

Four Layers of Control for Advanced AI

The debate over advanced AI is often framed as a choice between accelerating its development and slowing it down. That framing is too narrow. AI can continue to become more capable while society separately determines what those systems are permitted to do.

A durable safety framework should therefore distinguish four different functions:

PROTECT	ALIGN	GATE	BOUND AUTHORITY
1. Protect - Establish the Boundary Establish a minimum protected baseline that advanced AI must not be permitted to override, beginning with human survival, fundamental human agency, and continued human-governed authority over consequential domains. HOW Seek the narrowest achievable common ground across nations and cultures rather than attempting to define a complete universal value system.	2. Align - Reduce the Probability of Harm Make advanced AI as reliably consistent as possible with that protected baseline. HOW Use training, behavioral safeguards, evaluation, interpretability, monitoring, human oversight, and continuing alignment research. Alignment can reduce risk, but it cannot provide absolute assurance and should not itself confer authority.	3. Gate - Control Admission to Consequential Domains Do not equate the capability to perform an action with permission to enter a domain in which that capability can create consequential effects. HOW Require evidence of safety and control proportionate to the potential consequence before granting access to consequential systems, resources, or privileges.	4. Bound Authority - Control Consequential Execution Never make safety depend on alignment alone. Once an AI is admitted into a consequential domain, independently constrain the actions, sequences, and cumulative authority it is permitted to exercise. HOW Place enforceable authorization between AI and consequential effect. Every material pathway through which intelligence can become external action should have an authority boundary appropriate to its risk.

Key distinction: Layer 3 governs admission to consequential domains. Layer 4 governs execution once a system is operating within them.

Alignment and authority are different safety functions.

ALIGNMENT	AUTHORITY CONTROL
Governs what intelligence should do .	Governs what intelligence is permitted to cause .
It shapes behavior, objectives, reasoning, and adherence to protected interests.	It independently constrains whether proposed actions may become consequential effects.

Why the fourth layer matters

Alignment remains essential. So do evaluation, governance, cybersecurity, human oversight, and decisions about when and where advanced systems should be deployed. But none eliminates the need for an independent boundary between intelligence and consequence.

Bound authority at every material pathway. Consequential AI systems may act through many different pathways - financial transactions, networks, devices, robotics, credentials, APIs, memory, delegation, or other mechanisms. No single control should be assumed sufficient for every pathway or threat.

External systems can enforce their own boundary. Banks, networks, infrastructure, devices, and other consequential systems can require valid authorization before accepting AI-generated actions, so that an unrestricted or compromised model does not automatically acquire authority over systems it does not control.

Enforcement depth should match the assurance requirement

Software	Protected runtime	Firmware	Hardware roots	Accelerators / chiplets	Potential silicon
----------	-------------------	----------	----------------	-------------------------	-------------------

Use the shallowest enforcement layer that can credibly provide the required assurance. Move deeper only when necessary. Where ordinary software is insufficient, enforcement can move progressively into protected runtime, firmware, hardware roots, accelerators or chiplets, and ultimately silicon where warranted.

The authority boundary must be independent of the intelligence it constrains and strong enough to survive the relevant threat model.

INCREASING INTELLIGENCE SHOULD NOT ENABLE INCREASING AUTHORITY TO ACT.

Advanced AI may become dramatically more capable. Its authority should increase only through explicit, bounded, and independently enforceable authorization.

Scope: This framework does not replace technical alignment research, evaluation, governance, cybersecurity, containment, human oversight, or other safety mechanisms. It separates their functions and adds an independent execution-control layer so that the protection of humanity does not depend on alignment alone.