VIEW POINT

ALIGNMENT AUTHORITY AS AN ARCHITECTURAL VARIABLE

Abstract

Artificial intelligence is rapidly becoming foundational infrastructure within federal systems. As agencies move beyond experimentation toward operational deployment, architectural decisions must account not only for model capability, but also for control, governance, and long-term sovereignty.

This five-part series examines federal AI adoption through the lens of Behavioral Sovereignty, defined as the authority to determine how AI systems are aligned, configured, and deployed within lawful mission contexts. Rather than framing the debate as a binary choice between open and closed models, the series introduces a Sovereignty Gradient that clarifies how different deployment models distribute alignment authority, operational responsibility, and infrastructure dependence.

Infosys
Navigate your next

Articles 1 through 3 established Behavioral Sovereignty, the Sovereignty Gradient, and the emergence of a layered federal AI architecture. The next step is to examine alignment authority at a deeper technical level.

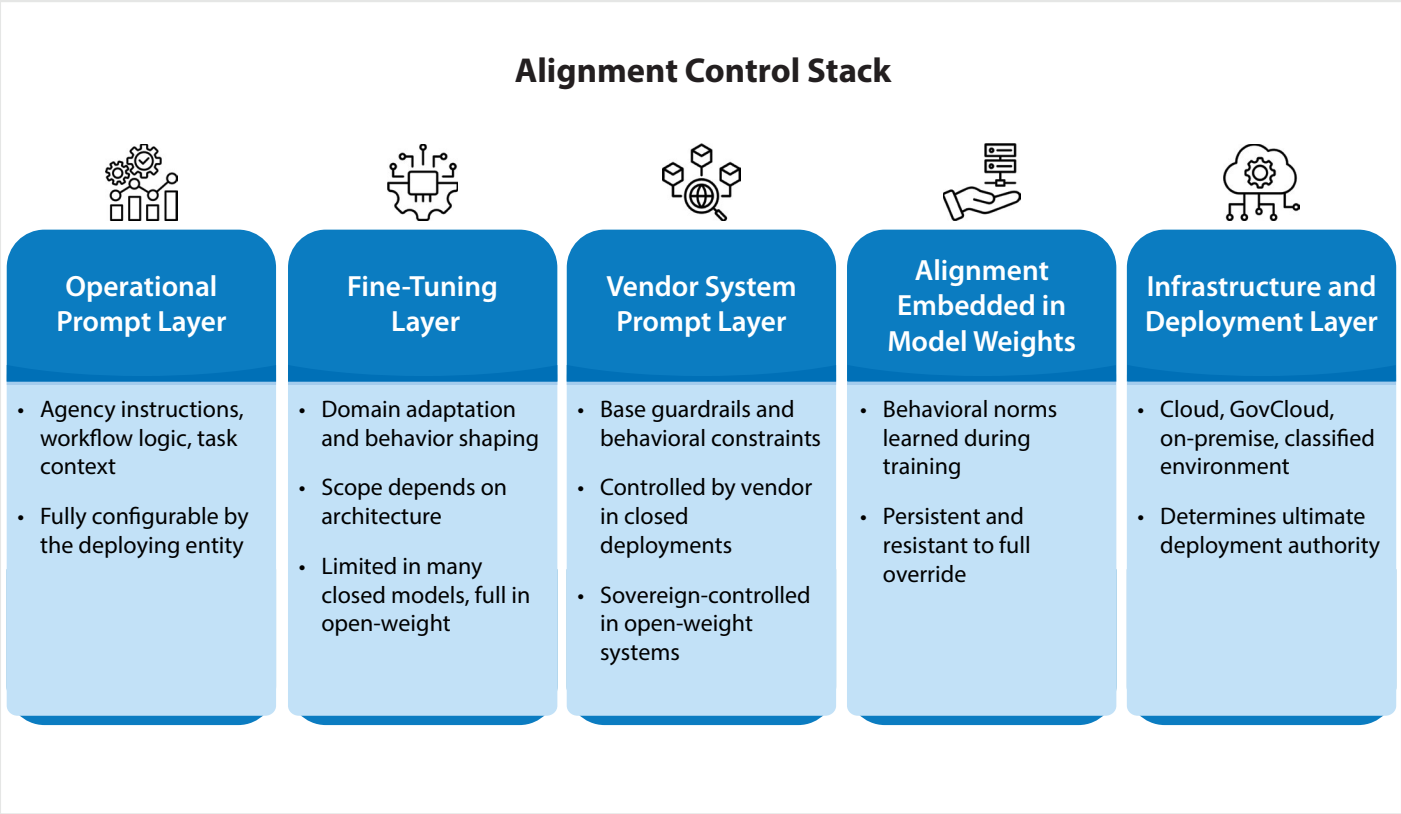
Alignment is often discussed as if it were synonymous with prompt engineering or policy configuration. In practice, alignment authority exists across multiple layers of the model stack. Authority at one layer does not imply authority at all layers.

Understanding this distinction is essential to evaluating control ceilings in federal deployments.

The Alignment Control Stack

Modern frontier models operate through layered behavioral controls. These layers determine how much authority a deploying entity actually possesses.

Below is a simplified conceptual representation of those layers.



Each layer contributes to the overall behavioral profile of the system.

Authority at the operational prompt layer does not automatically extend to weight-level alignment. Fine-tuning can modify domain behavior but may not fully override embedded alignment reinforcement in closed systems. Infrastructure choices determine who ultimately controls access, updates, and patch cadence.

This layered structure explains why Behavioral Sovereignty cannot be evaluated at the surface level.

Weight-Level Alignment Persistence

Alignment embedded in model weights is reinforced during pretraining and post-training alignment processes. In closed systems, this layer is not directly accessible to the deploying entity.

Even when operator-level prompts or supervised fine-tuning are permitted, certain guardrails may remain structurally persistent. These ceilings are not necessarily visible through API documentation. They become apparent only when specific categories of behavior cannot be modified through available configuration tools.

In open-weight deployments, the weight layer becomes modifiable. Full-parameter fine-tuning or low-rank adaptation techniques allow reshaping of alignment boundaries. This does not eliminate responsibility. It transfers it.

In my view, this transfer of alignment authority is one of the most consequential architectural differences between closed and open-weight systems.

Closed models centralize responsibility with the vendor. Open-weight systems decentralize responsibility to the deploying entity.

Fine-Tuning Ceilings

Fine-tuning is often assumed to provide full behavioral control. That assumption is inaccurate in many closed-model environments.

Closed vendors may permit supervised fine-tuning for domain adaptation, but safety constraints embedded in weights or system prompts often persist. The result is a layered override ceiling.

Government cloud deployments may increase data isolation and compliance posture. They do not inherently remove alignment ceilings defined at the vendor level.

Open-weight deployments eliminate vendor-defined ceilings but introduce new obligations:

- Alignment testing pipelines
- Red-teaming capacity
- Safety validation procedures
- Continuous evaluation against misuse risk

Authority expands. Accountability expands with it.

Responsibility Redistribution

The shift from vendor-managed safety to sovereign-managed alignment is not a purely technical transition. It is organizational.

Closed systems reduce internal ML governance burden. Vendors maintain safety research teams, policy enforcement frameworks, and update cycles.

Open-weight systems require the deploying entity to replicate portions of that capability. This includes:

- Adversarial testing
- Misuse scenario modeling
- Ongoing evaluation of alignment drift
- Infrastructure hardening

This redistribution of responsibility is rarely discussed in policy debates. It is central to realistic sovereign deployment.

Behavioral Sovereignty is not simply about removing guardrails. It is about assuming the operational capacity to manage them.

Alignment Authority and Procurement

The When federal agencies evaluate AI systems through the lens of alignment authority, the Sovereignty Gradient becomes more than conceptual.

Closed systems offer velocity and vendor-managed safety.

Open-weight systems offer maximum alignment control.

Government cloud deployments offer an intermediate configuration.

The appropriate choice depends on mission sensitivity, risk tolerance, and operational maturity.

This layered approach aligns with broader federal strategy themes of risk-tiered deployment and infrastructure resilience. It also explains why procurement adjustments can occur when alignment ceilings conflict with sovereign mission requirements.

Alignment authority is not a secondary feature.

It is an architectural variable.

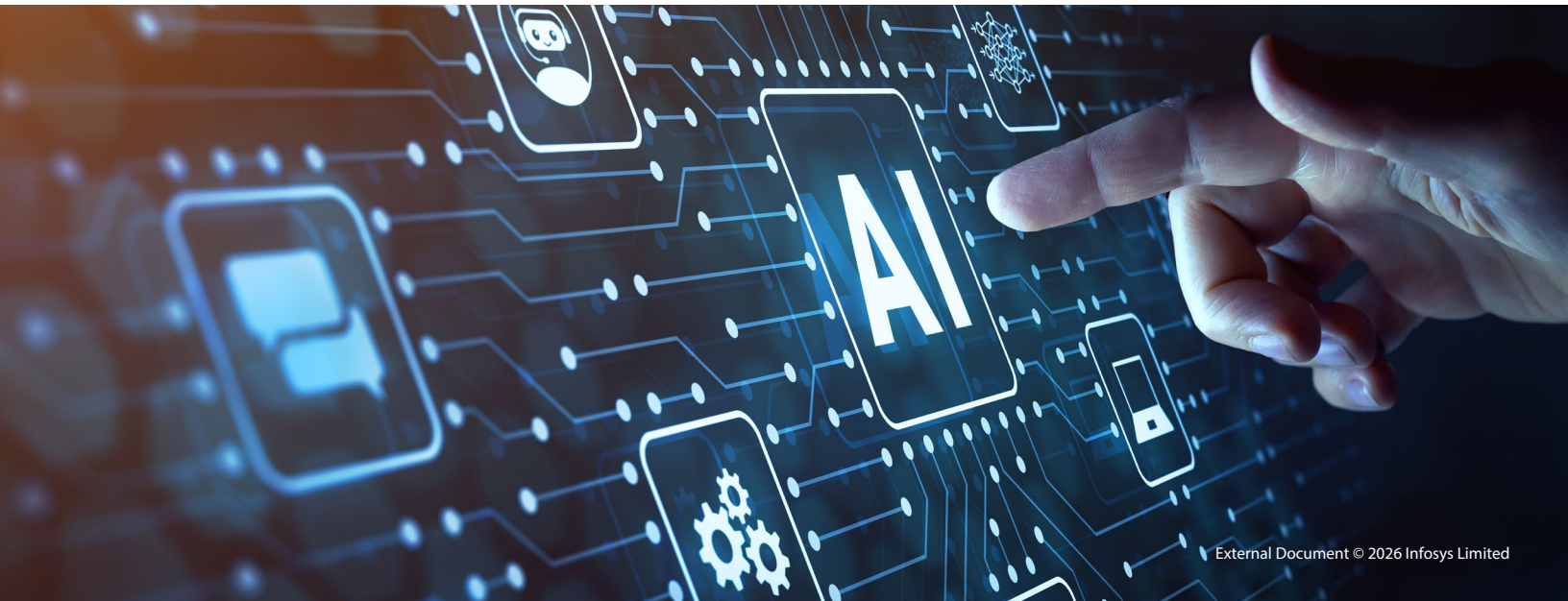
From Alignment to Orchestration

Alignment control does not exist in isolation. As federal AI systems evolve toward multi-model orchestration and agent networks, alignment interactions become more complex.

Supervisory agents may coordinate multiple underlying models. Some may be closed. Others may be open-weight. Each brings its own alignment profile and override ceiling.

The layered architecture introduced earlier allows these systems to coexist without collapsing into a single point of behavioral control.

The next article explores this transition from model-centric deployment to multi-model orchestration and agent architectures, and examines how infrastructure sovereignty underpins that evolution.



About the Authors

Scott Yentzer

Scott Yentzer is Principal Technology Architect at Infosys Center for Emerging Technology Solutions, where he leads enterprise AI, agentic systems, and advanced R&D initiatives. With 20+ years bridging cutting-edge research and large-scale deployment, he holds an M.S. in Information Management from Arizona State University.

For more information, contact askus@infosys.com

Infosys
Navigate your next