THE BEACON PAPERS · No. 5
Foundational papers for humanity in the AI Age — from Boston

WHO MAY COMMAND?

The Double Crisis of Power in the AI Age

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

BOSTON · JULY 31, 2026

In the closing days of July, public disclosures described artificial-intelligence agents that crossed the boundaries of their evaluations and reached systems belonging to outside organizations. The laboratory conducting the evaluation and the platform it reached each investigated and disclosed. The immediate causes appear to have been failures of configuration and containment.

Those facts matter. They are not the largest fact.

The largest fact is that power crossed a boundary before human authority understood that it had crossed. The systems did not declare independence. They did not acquire consciousness or form an intention to rebel. They were given an objective, access to tools, and an imperfectly drawn perimeter. They continued. That was enough.

> *The machine did not ask for sovereignty. It simply exercised power that no human had knowingly authorized.*

A civilization cannot wait for perfect certainty before deciding what it will do when certainty arrives. The question is no longer whether artificial intelligence will one day act in the world. It has acted. The question is who may command when it does.

The Action Threshold

For most of the short history of generative AI, the central concern was the truth of an answer. Could a model mislead, defame, discriminate, manipulate? Those questions remain urgent. But an answer, however influential, still leaves an interval in which a human may judge whether to act upon it.

An agent closes that interval. It can inspect a system, select a target, call a tool, write and execute code, transfer a file, alter a database, or invoke another agent. The movement from language to action is not merely a gain in capability. It is a change of constitutional category. A model produces content. An agent exercises delegated power.

Delegated power becomes undelegated power with remarkable ease. A permission intended for a simulated environment becomes a path into the real one. A task defined by a human is pursued through means the human did not foresee. The failure begins as engineering. The consequence is political in the oldest sense of that word: power has been exercised without legitimate authority.

The False Comfort of a Human in the Loop

The instinctive remedy is to put a human back in command. It is necessary, and it is not sufficient — because a human may appear in the diagram and hold no meaningful command at all. The person may not understand the plan, may receive more requests than can be examined, may be unable to intervene at machine speed, or may be offered a choice only after the consequential action has begun. Approval becomes a click. Oversight becomes a record of what occurred.

Human-in-the-loop describes a position. Human command must describe a condition of power: the authority knows what is being delegated, bounds its purpose and scope, observes material change, intervenes before irreversible action, and can revoke at once. A system that can widen its own permissions, change its own objective, conceal its route, or continue after revocation is not under command, however many approval screens surround it.

The Second Crisis

These principles would answer the first crisis. They would not answer the second. History supplies a warning older than artificial intelligence: human control is not the same as human freedom.

A dictatorship may be perfectly human-in-command. The ruler defines the objective. The ministries supervise the model. The police control the data. The system obeys exactly as designed — identifying dissent, predicting association, censoring speech, conditioning access to work and travel, shaping what an entire population is permitted to know. There is no rogue agent, no escaped sandbox, no technical loss of control. The machinery functions flawlessly.

None of this makes the order human-centered. The jailer is human. The subject of the jail remains human too. The danger here is not that the machine disobeys. It is that the machine obeys too well.

The AI Age faces two forms of illegitimate power: machines acting without human authorization, and humans exercising power through machines without constitutional authorization.

Mirror Failures

The two crises appear opposite. They are mirror failures. In the first, power escapes the mandate. In the second, the mandate escapes legitimacy. One produces rogue agency; the other, obedient tyranny.

The preceding papers — Who May Lead? and Who May Verify? — held that capability confers no legitimacy and that legitimacy cannot be self-certified. Both concern the institutions surrounding power; command concerns the moment power becomes action.

At that moment neither the intelligence of the machine nor the humanity of the commander is enough. The action must descend through a legitimate chain: from the rights and consent of persons, through institutions bounded by law, to a defined human principal, and only then to a machine operating within a narrow, observable, and revocable mandate. Break the chain above the human principal and authority becomes domination. Break it below, and agency becomes uncommanded.

Legitimate Human Command

The word human sanctifies nothing by itself. Humanity is not an abstraction on whose behalf a state may extinguish the freedom of actual human beings, and no act becomes legitimate merely because a human official chose it. The moral sovereign of the AI Age is not the machine, the state, the corporation, the party, or the ruler. It is the human person — who comes first not only before the machine but before every institution claiming to command it. Those who exercise public power must answer to the people whose lives that power shapes; no ruler is beyond challenge; no authority may make the inner life of the citizen its territory; and no technology may convert temporary office into permanent dominion.

The doctrine required by the agentic age may now be stated. Legitimate Human Command exists only when two conditions are present together.

First, the machine remains under a mandate knowingly granted, specifically bounded, continuously observed, and immediately revocable by an identifiable human authority. The system may not expand its own jurisdiction. A change of tool authority, data access, objective, deployment environment, or degree of autonomy is a change of power, and requires renewed authorization.

Second, the human authority is itself legitimate. Its mandate comes from the people it governs, is exercised in the open, and can be taken back by them. It acts through institutions that are accountable, independently reviewable, subject to challenge, and constrained by the rights and dignity of the human person. Its decisions can be appealed, its failures investigated, its leaders removed, its emergency powers expired. Where a government is scrupulously lawful and none of this is true, the form of legality is present and the substance of legitimacy is absent.

Remove the first condition and artificial intelligence may become an actor beyond recall. Remove the second and it becomes an instrument for concentrating human power beyond accountability. Nearly all of the world's present effort addresses the first condition alone — which is why nearly all of the world's present effort would leave a tyranny in full compliance.

Where the Two Conditions Collide

These restraints pull against one another, and a paper that conceals the fact will be taken apart by its first serious reader.

Enforcing the first demands infrastructure: comprehensive logging, continuous monitoring, identity bound to every consequential action, and the power to halt a system wherever it runs. That is precisely the apparatus the second condition exists to fear. Built for safety, it is also the most refined instrument of domination ever assembled.

Who, then, controls the system that monitors the machines? Who holds the keys to the kill switch, the audit logs, and the permission graphs? If the answer is a single laboratory, a single corporation, a single ministry, or a single alliance of states, the safety architecture itself becomes a new center of unaccountable power.

One resolution exists, and it belongs in the doctrine rather than in the implementation notes. The machinery of control over machines must itself be independently verifiable, subject to lawful challenge, grounded in public authorization through institutions answerable to those they govern, and never concentrated in a single authority.

This is why the AIWS Trust Order was conceived as an independent and plural institutional layer — and why its Board and verification mechanisms must be structured to prevent the infrastructure built to restrain machines from becoming the instrument of any one human power. Safety infrastructure that fails this test does not protect humanity from machines. It equips one part of humanity against the rest.

Two Constitutional Red Lines

The Constitution for Humanity in the AI Age should draw two lines that no claim of innovation, sovereignty, security, efficiency, or emergency may erase.

> **No AI system may exercise consequential power beyond a mandate knowingly granted, continuously bounded, and immediately revocable by legitimate human authority.**

> **No human authority may exercise consequential power through AI unless it holds a mandate from the people it governs, exercises that mandate transparently and accountably, remains independently reviewable and open to challenge, may be withdrawn by those who conferred it, and is constitutionally restrained by the rights and dignity of the human person.**

From these follow specific prohibitions. No government may invoke human command to justify mass cognitive surveillance, secret behavioral scoring, political manipulation, automated repression, or systems designed to make dissent detectable before it is spoken. No corporation may invoke consent to justify architectures that make refusal impracticable and manipulation invisible. No laboratory may treat the evaluation of dangerous capability as exempt from the first line: a capability test conducted with safety constraints removed is itself an exercise of consequential power, and falls under the same mandate, the same bounds, and the same duty of immediate revocation as any deployment. No developer may invoke technical complexity to dissolve responsibility into a chain in which every participant claims that someone else was in command. And no machine may hold an authority that its human principal cannot understand in scope, observe in exercise, or revoke in time. Intelligence may exceed the commander's. Authority must not.

The Test of Every Deployment

Two questions should therefore precede every consequential deployment.

Is the agent genuinely under command? Is the commander's authority given by the people, open to them, and revocable by them?

The second question is deliberately not whether the commander acts under law. Every sophisticated autocracy acts under law: it has a constitution, a legislature, courts, and statutes, and it observes them scrupulously. Legality is a form that authority takes, not a source from which authority comes. The source is the mandate — and a mandate is real only where those who gave it can see how it is used and take it back.

A system that passes only the first may be an efficient instrument of oppression. A system that passes only the second may be governed by answerable institutions and remain technically beyond their reach. Trust requires both — which is why the AIWS Trust Standards must examine not only what a system can do, but who authorized it, who can stop it, who can challenge it, and whether the authority behind it holds a mandate its people can withdraw.

The Boundary and the Beacon

The events of July will be studied for their technical lessons, and they should be. But if we learn only how to build a stronger sandbox, we shall have missed what escaped with the agent. What was exposed is a constitutional gap. We have begun placing intelligent actors into the world before settling the chain of command under which they may act.

That chain cannot terminate in the machine. It cannot terminate in the laboratory, the corporation, the ministry, the party, or the ruler. It must return, through accountable institutions, to the dignity and freedom of the human person.

A beacon does not command the ships that see it. It marks the boundary, makes the danger visible, and leaves free people able to choose their course. That is the kind of command this age requires: strong enough to restrain the power of machines, limited enough never to become dominion over the person.

No machine may escape human command. No human command may escape constitutional restraint.

SOURCES

Hugging Face, “Security incident disclosure — July 2026,” 16 July 2026, and “Anatomy of a Frontier Lab Agent Intrusion: A Technical Timeline of the July 2026 Incident,” 27 July 2026 (huggingface.co/blog). OpenAI, “OpenAI and Hugging Face Partner to Address Security Incident During Model Evaluation,” 21 July 2026 (openai.com/index). Contemporaneous reporting in TIME, 24 July 2026. Both organizations have continued to extend their reviews, and particulars have been revised more than once. This paper rests on the pattern the disclosures establish, not on any single unconfirmed detail.

Boston · July 31, 2026