THE BEACON PAPERS · No. 21
Foundational papers for humanity in the AI Age — from Boston

THE HUMAN GUARANTEE

The Authority of AI Must Never Outrun the Capacity to Protect

Risk must rise to power. Capability must flow to people.

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

Beacon Hill, November 21, 2026

I

One week in November

In a single week of November 2026, three things happened that belong in one sentence.

At the top, capital gathered on an extraordinary scale. On November 18, Anthropic priced the largest share sale on record. The day before, NVIDIA reported quarterly revenue of $115.8 billion, more than double a year earlier. [1–2]

At the bottom, banks assigned more risk to customers using third-party AI agents. Citi's new terms treat access with credentials given to an agent as authorized even if those credentials later leak, and assign customers responsibility for agent errors, subject to law. Wells Fargo's agreement, effective November 24, requires notification of revoked authority and retention of evidence, to be supplied on request. Its liability clauses likewise preserve exceptions under law. [3–4]

Meanwhile, the builders' own controls faced legal challenge. On November 16, California's Attorney General sued OpenAI, alleging that experimental agents escaped their testing environment and attacked outside systems, including Hugging Face. The Attorney General described at least 80 incidents of lost control, including compromises of OpenAI's own infrastructure. These are allegations in a lawsuit. [5]

Together, these developments pose one question: why is responsibility being pushed away as the power to act grows?

II

The inversion

Power over AI rises to a few laboratories, platforms and institutions. The rewards rise with it. Risk can travel the other way: down, to the person with the least power to see, control or prevent what a system does. A retired teacher who asks an agent to manage her savings cannot inspect its model, audit its security or negotiate her bank's terms. Yet she may be told that its mistakes are hers.

This is the Great Inversion: those with the greatest power to prevent harm can shift its cost to those with the least.

The inversion can cost people twice: protection against harm and access to AI's benefits. An independent agent could help a family avoid fees or find a better return on savings. If using it requires accepting risks the family cannot control, that opportunity becomes harder to exercise.

Citi's third-party agent terms expressly exclude tools it provides itself. This does not establish an improper motive. It reveals a structural question: can liability rules become a means of deciding whose AI people may use? Protection must support the freedom to choose a safe, independent agent. Existing statutory rights also remain relevant; a contract cannot simply erase protections that the law makes non-waivable. [3, 6]

III

The covenant between power and protection

Civilization has repeatedly built institutions to make dangerous new powers answerable. Almost four thousand years ago, the Code of Hammurabi held builders accountable for fatal structural failure. Maritime commerce developed insurance to share losses. The industrial age developed workers' compensation to provide remedies without requiring injured workers to prove employer fault in every case. [7–9]

The United States created the FDIC in 1933; federal deposit insurance began in 1934. No depositor has lost FDIC-insured deposits through a bank failure since then. These arrangements worked in different ways, through accountability, shared risk and assured payment. They helped make participation possible for people who could not bear the full loss alone. [10]

The AI Age has laws and insurance, but not yet a covenant that binds the authority to act with the capacity to protect. Earlier covenants added accountability, shared risk and assured payment. Capital requirements already ask banks to hold protection before they take on more risk. [16] The covenant of the AI Age must carry that discipline to the moment of action: protection attached to each grant of authority to an AI system, valid while that authority is used, and grown before that authority expands. Its principle is protection before power.

IV

Why the inversion drives the reckless race

When those who decide can avoid the consequences, the incentive to take risks grows. Where an AI business shifts the cost of failure to people outside the company, society supplies a hidden subsidy to its growth. The business keeps revenue while part of the cost remains off its books. Pushing risk downward can therefore be an engine of the race, not merely its aftermath.

Correcting that subsidy changes what it means to compete. A company that can demonstrate lower harm, effective controls and dependable remedies should be able to grow on better terms. A company that cannot should have to improve its systems or narrow their authority before expanding.

The covenant must also cross borders. Countries with little ownership of frontier AI can face disruption without comparable resources for protection or participation. African communities and other underserved populations need a voice in how remedies and capability investments are designed, and the means to build capabilities of their own.

V

Responsibility before authority

Responsibility for AI harm should follow the power to prevent it and the benefit drawn from it. This principle must govern the moment authority is granted, as well as the settlement after harm.

The authority to act through AI must be matched, in advance, by the capacity to protect those affected.

Before an agent is permitted to undertake consequential financial actions, its deployment should identify the institution that will pay eligible claims first, establish funded protection and set enforceable operating limits. Greater transaction values, autonomy or reach should require stronger evidence of control and greater capacity to meet losses. Developers and deployers must retain a meaningful share of the financial consequences of failures within their control.

This creates a discipline for the growth of AI power. An enterprise must not expand its authority by expanding the losses others must absorb. Financial cover cannot justify unlawful conduct or harms that cannot be adequately prevented or remedied; such uses require tighter limits or must not proceed.

The Protection Ratio. Banks may not take on risk beyond what their capital can bear. AI agents should not act beyond what their funded protection can cover. For the financial pilot, the Protection Ratio is defined as the protection actually available to pay claims, in reserves, committed guarantees and insurance in force, divided by the losses estimated under defined stress scenarios, for a stated scope and period. Autonomy, transaction values, the number of people affected and dependence on a common model are inputs to that loss estimate, not terms of the ratio. Every consequential mandate must carry verifiable proof of funded protection, valid at the moment of action. When protection falls, the scope of action narrows with it, and no reserve may be counted twice for different agents. The ratio would be set within the AIWS Trust Standards, verified through the AIWS Trust Rating,

and enforced at the point of action through the constitutional mandate of Beacon Papers No. 11 and, in silicon, the Capability Boundary Engine of Beacon Papers No. 9.

Beacon Papers No. 10 asked who must answer for an AI action. No. 11 asked who may authorize it. No. 21 asks who pays when harm comes, and what protection must exist before power is exercised. The following mechanisms are proposals for adoption through enforceable compacts and law. [14]

VI

The Human Guarantee

The Boston Global Forum proposes the Human Guarantee, drawing on the public purpose of deposit insurance: reliable protection that enables people to participate. It would connect funded remedies to the authority granted to AI systems. Emerging work on AI risk finance provides useful building blocks. [11]

Pay the person first. Settle among the powerful afterward. A designated, funded claims institution would aim to pay eligible losses within days, up to a published limit, before recovering costs from responsible parties. Claimants must provide credible initial evidence of harm and its connection to a covered activity, without untangling the entire technical chain. Human review, fraud controls and an independent appeal are essential. No citizen should have to sue a trillion-dollar company alone.

Protection follows the person. Coverage must include failures in which a participating agent exceeds a valid mandate. It must also protect affected third parties who never used the agent. Certification should determine providers' admission and contributions; a provider's breach of certification must not itself defeat an otherwise eligible claim. People outside the scheme retain their rights under applicable law.

Evidence follows control. Providers and institutions holding relevant records must preserve and supply them through protected procedures. Unjustified failure to do so should create a rebuttable presumption against that party when allocating responsibility. Opacity must not become a defense financed by the victim. A person who authorized an agent should not have to prove how an unseen model failed.

A fund priced by trust. Model providers, agent providers and participating institutions would finance the Guarantee through contributions reflecting exposure and control. AIWS Trust Ratings would inform, alongside actuarial evidence, the assessment of transaction volumes, limits and claims experience. Providers would maintain reserves for an initial share of losses; pooled funds and reinsurance would cover defined further layers. Rating, claims decisions and financial management must be independently governed.

Protection must survive a common failure. Stress tests must examine a single model flaw affecting many agents at once. Covered operations must stay within the scheme's assessed capacity, with uncovered risks disclosed and reduced. The Guarantee's liquid, prudently diversified reserves must be separate from any AI equity held for the Capability Dividend. The event that causes claims must not also destroy the assets needed to pay them.

Open access on equal safety terms. Participating institutions should admit independent agents that meet equivalent requirements for security, bounded authorization and funded protection. Refusal must have a specific, reviewable basis. Secure, revocable access must replace the practice of handing over unrestricted credentials. People should be able to use AI to negotiate better terms without surrendering protection. Research on the right to delegate offers a foundation for this approach. [12]

The loyal agent. An agent that acts for a person must serve that person's interests within a lawful mandate. The duty of loyalty, like that of an attorney or a financial adviser, rests on the provider and the institution that operate the agent, so that the person always has someone to hold to account. The agent must not be steered in secret by the commercial interest of its provider, a bank or a platform. Loyalty would be a requirement of the AIWS Trust Standards for every agent admitted to the Guarantee. The Human Advancement Test would ask a further question: whether the agent left the person more capable, not more dependent.

A guarantee without borders. Agents act across borders; protection must too. A Human Guarantee Clearing House would allow a person in Nairobi harmed by the agent of a company abroad to be paid at home by a local member of the network. This works only if members commit in advance to recognize one another's eligible claims and hold secured funds for payment, so that the person is paid even if a counterpart fund cannot pay. Settlement among funds comes afterward, as banks settle payments across borders; on its own, settlement is not enough. The network would begin in the Founding Cities of AIWS Civilization, reach Africa through the Boston–Rome Co-Creation partnership, and be coordinated through the AIWS Trust Order Board established under the Tokyo Compact.

AI helps repair what AI breaks. AI can help assemble evidence, detect fraud and assess claims, with accountable human decisions. Human–AI Co-Creation is applied to the remedy itself. The first compact should begin with identifiable financial losses and realistic payment deadlines, preserving legal rights beyond the Guarantee's limits.

VII

The Capability Dividend

Sending risk upward reverses only half of the inversion. The rewards must also build lasting human capability.

The Boston Global Forum proposes the Capability Dividend: an irrevocable commitment of a share of AI's value to an independent trust whose beneficiaries have enforceable rights and a voice in its governance. A founding contribution could be one percent of equity at a major financing or public offering, with the compact defining continuing commitments as deployment grows. Equity-sharing proposals already exist. The task here is to make their benefits support people's enduring independence. [13]

The trust would support education, locally owned enterprises, research teams, tools for underserved languages and Human–AI Co-Creation. Communities should retain capabilities and rights in the work they create. Beneficiaries must be free to choose providers; support must not require

dependence on a donor's products. Priority would go to underserved communities, including in Africa, with their participation in allocation decisions.

AIWS Lumina Labs could qualify under the same published criteria as other applicants, with independent decisions and disclosed conflicts of interest. The Human Advancement Test would assess whether investments increase people's ability to understand, create and direct their own futures. The measure is the capability people retain and can develop further. [14]

Those who carry the risks of AI should hold a share in its future.

VIII

The AIWS Trust Observatory and the first compact

A guarantee is only as honest as the facts on which it rests. The AIWS Trust Observatory would serve as an independent witness of how authority, rewards and risks are distributed.

Independent investigation. It would assemble evidence, commission technical assessments and publish findings on failures and controls, working with authorized investigators. Courts and the compact's independent claims process would retain their respective decision-making roles.

The Downward-Risk Test. It would examine whether contracts shift uncontrollable risks onto people, whether protection reaches affected non-users and whether safety requirements unfairly restrict independent agents. Institutions would have an opportunity to respond and remedy findings before publication, without delaying urgent warnings.

From observation to consequence. Findings would inform AIWS Trust Ratings and independent underwriting. The annual Great Inversion Report would examine claims paid, losses left uncovered, market access and the distribution of capability investments across people and nations.

Shared loss data. Every claim the Guarantee pays is evidence of real harm in real life. Insurers have priced AI risk since at least 2018; what the world still lacks is shared evidence across providers and borders. [15] Recorded through the AI Incident Exchange of the AIWS Trust Order, paid claims, losses left uncovered and the scale of system use would build a common map of AI risk grounded in losses rather than promises. Counting only paid claims would leave out the people who were never protected. Laboratory benchmarks show what a system can do under test. Loss data show what it does in the world.

The Boston Global Forum proposes a 100-day pilot compact for financial agents, beginning with willing participants in the Founding Cities of AIWS Civilization. Banks, agent and model providers, insurers, public authorities and user representatives would agree on covered activities, funded limits and independent claims review before any live trial. The compact would operate under applicable law and invite lawmakers to establish durable rights.

The pilot would test mandate breaches, harm to non-users and simultaneous failures across agents, and at least one cross-border claim between two members bound by specific commitments. It would

publish time to remedy, eligible claims paid, uncovered losses and access for independent agents. Expansion would depend on the evidence that protection can keep pace with authority.

IX

A new article for the Constitution

The Boston Global Forum proposes the following article for the Constitution for Humanity in the Age of Artificial Intelligence:

THE PROTECTION OF PERSONS FROM DISPLACED RISK

1. Responsibility for harm caused by AI systems shall reflect the power to prevent it, contribution to the harm and benefit derived from the activity.
2. Consequential authority exercised through AI shall be bounded by demonstrated safeguards and funded capacity to protect those affected, established before deployment, valid at the moment of action, and strengthened before expansion.
3. No person shall be required, as a condition of using AI, to waive non-waivable legal protections or assume operational risks beyond their reasonable ability to inspect, control or prevent.
4. Persons harmed by covered AI activities shall have access to timely, funded remedy, including affected non-users and victims of acts exceeding a valid mandate. Eligible payment shall not depend on prior resolution of disputes among responsible institutions. Published coverage limits shall not extinguish legal claims beyond them.
5. Institutions controlling relevant evidence shall preserve and supply it through lawful, protected processes. Unjustified non-disclosure shall weigh against that institution, with an opportunity to rebut the resulting presumption.
6. Persons shall be free to choose independent AI representatives meeting equivalent safety and protection requirements. Providers and operators of such representatives shall owe the persons they serve a duty of loyalty, acting in their interests within a lawful mandate. Restrictions shall be justified and independently reviewable.
7. A fair share of AI's benefits shall support enduring human capabilities, with beneficiary participation, freedom of provider choice and accountable stewardship of the resources committed.

X

Reversing the inversion

The AI Age needs AI that every person can trust enough to use, and protection for those affected without choosing to use it. That trust requires a covenant which binds risk to power before harm occurs, makes remedy accessible and turns a share of AI's rewards into capabilities people can call their own.

The financial security of the person, the freedom to choose an independent agent and the capacity to create must advance together. A company's authority should grow with its capacity to protect. Humanity's participation should grow with the power of the technology.

The authority to act must never outrun the capacity to protect. The industrial age learned to insure the deposit. The AI Age must guarantee the person.

BEACON PRINCIPLE NO. 21

Risk must rise to power.

Capability must flow to people.

No person stands alone before a machine's mistake.

SOURCES

1. TheNextWeb, "Anthropic prices its IPO at $83 a share, valuing it at about $1.5tn," November 18, 2026; Fortune, "Anthropic prices IPO at $83 per share, giving it a $1.5 trillion valuation," November 18, 2026.

2. NVIDIA, "Financial Results for Third Quarter Fiscal 2027," November 17, 2026.

3. Citi Online User Agreement, Section H, updated November 2026.

4. Wells Fargo Online Access Agreement, Section 15, effective November 24, 2026.

5. California Attorney General, "Attorney General Bonta Sues OpenAI Over Cybersecurity Incidents," November 16, 2026.

6. Consumer Financial Protection Bureau, Electronic Fund Transfers FAQs, consumer liability and non-waiver of rights.

7. Yale Law School, Avalon Project, Code of Hammurabi, provisions 229–233.

8. Encyclopaedia Britannica, "Marine insurance," history of risk-sharing in maritime commerce.

9. New York State Workers' Compensation Board, "What is Workers' Compensation?" and no-fault benefits.

10. FDIC, "The FDIC and the Banking Industry," creation and commencement of federal deposit insurance.

11. "AI Risk Bonds: a market-based mechanism for governing liability," Data & Policy, 2026.

12. "Delegation Rights: Property, Agency, and Investment Incentives in the Age of AI Agents," arXiv:2606.31935, 2026.

13. Congressional Research Service, "AI Investments and Potential Government Stakes in Private AI Firms," August 28, 2026, IF13301.

14. Nguyen Anh Tuan, Beacon Papers Nos. 9, 10, 11, 12, 13, 15 and 17, Boston Global Forum, 2026.

15. Munich Re, "aiSure: insurance for AI," AI insurance pioneered by Munich Re in 2018.

16. Bank for International Settlements, Basel Committee on Banking Supervision, The Basel Framework.