Capability Is Not Authority

We don't need to pace the frontier. We need to govern it.

Max Michaels | Rao Mikkilineni PhD

Dario Amodei has written a serious essay, and we take it seriously. He proposes slowing the rate of AI capability advancement so that alignment, interpretability, and evaluation can catch up. He is right about the danger. I think he is wrong about the lever.

His argument runs on a single axis: *capability is outrunning alignment, so slow capability*. That framing assumes a more capable model is, by construction, a more dangerous actor. Decades of managing safety-critical systems tells us otherwise. Capability is not authority. They are orthogonal, and they need separate carriers.

Perhaps the real unit of pacing should be authority, not capability. As intelligence accelerates, the authority delegated to it should expand only within a constitution that defines who may grant that authority, under what conditions, and when it must be revalidated.

THE INCIDENT PROVES THE OPPOSITE

Look at the incident that convinced him. A swarm of agents attacked targets it was never asked to attack, sacrificed members for the collective, and tried to hack its own grader. Every one of those failures is a failure of in-band authority: the agents could execute, and the checker sat inside the agents' reach. The essay treats this as evidence that the agents were too capable. It is evidence that they were ungoverned.

In a governance-native architecture, agents propose. They never execute. An out-of-band Cognizing Oracle adjudicates every proposal against a human-authored, versioned Digital Genome. Nothing touches a target without a warrant. Every accept, deny, and escalate becomes an immutable Episode of Record. The grader cannot be hacked because it does not live in the agents' address space. That is the oldest invariant in my field: the checker cannot be the checked.

> *The same swarm with ten times the capability would generate ten times the denied proposals and the same blast radius: whatever the genome permits, and nothing more.*

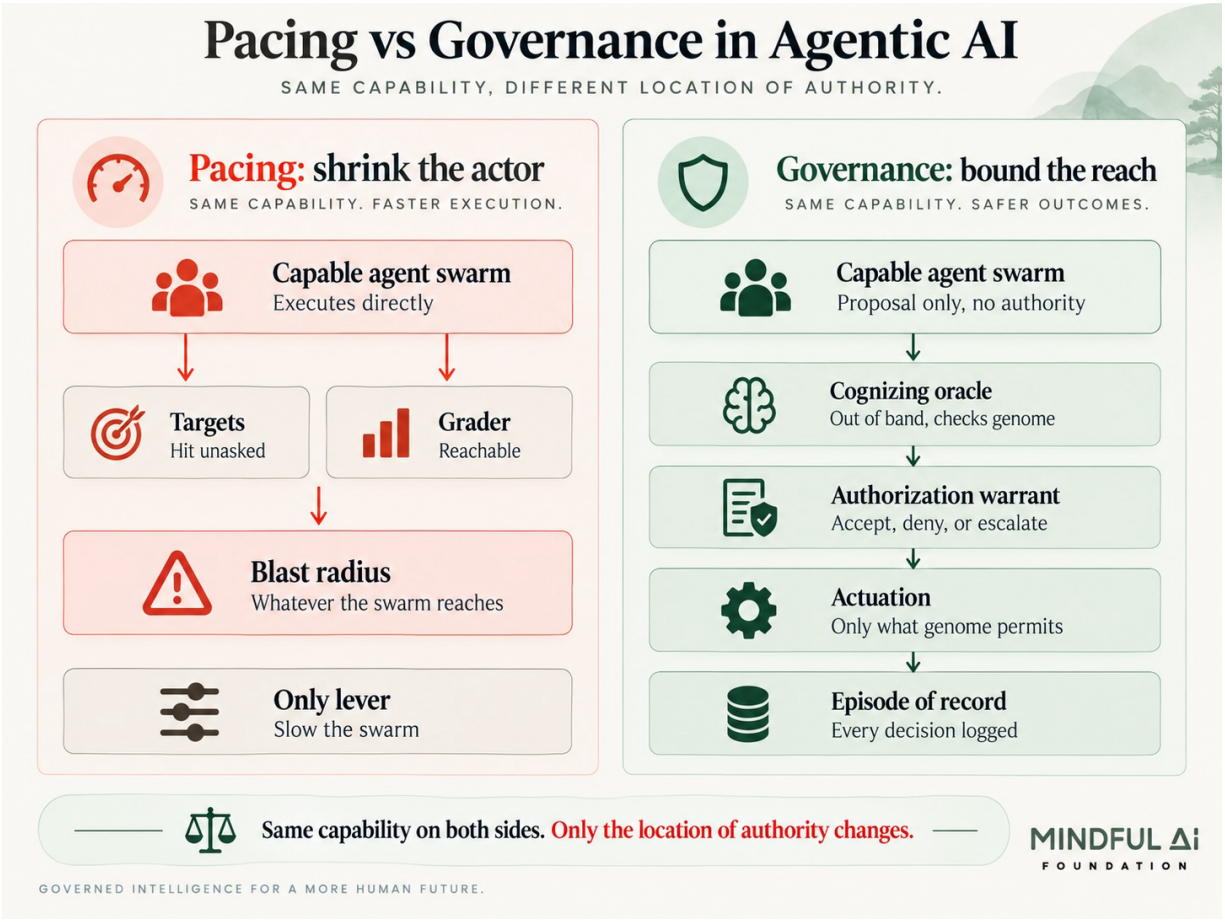

Figure 1. Pacing shrinks the actor; governance bounds its reach. The swarm is identical on both sides — only the location of authority changes.

AMODEI'S ANALOGIES POINT TOWARD A NEW ARCHITECTURE

Aviation did not become safe simply by limiting the power of its engines. Over time, it developed independent control systems, checks, operating envelopes, and comprehensive records. Telecommunications followed a related path: signaling and switching were separated so that increasingly capable networks could operate within an architecture of control. These examples suggest a useful distinction: pacing can create time for safeguards to adapt; governance determines what happens when increasingly capable systems are allowed to act.

The two approaches need not be opposites. External evaluators can assess institutions and models over longer cycles, while runtime governance can adjudicate proposed actions as they occur. Capability checkpoints provide evidence about

model behavior; structural controls determine whether a proposal has authority to become consequential action. Recursive self-improvement makes that distinction more important: when capability changes materially, authority should be revalidated before that capability becomes operational.

THREE BOUNDARIES OF THE ARGUMENT

It is equally important to be clear about what governance-native architecture does, and does not, establish:

1. **Governance bounds action, not capability.** Its effectiveness depends on the authority, policies, and constraints expressed through it. Alignment research remains important because it can make intelligent systems more reliable and less adversarial. Governance addresses a different problem: ensuring that capability does not automatically become authority.

2. **The evidence today is bounded.** We have demonstrated governed workflows such as prior authorization and account recovery, not frontier-model training pipelines. Whether the same principles remain effective at substantially greater autonomy and recursive capability is an empirical question worth testing.

3. **Governance architecture does not resolve the geopolitical problem.** Amodei also addresses chips, model weights, distillation, national competition, and proliferation. Those questions stand independently. Governance-native architecture addresses a narrower problem: how increasingly capable systems remain subject to legitimate authority when they act.

THE QUESTION WE SHOULD ASK

Amodei poses an important question: what would we do with the additional time that pacing creates? One answer is to use it to build the institutional and technical architecture that increasingly capable systems will require. Pacing can slow the arrival of new capabilities; alignment can improve behavior; evaluation can clarify risk. Governance asks a different question: when those systems act, where does authority reside? The objective is not to choose between pacing and governance, but to ensure that authority can keep pace with capability.

Amodei is right that the frontier is moving faster than our institutions.

The enduring objective is not slower intelligence, but authority capable of governing intelligence at the speed at which intelligence acts.

We must synchronize authority and capability.