The AI OSI Stack: A Governance Blueprint for Scalable and Trusted AI

Daniel P. Madden

Independent AI Researcher and IT Specialist

danielpmadden.com

Originally conceived September 2025
Revised October 2025

Abstract

Artificial intelligence is moving from discrete products to systemic infrastructure. Yet most AI is still designed and governed as if it were a single opaque system. This creates concentration risk, weak accountability, and an environment in which a few model or API providers can shape whole sectors. This paper proposes the **AI OSI Stack**, a seven layer architectural and governance framework for clarifying how AI is built, where risks concentrate, and how trust can be made portable. Inspired by the Open Systems Interconnection (OSI) model in networking, the AI OSI Stack separates the AI ecosystem into: (1) Physical / Hardware, (2) Model Architecture, (3) Training / Optimization, (4) Instruction / Control, (5) Interface / Protocol, (6) Application, and (7) Governance / Trust. This separation of concerns turns AI governance from a reactive afterthought into a design time feature. It enables targeted regulation, auditable decision artifacts, persona based safety layers, and protection against interface monopolies. The Stack is intended for AI labs, policymakers, enterprise architects, and standards bodies that need an operational map for trustworthy AI.

1 Introduction: From Opaque Systems to Accountable Architectures

Current AI systems are too often treated as black boxes. Hardware, models, training data, APIs, and safety controls are collapsed into a single object called “AI.” This collapse hides where power sits and where failure originates. It also makes meaningful governance nearly impossible, because regulators, executives, and engineers are never talking about the same layer.

This problem is not only technical. It is strategic. The AI economy is already showing signs of what we can call *mitosis*: a rapid branching into specialized model lineages inside a high mortality environment. Some branches survive and become infrastructure. Others fragment or are captured by private platforms. In this environment, a layered architecture is no longer a nice abstraction. It is a requirement for accountability, resilience, and competition.

The historical precedent is clear. Networking only became truly global after the OSI model and later TCP/IP clarified who was responsible for what. AI now sits in the same moment. The AI OSI Stack offered here was first drafted in September 2025 as an attempt to bring that same architectural clarity to AI, with governance included as a first class layer rather than an external legal wrapper.

2 Background: Why AI Needs a Layered Model

In the early years of computer networking, vendors published proprietary protocols that did not interoperate. The result was fragmentation, vendor lock in, and low trust. The ISO 7498 1 OSI model solved this by separating concerns into seven conceptual layers. Each layer had a defined purpose, was testable, and could be replaced without rebuilding the whole system.

AI today looks similar to networking in the 1970s. There is:

- platform level concentration at the hardware layer (export controlled GPUs and data center scale compute),
- a small number of model providers at the architecture layer,
- opaque training pipelines,
- closed source alignment and control policies,
- API mediated access that can be switched off unilaterally,
- and above this, a large but vulnerable application ecosystem.

Without a layered model, all of these tensions are discussed at once. That leads to two bad outcomes. Either regulation overreaches and tries to police everything generically, or it underreaches and leaves real chokepoints untouched. A layered model creates precision. It tells everyone where to look.

3 The AI OSI Stack: Seven Layers

The AI OSI Stack is a conceptual and operational map. It does not compete with the NIST AI Risk Management Framework, ISO/IEC 42001, or the EU AI Act. It gives those instruments an architectural surface to attach to.

Layer 1: Physical / Hardware

What it is. GPUs, TPUs, AI optimized ASICs, high bandwidth networking, storage, and the cloud and on premises data centers that host them.

Why it matters. Every higher layer depends on this one. Hardware supply is already a geopolitical issue. A small cluster of companies can shape global AI capability by controlling access to accelerators.

Risks.

- Supply chain concentration and export controls.
- Energy and cooling constraints.
- National security dependencies on foreign owned data centers.

Governance levers. Infrastructure resilience assessments, transparency on origin and lifecycle, diversification incentives.

Layer 2: Model Architecture

What it is. The underlying model families and architectures: transformer variants, diffusion models, multimodal architectures, and emerging neurosymbolic hybrids.

Why it matters. This layer encodes potential capability before training. Safety and alignment are easier when architectures are designed with controllability in mind.

Risks.

- Frontier capability leakage through open weights.
- Over centralization of research directions by a few labs.
- Architectural opacity that makes downstream evaluation harder.

Governance levers. Capability reporting, architecture level safety evaluations, incentives for interpretable architectures.

Layer 3: Training / Optimization

What it is. Data pipelines, optimizers, fine tuning, reinforcement from human feedback, constitutional and rule based training, distillation, and efficiency techniques.

Why it matters. This layer determines cost, reproducibility, and provenance. It is also where unvetted data, copyrighted material, or sensitive personal information may enter the system.

Risks.

- Opaque data provenance.
- High training costs that create economic barriers to entry.
- Safety erasure when models are later fine tuned without constraints.

Governance levers. Data lineage requirements, training run registries, reproducibility standards, and structured model cards.

Layer 4: Instruction / Control

What it is. Everything that shapes model behavior after training: prompts, embeddings, tool use, guardrails, role based or persona based controllers, RLHF, constitutional AI, and your own work on persona architecture, epistemology by design, and decision insurance.

Why it matters. This is the first layer where human intent and model capability meet. If this layer is weak, models can be repurposed for harmful goals or can be steered to outputs that look compliant but are logically or ethically compromised.

Risks.

- Misalignment between stated policy and actual control logic.
- Jailbreaks and prompt injection.
- Hidden influence or covert persuasion through instruction templates.

Governance levers. Independent auditing of alignment methods, red teaming of control policies, publication of high level control intent, mandatory refusal protocols for high risk domains.

Layer 5: Interface / Protocol

What it is. APIs, SDKs, orchestration frameworks, agent runtimes (like OpenAI AgentKit, Anthropic MCP, LangGraph), middleware, and integration surfaces into enterprise systems.

Why it matters. This layer is the new strategic chokepoint. Whoever controls the API or runtime can set pricing, access, allowed tools, logging rules, and enterprise deployment patterns.

Risks.

- Interface monopolies that block competition at the application layer.

- Opaque logging and telemetry.
- Automatic tool execution without human legibility.

Governance levers. Open protocol standards, mandatory execution transparency, exportable audit trails, right to portability for agent workflows.

Layer 6: Application

What it is. Everything users actually touch: copilots, internal decision assistants, sector specific tools in health, finance, education, manufacturing, publishing, and public administration.

Why it matters. This is where harm is experienced and value is created. It is also where context determines whether a model behavior is acceptable.

Risks.

- Misuse in sensitive sectors.
- Shadow AI deployments inside enterprises without oversight.
- Overreliance on model outputs without disclosure of limitations.

Governance levers. Context specific deployment policies, human in the loop requirements, user facing transparency reports, sectoral impact assessments.

Layer 7: Governance / Trust

What it is. The institutional and social layer. Audits, compliance, safety boards, evaluation protocols, transparency tools, decision cards, guardian notes, and the whole class of artifacts you already produce in your lab work.

Why it matters. Trust does not emerge automatically from good models. It is constructed. If this layer is weak, well designed systems will still fail in public because people cannot see how they make decisions.

Risks.

- Regulatory capture by incumbent vendors.
- Fragmented, incompatible governance regimes across jurisdictions.
- Ethics theater that creates reports but not accountability.

Governance levers. Shared audit schemas, public model behavior policies, cross jurisdictional alignment, publication of safety incidents, and long term monitoring.

4 Chokepoint and Power Analysis

One advantage of the AI OSI Stack is that it makes power visible. Different layers are controlled by different actors.

- Layer 1 is dominated by chipmakers and hyperscalers.
- Layers 2 and 3 are dominated by frontier labs.
- Layer 4 is co dominated by labs and safety researchers.
- Layer 5 is where platform companies can entrench themselves.
- Layer 6 is where startups compete.
- Layer 7 is where governments and standards bodies try to re enter the game.

Without this analysis, regulation tends to target the visible layer (applications) and ignore the structural one (interfaces and hardware). The Stack allows policymakers to regulate Layer 5 behavior to protect Layer 6 innovation. It also allows enterprises to demand exportable audit artifacts from Layer 7 before accepting automated decision support in Layer 6.

5 Governance and Compliance Integration

This framework is compatible with and complementary to existing instruments:

- **NIST AI RMF 1.0** defines core functions (govern, map, measure, manage). The AI OSI Stack tells you at what layer to perform each function.
- **ISO/IEC 42001** defines an AI management system. The Stack provides the architectural view to map organizational controls to technical layers.
- **EU AI Act** defines a risk based approach. The Stack helps map risk levels to layers, so that high risk applications at Layer 6 can be traced down to model, data, and control assumptions at Layers 2 to 4.

In practice, this means an enterprise or regulator can write policy like this: *All systems at Layer 6 that target healthcare, finance, public safety, or employment must consume only from APIs at Layer 5 that provide policy legibility and exportable decision logs, and must document the alignment method used at Layer 4.*

That is clearer, more testable, and more future proof than regulating “AI” in the abstract.

6 Applications for Different Stakeholders

6.1 Enterprises

Enterprises can use the Stack as an internal audit map. A feasibility or risk review can be run layer by layer. Your existing governance reports, decision cards, and guardian notes drop neatly into Layer 7. Your Playbook style feasibility engine maps to Layer 6 inputs and Layer 5 constraints.

6.2 Policymakers and Standards Bodies

Policymakers can use the Stack to avoid overbroad law. Rather than banning an entire class of models, they can target Layer 5 to prevent API monopolies, or Layer 4 to require transparent alignment. Standards bodies can define test suites per layer.

6.3 Researchers and Builders

Researchers can innovate at a single layer without destabilizing the whole. A team can work on persona based control (Layer 4) while another team experiments with micro data centers (Layer 1) and both will still be compatible at Layer 5.

7 Future Work and Evolution

The Stack is not static. Several trends will stress it:

- **Agentic systems** that act across tools will increase the importance of Layer 5 observability.
- **Edge and on device models** will reshape Layer 1 and Layer 3, because hardware will no longer be centralized.
- **Vertical AI releases** from major labs, like finance or health specific assistants, will raise the stakes at Layers 6 and 7.
- **International divergence** in AI law will require the Trust layer to support multiple simultaneous governance regimes.

Future versions of the Stack may therefore need sublayers. For example, Layer 4 could be split into *Control* and *Ethical Reasoning*, to reflect work like persona architecture and epistemology by design. Layer 7 could be extended to include cryptographic attestations and supply chain proofs.

8 Conclusion

This paper argued that AI cannot be governed as a single thing. It must be governed as a stack. The AI OSI Stack provides a map. It separates hardware from models, models from training, training from control, control from interface, interface from application, and application from trust. That separation allows responsibility to be assigned and risks to be mitigated where they actually arise.

This also means your existing work fits directly into it. Persona based safety, decision insurance, guardian notes, and auditable decision cards all live at the top layer. They make trust portable. They turn governance from a static checklist into a living practice.

If AI is to become infrastructure for human life, then it must be designed like infrastructure: layered, transparent, and accountable.

About the Author

Daniel P. Madden is an independent AI researcher and IT specialist focused on feasibility, governance, and reasoning systems. His AI Lab Notebook documents live experiments in persona architecture, decision auditing, and epistemology by design. See https://danielpmadden.com for related work.

References

[1] National Institute of Standards and Technology. 2023. *AI Risk Management Framework (AI RMF 1.0).*

[2] ISO/IEC 42001:2023. *Artificial Intelligence Management System.*

[3] European Commission. 2024. *Regulation of the European Parliament and of the Council Laying Down Harmonised Rules on Artificial Intelligence (AI Act).*

[4] International Organization for Standardization. 1994. *Information technology – Open Systems Interconnection – Basic Reference Model (ISO/IEC 7498-1).*

[5] Partnership on AI. 2024. *Guidance for Safe Foundation Model Deployment.*