



THE BEACON PAPERS · No. 4
Foundational papers for humanity in the AI Age — from Boston

WHO MAY VERIFY?

The Missing Institution of the AI Age

Nguyen Anh Tuan

Co-Founder, Co-Chair, and CEO, Boston Global Forum · Founder and Chief Architect, AIWS

BOSTON · JULY 27, 2026

In the space of a single month, the world's most powerful institutions arrived — independently, and almost simultaneously — at the same conclusion: artificial intelligence must be verified by someone other than the people who build it. Illinois became the first American state to require large frontier developers to submit to annual independent third-party audits, enacted in July and beginning in 2028. A bipartisan draft in the United States House proposed transparency mandates and third-party audits at the federal level. OpenAI advanced a proposal for mandatory pre-release evaluation and annual outside audit; the leadership of Google and Google DeepMind advanced another, calling for an industry safety body under federal supervision. In Europe, obligations on general-purpose models have applied for a year, and on the second of August the European Commission acquires the power to enforce them. Rarely has an idea traveled so far so fast.

And yet not one of these proposals answers the question upon which every one of them depends. Verified by whom?

The Question Beneath the Consensus

Examine the proposals together and a pattern appears that none of them acknowledges. In each, the verifier depends on what it must judge. Laboratories propose audits designed by laboratories. Governments propose agencies accountable to governments. A cooperation organization headquartered in a single capital will, if it issues standards at all, certify its own members against them. Every arrangement is defensible on its own terms. Taken together they describe a world in which everyone is checked, and no one is checked from outside.

There is a second omission, and it is the larger of the two. Every proposal now under discussion concerns the verification of models — of systems, weights, capabilities, evaluations. But the model is the smallest object in this field and the least of what requires verification. Four tiers of power are shaping the intelligent century, and each of them makes claims that no one is presently positioned to test.

The models make claims about what they can and cannot do. The corporations that build them make claims about their own conduct: that their safety practices are real, their internal governance independent, their incidents reported rather than absorbed. The governments that regulate them make claims of their own: that their laws protect citizens rather than watch them, that their oversight bodies are genuinely independent of the ministries and industries that sustain them. And the certifying bodies now being proposed will make the largest claim of all — that their judgments may be relied upon by everyone else.

Verify the model alone and leave those three tiers unexamined, and we shall have inspected the instrument while trusting, on faith, every hand that holds it.

This is not hypocrisy. It is the ordinary gravity of institutions, which build what they can reach. But gravity is not an argument, and the consensus now forming will harden into architecture within a year. What is built in this window will be difficult to unbuild.

What Humanity Has Already Paid to Learn

We are not the first generation to face this problem. We are merely the first to face it at the speed of software.

Consider the audit of corporations. For most of the twentieth century, the accounts of great firms were examined by auditors selected, paid, and retained by the firms they examined. The arrangement worked until it spectacularly did not, and the reckoning at the turn of this century destroyed not only a celebrated corporation but one of the largest audit houses on earth. The remedy that followed was not more auditing. It was the creation of an oversight body that the audited did not control.

Consider the rating of credit. In the years before 2008, the instruments at the center of the world financial system were rated by firms compensated by the issuers of those instruments. No individual rater need have been corrupt for the outcome to be catastrophic. The structure supplied the result.

Now consider the one great verification regime that has worked comparatively well: the international inspection of nuclear materials. It succeeded, where the others failed, for reasons that can be stated precisely. Its inspectors belonged to no inspected state. Its presence was continuous rather than announced. Its methods were published, while its findings were reported to a body standing above any single government.

The lesson across all three is consistent, and it is structural rather than moral. Verification fails when the verifier's existence depends upon the verified. It succeeds when it does not. Nothing

about artificial intelligence exempts it from this law; its speed and opacity only raise the penalty for ignoring it.

The Regress, and How It Ends

State the principle plainly and an ancient objection arrives at once. If the corporations must be verified, and the governments, and the certifying bodies, then who verifies the verifiers? And who verifies those? The question is not a debating trick. It is the oldest problem in the theory of authority, and any proposal unable to answer it deserves to fail.

The answer is not another tier. A supreme verifier placed above all others would merely relocate the difficulty, and would concentrate in one unaccountable body precisely the kind of power that a constitution for humanity exists to restrain. The regress cannot be ended by climbing.

It is ended by closing the circle. Three properties, held together, terminate the chain.

Plurality. There must never be a single verifier. Where several independent bodies assess the same subject against published criteria, their agreement carries weight and their disagreement is itself evidence. Capture then becomes not merely wrong but visible, for a captured verifier must diverge from its peers in ways that can be seen and asked about.

Publicity. A verifier that publishes its methods has submitted to examination by everyone competent to examine them — rival institutions, researchers, journalists, and the assessed themselves. Publication converts a private judgment into a public claim, and a public claim can be tested, replicated, and refuted. Secrecy about method is the condition under which capture survives.

Contestability. Any party subject to a verdict must be able to challenge it, to require re-assessment, and to have that challenge heard by persons other than those who rendered the original judgment. A right of appeal is not administrative courtesy. It is what prevents verification from hardening into decree.

This is not a novel invention. It is how constitutional orders have always answered the question of who stands at the summit — by declining the question. The framers of 1787 did not discover a final arbiter. They divided authority, exposed each part to the scrutiny of the others, and made every exercise of power answerable somewhere. The AI Age requires that same architecture, rebuilt for a domain in which power moves at machine speed and crosses every border without asking.

Five Conditions on Any Verifier

The three properties above describe the system of verification as a whole. What follows describes what any single body must be in order to take part in it. Five conditions follow from that history and that logic. They are not a wish list. They are the minimum below which the word verification should not be used — and they bind the verifiers as strictly as anyone the verifiers assess.

Independence of legal person. The verifier must not be owned by, funded at the discretion of, or dissolvable by those it assesses. Fees may pay for the work of assessment; they may never purchase its outcome. Where that line is not visible from the outside, it does not exist. Interests that could bear upon a judgment must be declared before the judgment, not discovered after it.

Continuity rather than ceremony. An annual audit measures a system on the day it is measured. But these systems change weekly; capabilities emerge after deployment that were absent at release; behavior drifts under retraining and under pressure. Periodic certification of a continuously changing system is a photograph offered as a film. Verification worthy of the name must be instrumented and continuous.

Transparency of method. The yardstick must be public even where particular findings are confidential. A standard whose criteria are secret cannot be contested, improved, or trusted. Openness about how one measures is the discipline that makes a measurement mean something.

A route of challenge. Contestability binds the verifier as strictly as it binds anyone else. A body that renders judgments must provide a route by which those judgments may be challenged, and the challenge must be heard by persons other than those who rendered the original finding. A verifier that cannot be contested is not a verifier. It is an authority, and this paper has already declined to build one.

Legitimacy beyond any single jurisdiction. A national verifier can bind its own market. It cannot confer legitimacy upon a global technology. Frontier systems cross every border in the time it takes to read this sentence, and no verdict rendered inside one legal order can travel with them.

Why This Is a Constitutional Question

The preceding paper in this series held that legitimacy in the governance of artificial intelligence is earned rather than inherited, and demonstrated in practice rather than in proclamation. That claim carries an obligation, and it falls due here.

If leadership must be demonstrated, there must exist a means of demonstration not controlled by the one demonstrating. Otherwise demonstrated fidelity collapses quietly into self-report, and the open door of the Constitution for Humanity — membership earned by commitment and verified through trust standards — becomes an honor system among the powerful. A constitution cannot promise verification and leave the verifier unbuilt.

This is why independent verification belongs not in an administrative annex but among the load-bearing elements of any constitutional order for the AI Age. Rights that cannot be checked are aspirations. Standards that cannot be checked are advertisements. And trust that cannot be checked is not trust at all; it is hope, wearing the costume of assurance.

An Architecture, Not an Authority

What follows from this is not a proposal for one more center of power. It is a proposal for an ecosystem: independent, multi-tiered, and capable of examining itself.

The technical means already exist in outline. The AIWS Trust Infrastructure was built to supply what the four conditions require — standards developed in the open rather than behind closed doors; a rating methodology whose criteria are published rather than held as proprietary advantage; continuous instrumented monitoring in place of the annual inspection; and evidence that can be checked rather than assurances that must be believed.

The institutional means are the harder half. The AIWS Trust Order exists to provide them: a governing body drawn from many nations, so that no single government may direct a verdict; a common framework within which several verifiers may operate and check one another; and a standing to which the assessed may appeal, distinct from those who assess.

Neither is offered as the answer. Both are offered as a working demonstration that an answer is possible — and as an invitation to others to build alongside. A single verification infrastructure, however careful and however well-intentioned, would violate the first of the three properties named above. Plurality begins with a willingness to be one among several. We intend to be one among several.

We Submit Ourselves to This

A paper that demands verification of everyone and exempts its author is worth nothing. The standard proposed here binds the institution proposing it, and it should bind us first.

Accordingly: the criteria by which trust is assessed under this architecture are published rather than proprietary, and are open to challenge by anyone. The sources of funding for this work will be disclosed, so that any interest which might bear upon a judgment can be weighed by those who read it. External review of our methods is invited, including — especially — by institutions that disagree with us. And any party assessed under these standards may contest the result and require that the contest be heard by persons who did not render it.

We do not claim that these commitments make us trustworthy. That is precisely the claim no institution may make about itself, and the claim this paper exists to reject. We claim only that we have placed ourselves where the question can be asked, and answered by others.

The Instrument and the Claim

Every age of power has produced its claims, and each has had to produce, alongside them, the instrument by which claims are tested. The claims of the AI Age have arrived in abundance — from capitals, from laboratories, from newly founded organizations, each announcing itself as the trustworthy steward of the intelligent century. The instrument has not yet arrived.

Let it arrive with a principle stated plainly enough to be held against everyone, its authors not excepted.

No power over humanity should remain beyond verification — not the power of machines, not of governments, not of corporations, and not of those entrusted to verify them.

A beacon is trusted for one reason only: because its light can be checked against the chart. The light that cannot be checked is not a guide. It is merely a bright thing in the dark, asking to be believed.

Let us build the chart.

Boston Global Forum · AIWS · Boston, July 27, 2026