THE BEACON PAPERS · No. 20
Foundational papers for humanity in the AI Age — from Boston

AI SAFETY CANNOT RUN ON PROMISES

A Constitution, an Observatory, and a Movement for AI Safety

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

Beacon Hill, Boston • November 15, 2026

A principle tells power where to stop. Only an institution can show whether it stopped.

I. Three Dates in November

On November 10, 2026, at Loeb House, Harvard University, the Boston Global Forum Conference "New Civilization in the Age of Artificial Intelligence: Human–AI Co-Creation" examined three pieces of work: the draft Constitution for Humanity in the Age of AI, the first Special Report of the AIWS Trust Observatory, and The Human–AI Co-Creation Atlas. Two days later, at the Massachusetts State House, President Egils Levits joined the Board of Thinkers of the Boston Global Forum, and we agreed on the next steps. On November 23, the Secretary-General of the United Nations, António Guterres, and the Prime Minister of Barbados, Mia Mottley, will host the Bridgetown Special Summit on AI Safety. There the Boston Global Forum will present its Special Report, *A Constitution, an Observatory, a Movement*, in Roundtable 2, "AI for Humanity."[1]

This paper sets out the argument that joins these three dates. For a decade, the world has written principles for artificial intelligence. Many of them are good, and some have become law. What the world has not yet built is a way for the public to check whether they are kept. A constitution sets the limits. An observatory shows whether the limits hold. A movement gives people the standing to insist that they do. None of the three is sufficient alone. Together, they turn safety from a promise into a practice.

II. The Decade of Principles

The record deserves credit before it receives criticism. Since 2017, when AIWS began and the Asilomar AI Principles were published, governments, scientists and companies have produced an extensive body of texts.[4] The OECD adopted its AI Principles in 2019. UNESCO's member states adopted the Recommendation on the Ethics of Artificial Intelligence in November 2021.[5] At Bletchley Park in November 2023, twenty-eight countries and the European Union signed a declaration on the safety of frontier AI.[6] The G7 issued the Hiroshima Process International Code of Conduct the same autumn.[7] In Seoul, in May 2024, sixteen companies made the Frontier AI Safety Commitments and promised to publish safety frameworks.[8] On August 26, 2025, the General Assembly of the United Nations established the Independent International Scientific Panel on AI and the Global Dialogue on AI Governance.[9] The Boston Global Forum contributed its own texts to this record, among them the Social Contract for the AI Age in 2020.[15]

Some of these principles have since become law. The European Union's AI Act is binding, and its AI Office oversees general-purpose models. California's SB 53 requires the largest frontier developers to publish their safety frameworks and to report critical safety incidents to the state.[10] A network of national AI safety and security institutes now tests frontier models.

These texts and laws did real work. They gave the world a shared vocabulary, placed safety on the agenda of heads of government, and gave the United Nations a scientific panel of its own. Yet three features recur across the record. Most commitments are voluntary. Compliance is measured mainly by what developers report about themselves. And where the law requires more, what regulators learn is largely reported in confidence: the public is rarely shown it, and no one independent of the developers checks it in public.

The consequence is now visible. The first Special Report of the AIWS Trust Observatory, "The State of AI Transparency at the Frontier," discussed at Loeb House on November 10, finds that AI power is growing while what society can see of it is shrinking.[2] Stanford's average transparency score for leading developers fell from 58 out of 100 in 2024 to 40 in 2025, although the indicators and the companies assessed changed between editions.[11] In the Future of Life Institute's AI Safety Index for Summer 2026, the highest grade any company received on existential safety was D+.[12] An August 2026 assessment by Guidelight AI Standards found few published plans for containing a model that evades control.[13] The White House Accord on Super Intelligence of September 29, 2026, provides for an independent auditor whose findings go to a board committee, but does not require those findings to be made public.[14]

The problem is not that the principles were wrong. It is that they were left unchecked. A principle that no one can check becomes, over time, a statement of intent. When the next model is more capable than the last, a statement of intent is not enough.

III. What a Principle Cannot Do

A principle cannot do three things.

It cannot bind. A principle says what should be done. It does not say who may decide, within what limits, and with what remedy when the limit is crossed. Beacon Papers No. 11 held that capability is not

authority.[19] Authority needs a mandate, and a mandate needs a written order of limits that does not depend on the goodwill of whoever holds the most capable systems.

It cannot see. A principle has no eyes. Whether a model was evaluated before release, whether an incident was disclosed, whether a containment plan exists: these are facts, and facts must be gathered by someone who does not depend on the party being observed. Beacon Papers No. 4 asked who may verify.[16] Beacon Papers No. 13 answered that AI power must not operate unseen.[21]

It cannot defend itself. A principle has no constituency. It survives when people know it, use it, and expect others to keep it. Where people become passive users of systems they do not understand, no principle will long survive the pressure of competition.

Each of these limits calls for an institution. Each institution also carries a failure of its own when it stands alone. A constitution without a means of observation becomes a text. An observatory without limits to measure against becomes a commentary. A movement without either becomes a mood. And there is one failure the whole architecture must avoid: the answer to unchecked power cannot be a new concentration of power in a few companies or in a single central authority. The institutions proposed here are designed to distribute the work of safety, not to collect it. That rule binds their authors first. The AIWS Trust Observatory is one observer among the many the world needs, and it invites others to observe the same commitments by their own methods. Each institution described here submits to the tests it asks of others.

There is a deeper reason why principles alone cannot keep AI safe. The world discusses safety at two levels. The safety of models is the work of laboratories. The safety of systems is the work of regulators. A third level belongs to no one: humanity's capacity to know what AI power is doing, to judge it, and to correct it. No laboratory can hold this level, because no one can verify itself. No single state can hold it, because the risks cross every border. We call it **civilizational safety**. A safe model is not yet a safe civilization. The three institutions that follow are its three parts: a constitution that sets its limits, an observatory that supplies its knowledge, and a movement that keeps the human capacity to use both.

IV. A Constitution: Limits on Power

The Constitution for Humanity in the Age of AI addresses power in two forms: the power of machines acting without legitimate human authorization, and the power of people acting through machines without constitutional restraint. Its design is stated in two sentences: "No machine may escape human command. No human command may escape constitutional restraint."[17]

The first Draft for Consultation was issued on Beacon Hill, Boston, on July 8, 2026. Version 1.7, the Constitutional Edition examined at Loeb House, contains 54 Articles in seven Parts and opens with the words "We, the People of Humanity."[3] Several of its Articles bear directly on safety. Article 5 establishes Human-in-Command: no AI system shall acquire unchecked authority over human beings, and command cannot be delegated where a decision is irreversible or affects liberty, rights or life at scale. Article 25 sets out the AIWS Trust Order: no frontier capability shall be deployed without registration and independent evaluation, and no significant AI incident shall go undisclosed.[20] Articles 26 and 28 require that there never be a single verifier, and that every consequential AI agent have a Designated Human Answerer.[18] Article 31 gives independent civil-society observation a constitutional place without making it a regulator. Article 52 keeps Human Dignity, Human Agency, mental

sovereignty and Human-in-Command inviolable, while requiring review of the text every two years against independent findings.

A fair question is how a text that no state has yet acceded to can bind anyone. President Levits answered it at Loeb House from the experience of Latvia, where the constitutional order kept its legitimacy through fifty years of occupation. Constitutional fidelity, he argued, can come before formal constitutional force. A "soft law" instrument gains real influence when it is legitimate, credible and useful, and he found that this text meets all three requirements. What distinguishes this Constitution from the decade of principles, then, is not yet the force of law. It is that every limit it sets is joined to a means of checking whether the limit is kept, and to a public able to insist that it is.

Legitimacy is earned by authorship. That is why the text is open to the Global Consultation until January 31, 2027, why the accession book has been open since September 23, 2026, and why the Working Group formed on November 12 will put specific questions to people who have led nations, institutions and fields of science. The Constitution is to be officially proclaimed at Loeb House, Harvard University, on April 27, 2027, marking ten years of AIWS.

V. An Observatory: Seeing Whether the Limits Hold

The AIWS Trust Observatory is an independent civil-society institution established by the Boston Global Forum and AIWS to observe consequential AI power, warn early when danger emerges, and call those responsible to act. It puts into practice independent monitoring, the first mechanism of the Tokyo Compact proclaimed on June 12, 2026.[24]

Its credibility rests on rules that bind the Observatory itself. It draws conclusions only from what can be verified. It accepts no funding from any party it observes, and no donor may exceed 20 percent of its annual budget. It approves nothing and prohibits nothing. It does not govern AI companies or replace national regulators. It keeps the public record, so that those who hold authority, and the public they answer to, can act on evidence.

Its first Special Report proposes a Transparency Floor: a minimum public disclosure for every frontier release. The Floor asks for a technical report on the day of release; a summary of training data sources and compute; the results of dangerous-capability evaluations; energy and water use, the burden Beacon Papers No. 19 traced to the communities that host AI;[23] whether a containment plan exists and who can trigger it; whether the model was used inside the company before release; and what the developer does not yet understand about the system. The last item matters most. A developer that cannot say what it does not know has not yet earned the trust it asks for.

The Observatory's instruments follow from this standard. A Commitment Register sets the public promises of AI organizations beside their conduct. A Blind Spot Map states what society cannot yet see, and why. Four alert levels distinguish matters to watch from dangers that call for action. The first observation stations begin pilot operation in December 2026. The first Quarterly State of Frontier AI is due at the end of January 2027, and on April 27, 2027, the Observatory is to be presented at Harvard University as the monitoring mechanism of the Constitution.

On November 10, Professor Cynthia Dwork received the World Leader in AIWS Award 2026. The citation recalled that, with her collaborators, she invented differential privacy, and said of that work:

"You turned privacy from a promise into a proof." The Observatory asks the same of AI safety. A safety claim should be stated in a form that can be checked, and then it should be checked.

VI. A Movement: People Who Keep Power Answerable

Institutions are kept by people. The AIWS Human–AI Co-Creation Movement, launched on September 23, 2026, at Cultural Diversity Day in the Massachusetts State House, asks each participant to sign three commitments: "I remain the author of what I create. I bring my own culture into what I create with AI. I use AI to become more capable, never less."

The third commitment is a safety commitment. A population that loses the capacity for independent judgment cannot hold any system to account, however good its laws. The Human Advancement Test of Beacon Papers No. 15 measures this capacity in six dimensions: Deeper Understanding, Sounder Judgment, Higher Creativity, Fuller Command, Elevating the AI Assistant, and Greater Love for People and Nature.[22] Article 40 of the draft Constitution makes the claim falsifiable: where people become less capable of independent thought, the design must change and the finding must be published.

The Movement also needs a record. The Human–AI Co-Creation Atlas, announced by Professor Thomas Patterson at Loeb House, will document what humanity creates with AI and what people become in the process, across eight fields of achievement. Every record carries documented sources and a dated history of revisions. Fame, affiliation, scale and sponsorship are never sufficient grounds for inclusion, and the work of the Boston Global Forum and AIWS is held to the same standard as every other entry. The Atlas opens on January 8, 2027, at atlas.aiws.net, with eight Founding AIWS Civilization Cities, beginning with Boston, Tokyo, Rome and Hà Nội, as starting points for an open global network.

This is how safety becomes shared ownership. A citizen who signs the three commitments, a teacher who submits a record to the Atlas, a city that reports what its people have done: each holds a part of the work that would otherwise rest with a handful of laboratories and regulators. A movement of this kind does not replace law or observation. It gives both a public that expects them to work.

VII. One Architecture

The three institutions are designed to hold one another. The Constitution gives the Observatory its standard: Article 25 says what must be registered and disclosed, and Article 31 secures the Observatory's place without giving it power over those it observes. The Observatory gives the Constitution its evidence: under Article 52, the text is reviewed every two years against independent findings, so that it can learn from what actually happens. The Movement gives both their constituency, and the Atlas gives the Movement a record that can be examined. The Conference report summarized the relationship in three sentences: "A constitution restrains power. An observatory makes power visible. A movement awakens people."

The steps agreed at the Massachusetts State House on November 12 are the first working form of this architecture. A Working Group on the Constitution, whose scope includes the AIWS Trust Observatory, begins with President Levits and Professor Lawrence Lessig. Focused questions for

discussion will be put to the recipients of the Forum's two awards, the AI Pioneers community, leaders at the United Nations, and the leaders of the four Cornerstone Countries: the United States, Japan, Italy and Vietnam. A public exchange of signed letters, *Commonwealth of Minds: Letters for the AI Age*, which President Levits and I co-chair with Professor Lessig, opens the work to signed, public reflection, in the tradition of scholars who once corresponded across borders. With the consent of the writers, letters that address a specific Article will be entered into the Global Consultation. None of these steps grants anyone the power to rewrite the text alone. Changes to the Constitution pass through the Consultation, and the conclusions of the AIWS Trust Observatory follow its published method.

The architecture is guided by the four values of AIWS Civilization: Love, Creativity, Nobility and Wisdom. In matters of safety they are practical. Love asks who will bear the harm if the system fails. Creativity looks for the safeguard that does not yet exist. Nobility keeps a promise when breaking it would be cheaper. Wisdom knows which decisions must never leave human hands.

Institutions can restrain power and make it visible. They cannot, by themselves, give people a reason to rise. The deepest layer of civilizational safety is a shared spirit: that every advance of artificial intelligence must become an occasion for humanity to rise together, in understanding, in love for people and nature, and in the resolve to keep the decisions that matter in human hands. Built on these four values, this Spirit will be set out by the Boston Global Forum in January 2027.

VIII. What We Bring to Bridgetown

Bridgetown has already given its name to one proposal for reforming international institutions: the Bridgetown Initiative on development and climate finance.[25] The choice of Barbados to host a summit on AI safety carries a reminder. The countries with the least say over frontier AI are often those most exposed to its failures, and the least able to verify what they are told about it.

The oldest lighthouse in Barbados stands at South Point. Its cast-iron tower was shown at the Great Exhibition in London in 1851 and raised the next year at the island's southern tip.[26] A light first shown in London guides the ships of every nation along the coast of Barbados. Verification of AI must work the same way. A small island state cannot build a laboratory to evaluate frontier models. It can share a light. The public findings of the AIWS Trust Observatory will be open to every nation without charge, and small island and developing states will have first access to its briefings.

Verification also answers the fear that hangs over every discussion of AI safety: the race among great powers. The history of nuclear arms control shows that adversaries did not need to trust one another to reach agreement. They needed a method of checking that both could accept. Rivals cannot share their secrets. They can share a method of checking. Verifiability is not an obstacle to cooperation on AI. It is its precondition.

The Boston Global Forum will offer the Summit seven proposals, drawn from the results of November 10. They are proposals of the Forum alone and imply no commitment by the United Nations, by any government, or by any person named in this paper.

- **Make verifiability a pillar of AI safety**, alongside regulation and technical standards, and consider a common minimum of public disclosure for frontier releases along the lines of the Transparency Floor.

- **Let the public learn what regulators learn.** Where serious incidents and risk assessments are reported in confidence, a public summary should follow within a fixed time.
- **Measure promises against conduct**, by publishing voluntary safety commitments in a form that independent observers can check over time. The Forum offers the Summit and interested UN bodies the public outputs of the AIWS Trust Observatory, beginning with its first Quarterly State of Frontier AI.
- **Take part in shaping the Constitution** by sending comments on any Article to the Global Consultation before January 31, 2027, so that the text carries the voice of every region, including developing and small island states.
- **Measure AI for Humanity by human advancement**, asking of every major program for inclusion, access and capacity whether an advance in artificial intelligence also advances human intelligence.
- **Open the record of human achievement to every nation** through The Human–AI Co-Creation Atlas and the AIWS Human–AI Co-Creation Movement.
- **Continue the dialogue at Harvard University on April 27, 2027**, at the official proclamation of the Constitution.

IX. A Proposed Article

For the Global Consultation on the Constitution for Humanity in the Age of AI, this paper proposes one article, to complement Articles 25 and 31:

Article: Commitments That Can Be Checked

1. Whoever develops or deploys frontier artificial intelligence and makes a public commitment about its safety shall state that commitment in a form that independent observers can check over time, and shall report each year on what was promised against what was done.

2. No frontier system shall be released without a minimum of public disclosure, including its evaluation for dangerous capabilities, the existence of a containment plan and the person who can trigger it, and what its developer does not yet understand about it.

3. Where serious incidents or risk assessments are reported in confidence to public authorities, a public summary shall follow within a fixed time, without detail that would enable harm.

4. Independent observation of these commitments shall be plural. No observer shall be funded by those it observes, and no observer shall hold the power to license or prohibit.

5. A commitment that cannot be checked shall not be counted as a safeguard.

Beacon Principle No. 20

A principle says where power must stop.
A constitution makes the limit binding.
An observatory shows whether it holds.
A movement gives people the standing to insist.

The decade of principles gave the world the right words. The years ahead will be judged by whether those words acquired institutions able to keep them, and people willing to hold those institutions to their work. The Boston Global Forum carries three such institutions from Harvard University to Bridgetown, and invites every nation to help build them. A lighthouse does not ask which flag a ship flies.

Humanity leads. AI elevates. Together, we create.

Love · Creativity · Nobility · Wisdom

SOURCES

1. Boston Global Forum, A Constitution, an Observatory, a Movement: Three Foundations for AI Safety and AI for Humanity. Results of the Boston Global Forum Conference at Harvard University, November 10, 2026. Special Report to the Bridgetown Special Summit on AI Safety, November 2026
2. AIWS Trust Observatory, The State of AI Transparency at the Frontier, First Special Report, November 2026
3. Boston Global Forum and AIWS, Constitution for Humanity in the Age of AI, Version 1.7, Constitutional Edition (draft for Global Consultation)
4. Future of Life Institute, Asilomar AI Principles, 2017
5. OECD, Recommendation of the Council on Artificial Intelligence, May 2019; UNESCO, Recommendation on the Ethics of Artificial Intelligence, November 23, 2021
6. The Bletchley Declaration by Countries Attending the AI Safety Summit, November 1, 2023
7. G7, Hiroshima Process International Code of Conduct for Organizations Developing Advanced AI Systems, October 30, 2023
8. Frontier AI Safety Commitments, AI Seoul Summit, May 21, 2024
9. United Nations General Assembly, Resolution A/RES/79/325, Independent International Scientific Panel on Artificial Intelligence and Global Dialogue on Artificial Intelligence Governance, August 26, 2025
10. Regulation (EU) 2024/1689, the Artificial Intelligence Act; California SB 53, Transparency in Frontier Artificial Intelligence Act, 2025
11. Stanford HAI, "Transparency in AI Is on the Decline," December 2025 (Foundation Model Transparency Index)
12. Future of Life Institute, AI Safety Index, Summer 2026
13. Guidelight AI Standards, August 2026 assessment of containment plans, as reported by Unite.AI
14. "How does Trump's White House AI accord work?," Al Jazeera, September 30, 2026
15. Boston Global Forum and partners, Social Contract for the AI Age, September 9, 2020
16. The Beacon Papers No. 4, Who May Verify?
17. The Beacon Papers No. 5, Who May Command?
18. The Beacon Papers No. 10, No AI Agent Without Identity
19. The Beacon Papers No. 11, No Power Without a Mandate
20. The Beacon Papers No. 12, The AI Race Needs a Trust Order
21. The Beacon Papers No. 13, AI Power Must Not Operate Unseen
22. The Beacon Papers No. 15, The Human Advancement Test
23. The Beacon Papers No. 19, The Cloud Lost 44 Elections
24. The Founding Framework of the AIWS Trust Observatory, Boston Global Forum, September 27, 2026; Tokyo Compact, June 12, 2026
25. "How the Bridgetown Initiative envisions global financial system reform," Devex
26. South Point Lighthouse, Barbados

Beacon Hill, Boston • November 15, 2026