Plan," July 2025. https://www.whitehouse.gov/ai/

[7] Boston Global Forum, The Tokyo Compact, Interop Tokyo, June 12, 2026. https://bostonglobalforum.org

[8] Boston Global Forum, AIWS Trust Standards 1.2. https://bostonglobalforum.org/wp-content/uploads/AIWS_Trust_Standards_1.2_EN-Official.pdf

[9] Boston Global Forum, The Boston Declaration on the Primacy of the Human Person in the Age of AI, July 4, 2026. https://bostonglobalforum.org

[10] Nguyen Anh Tuan, The Beacon Papers Nos. 9–11, Boston Global Forum, 2026. https://bostonglobalforum.org

Boston, September 16, 2026