AI does not scale like software.

It scales like industry.

A structural view of the constraints and dependencies behind AI

Igor Allinckx – May 2026

Contents

Preface

This white paper is part of a broader effort to reshape how leaders think about AI, responsibility, and judgment. Much of today's discourse treats AI as a technical capability or a collection of use cases. But leadership does not operate at the level of features. It operates at the level of systems, constraints, and consequences.

My work (from The AI Glass Maze to board-level governance insights) argues that responsible leadership begins with seeing reality clearly. AI introduces new forms of opacity and new distortions. To govern it, leaders must understand not only what AI can do, but what it requires: the industrial, geopolitical, and organisational structures that make it possible.

This document provides that foundation. It exposes the physical and economic architecture behind AI, the bottlenecks that shape its trajectory, and the control layers that determine who can scale and who cannot. It is designed to support better judgment in boardrooms, more grounded decision-making, and a more mature understanding of responsibility in an AI-shaped environment.

AI is not only a technological shift. It is a test of leadership. And leadership begins with literacy. Not in algorithms, but in the system that surrounds them.

Executive Summary

AI is often described as a software breakthrough, but its foundations are industrial. Behind every model lies a global system of materials, chips, energy, data centers, networks, capital, and standards. This system is unevenly distributed, structurally constrained, and increasingly shaped by geopolitics. For leaders, understanding this architecture is not technical detail but indeed strategic AI literacy.

AI does not scale like software. It scales through physics: rare materials, advanced lithography, multi-billion-dollar data centers, and gigawatts of energy. A handful of chokepoints and hyperscalers determine who can access compute, at what cost, and at what pace. Models attract attention, but value and control sit elsewhere: in bottlenecks, infrastructure, and interfaces.

This white paper maps the real AI supply chain and value architecture. It shows where power concentrates, where dependencies emerge, and what leaders must understand to make informed, responsible, and resilient decisions in an AI-shaped environment.

Introduction: Why Leaders Must Understand the AI Supply Chain

AI is moving rapidly into every industry, yet most leadership discussions still focus on models, use cases, and vendor selection. What is often missing is a clear understanding of the *industrial system* that makes AI possible: the materials, chips, energy, infrastructure, and capital flows that sit behind every model.

For boards and executives, this is not technical detail. It is **strategic AI literacy**.

Understanding the AI supply chain enables leaders to:

- **separate hype from reality** by seeing the physical and economic limits of AI scaling
- **anticipate dependencies and disruptions** that may affect operations, continuity, and cost
- **avoid structural lock-in** to vendors, regions, or infrastructures they do not control
- **evaluate risk exposure** across compute, energy, data, and geopolitics
- **govern AI responsibly** in an environment where capacity is unevenly distributed

AI is often presented as frictionless and infinitely scalable. In reality, it depends on a global network of bottlenecks, chokepoints, and control layers. Leaders who understand this system can make informed decisions and avoid building strategy on infrastructure they do not influence.

This document provides a clear, factual view of the AI supply chain and value architecture. It is not written for engineers, but for the people who must steer organisations through the next decade.

To understand why AI cannot be treated like traditional software, we must first correct the core misconception that shapes most executive discussions. The limits of AI are not digital. They are physical, economic, and geopolitical. This is where the real misunderstanding begins.

1. The Great Misunderstanding: AI Does Not Scale Like Software

Most organisations still assume AI scales like traditional software. This assumption is structurally wrong and strategically dangerous.

AI scaling is constrained by physics, capital, and geopolitics. Software scaling is not.

The first distinction leaders must grasp is the difference between software scaling and AI scaling. They operate under entirely different laws.

1.1 Software scales with code. AI scales with physics.

Software grows through replication. AI grows through hardware, energy, cooling, and land.

· elastic cloud vs fixed physical limits

· near-zero marginal cost vs rising marginal cost

· containers vs substations

Once we move beyond the software mindset, the physical constraints of AI become visible. These limits are not theoretical and in fact are already shaping global capacity.

1.2 Hard physical limits

AI scaling hits constraints that cannot be abstracted away:

· GPU availability

· grid capacity

· cooling density

· water access

· fiber bandwidth

· permitting timelines

Physical limits are only part of the story. AI also requires unprecedented levels of capital investment, far beyond anything seen in traditional IT.

1.3 Capital intensity

AI scale is built on heavy CAPEX:

· data centers: multi-billion

· substations (high-voltage grid facilities): hundreds of millions

· fabs (semiconductor fabrication plants): tens of billions

· EUV (Extreme Ultraviolet Lithography) machines: $150–300M

These capital requirements concentrate power in a few chokepoints. AI's industrial base depends on actors that cannot easily be replaced.

1.4 Bottleneck-driven

AI depends on a few irreplaceable actors:

· ASML (lithography)

· TSMC (advanced nodes)

· NVIDIA (AI compute)

· export controls

· grid interconnection queues

These chokepoints do not exist in a vacuum. They sit inside geopolitical blocs that increasingly determine who can access advanced compute.

1.5 Geopolitical dependency

AI capacity is increasingly shaped by geopolitical blocs, each exerting influence through different forms of power:

· **US-aligned compute:** access to leading-edge chips, hyperscaler infrastructure, and the export-control regime that governs who can train at frontier scale.

· **China-aligned compute:** a parallel ecosystem spanning chips, clouds, standards, and domestic supply chains designed to reduce exposure to Western controls.

· **EU regulatory power:** not a compute superpower, but a *standards* superpower. Through GDPR, the AI Act, data-adequacy rules, and cloud sovereignty requirements, the EU shapes how data moves, how AI is deployed, and what compliance looks like globally.

· **Non-aligned states:** India, Brazil, UAE, Singapore and others that negotiate access with both sides, balancing compute availability, cost, and geopolitical exposure.

These blocs do not compete on the same axis: the US dominates compute, China dominates integrated industrial capacity, the EU dominates regulation, and non-aligned states navigate between them. Together, they define the global conditions under which AI can be trained, deployed, and scaled.

1.6 Multi-layer interdependence

Finally, AI is not one system but many interdependent systems. It requires simultaneous scaling of materials, chips, energy, data centers, networks, data, talent, and governance. If one layer stalls, the entire system stalls.

With the misconception corrected, we can now examine the real AI supply chain. It is a tightly coupled industrial system with circular dependencies and control layers.

2. The Real AI Supply Chain

AI is not a linear chain but a system of interdependent layers. Each layer introduces constraints, dependencies, and control points. Together they form the industrial backbone of modern AI.

The supply chain begins with the most basic layer: the materials that make computation possible.

2.1 Extraction & Refining

AI begins with materials but control sits in refining, not mining. The world extracts rare earths in many places, but transforms them in very few.

· silicon, rare earths, gallium, germanium, cobalt, lithium

· extraction: China, Australia, DRC, Brazil

· refining: overwhelmingly China

Insight: Transformation, not extraction, defines control.

From raw materials, the chain moves into one of its most strategic layers: semiconductors.

2.2 Semiconductors

Chips are the physical brain of AI. Their production is a global choreography of design, lithography, and fabrication … with chokepoints at every step.

· design: NVIDIA, AMD, ARM

· lithography: ASML (EUV monopoly)

· foundries: TSMC, Samsung

· logistics: air transport for high-value chips

Insight: No actor controls the full chain → interdependence = power + fragility.

Once chips exist, they must be assembled into systems capable of operating at scale.

2.3 Assembly & System Integration

Once chips exist, they must be assembled into functional systems. This is where engineering choices determine performance, efficiency, and cost per token.

· motherboards, cooling, power, chassis

· assembly: Foxconn, Quanta, Supermicro

· integration: thermal design, GPU interconnects, cluster architecture

Insight: System integration defines real-world performance.

These systems require a physical home. Data centers are where AI becomes real and where its constraints become unavoidable.

2.4 Data Centers

Data centers are the physical homes of AI and the primary constraints on its growth. They depend on energy, cooling, land, and regulation.

· locations (actual major ones): Virginia, Dublin, Singapore

· constraints: energy (hard limit) – cooling (physical limit) – geography + regulation (political limit)

Insight: Scaling AI is increasingly an energy and location problem.

Inside these facilities, data flows and model operations define the computational life of AI.

2.5 Data Flows & Model Layer

Models depend on global data flows and massive compute cycles. Training is episodic; inference is continuous and exponential.

· submarine cables: the physical backbone of global data flow; any disruption directly affects training pipelines and real-time inference delivery

· training: months of compute

· inference (users): real-time, exponential load

Insight: Inference, not training, is the real scaling challenge.

Once models produce outputs, they must be delivered to users through global distribution networks.

2.6 Distribution

AI outputs travel through Content Delivery Networks (CDNs) and Internet Service Providers (ISPs). These layers indirectly control access, latency, and the reliability of AI-driven services.

· **CDNs:** optimise delivery by caching outputs closer to users, but create regional performance asymmetries and introduce dependency on a small number of global providers.

· **ISPs:** shape the actual user experience through routing, congestion management, and peering agreements that organisations do not control.

· **Regional bandwidth constraints:** limit the speed and consistency of AI deployment, especially for inference-heavy applications that require low-latency access.

Insight: Together, these distribution layers form the last mile of AI infrastructure. They are largely invisible, yet decisive in determining who can access AI capabilities, at what speed, and under which constraints.

Behind all these layers lies a resource that cannot be automated: talent.

2.7 Talent

Talent is the invisible but decisive layer. It is scarce, clustered, and mobile.

· researchers

· infrastructure engineers

· system architects

Insight: Talent follows capital, infrastructure, and projects ... not geography.

Talent alone is not enough. AI scale depends on capital allocation and governance decisions made by a small number of actors.

2.8 Capital & Governance

Capital determines what becomes real. Governance determines who gets access.

· **hyperscalers** (AWS, Azure, Google Cloud): They operate global cloud infrastructures capable of running AI at industrial scale.

They provide:

- massive GPU clusters
- global data centers
- high-density networking
- energy procurement
- security, compliance, and orchestration layers

They are called hyperscalers because they operate at a scale no other organisations can match (tens of millions of servers, hundreds of data centers, and multi-gigawatt energy footprints)

· **secondary**: Alibaba, Tencent, Oracle

· **states** and sovereign funds

Insight: Hyperscalers sit between chips and models → They are structural gatekeepers.

Beyond physical infrastructure, software frameworks and interfaces shape how AI is built and accessed.

2.9 Standards & Software Ecosystem

Frameworks and APIs define how AI is built and accessed. They are soft control layers with hard consequences.

· PyTorch, TensorFlow (dominant machine-learning frameworks used to build, train, and deploy AI models). Most large-scale AI systems (OpenAI, Meta, Google, Anthropic) rely on one of them

· APIs (Application Programming Interfaces). They create *platform lock-in* because switching providers requires rewriting integrations. APIs are the main monetisation layer for model companies and hyperscalers

· platform ecosystems

Insight: Control operates through interfaces.

Across all these layers, one structural feature creates persistent tension: time asymmetry.

2.10 Time Asymmetry

Time asymmetry is one of the most overlooked constraints in AI: the physical layers scale over years, while the digital layers evolve in months or seconds.

· chips: years

· data centers: years

· models: months

· inference: real-time

Insight: Model innovation outpaces the physical capacity needed to run it. This creates bottlenecks, shortages, and strategic dependency on actors who can scale infrastructure fast enough.

The visible supply chain tells only half the story. Several critical layers operate behind the scenes, shaping constraints and determining who can scale.

3. The Invisible Control Layers

These layers are part of the supply chain but operate differently. They shape constraints, dependencies, and strategic exposure. The first invisible layer is also the most fundamental: energy.

3.1 Energy Supply Chain

Energy is becoming the primary bottleneck for AI scale. Compute growth now outpaces grid growth.

· grid capacity

· renewable PPAs: (Power Purchase Agreements) Long-term contracts where hyperscalers buy electricity directly from renewable producers (solar, wind).

· nuclear (SMRs): Small Modular Reactors. Compact nuclear reactors designed for industrial sites. Hyperscalers are exploring SMRs to secure stable, high-density power for future AI clusters.

· water rights

· transformers

· substations

Insight: AI scale is energy scale and it is increasingly limited. PPAs and SMRs are emerging as strategic assets.

The second invisible layer is data (once abundant, now regulated, monetised, and contested).

3.2 Data Supply Chain

Data used to be treated as abundant and freely accessible. This is no longer the case as illustrated by numerous examples:

Reddit now charges for API access to its content. X restricts data scraping and sells access tiers. News publishers (e.g., The New York Times) are suing model developers over training data. YouTube and LinkedIn explicitly prohibit training on their content. EU data laws (GDPR, DMA) restrict how data can be moved and processed.

Data flow is increasingly limited by:

· licensing markets (data bought or licensed for training)

· data localization laws (local rules requiring data to stay within borders)

· copyright constraints

· proprietary corporate data (valuable training data, but locked inside organisations)

· synthetic data pipelines (AI-generated data used to fill gaps)

Insight: Data is shifting from an open resource to a regulated commodity.

A third invisible layer is emerging rapidly: safety and alignment.

3.3 Safety & Alignment Supply Chain

Safety is becoming a mandatory production step.

· evaluations (tests measuring model behaviour and risks before deployment)

· red-teaming (adversarial testing to find failures)

· interpretability tools (methods to understand model reasoning)

· safety-tuned models (models adjusted to reduce harmful behaviour)

· compliance frameworks (EU AI Act, sector rules, audits)

Insight: Safety is no longer optional but becomes operational.

Finally, the entire system is shaped by geopolitical alignment, which increasingly determines access, rules, and strategic exposure.

3.4 Geopolitical Blocs

AI capacity is shaped by geopolitical alignment.

· US-aligned compute

· China-aligned compute

· EU regulation efforts

· non-aligned states

The supply chain is fragmenting into blocs with different rules, access rights, and dependencies.

Effects:

· Export controls determine who can buy advanced chips.

· Cloud sovereignty laws restrict where models can be trained or deployed.

· EU regulatory power shapes global AI rules and data flows.

· Alliances (US–Japan–Netherlands) control lithography and fabrication.

· China builds a parallel ecosystem (chips, clouds, standards).

· Non-aligned states (India, Brazil, UAE) negotiate access with both sides.

Insight:
Geopolitics now determines who can scale AI (and who cannot).
Understanding the supply chain allows us to see where value actually accumulates. It does not follow the chain evenly but concentrates in specific leverage points.

4. The Real AI Value Architecture

Value does not follow the chain. It concentrates in specific zones where substitution is hardest and control is strongest. Each layer plays a different role in shaping who captures value and who becomes dependent.

The first value zone sits at the bottom of the chain: essential but low-margin layers.

4.1 Low-Value Essential Layers

(mining, basic processing)

These layers provide the raw materials that make AI possible. They are essential but heavily commoditized, with global competition and low margins. Their importance is structural, not strategic.

· mining, refining, basic processing
· globally distributed, price-driven
· limited differentiation

Insight: Essential, but not where power sits.

Above them lie the true chokepoints: the layers with the highest strategic leverage.

4.2 High-Value Chokepoints

(ASML, TSMC, NVIDIA)

These layers have extremely high barriers to entry and almost no substitutes. They define the pace, cost, and feasibility of AI scaling. Their leverage comes from scarcity and specialization.

· ASML: EUV lithography monopoly
· TSMC: advanced node fabrication
· NVIDIA: AI compute dominance

Insight: Value concentrates where substitution is hardest.

Between chokepoints and infrastructure sit competitive layers with thin margins.

4.3 Thin-Margin Competitive Layers

(assembly, integration)

These layers sit between chokepoints and infrastructure. They require engineering excellence but face intense competition and margin pressure. Efficiency matters more than differentiation.

· assembly (Foxconn, Quanta, Supermicro)

· system integration (thermal, interconnects, cluster design)

· cost-driven, operationally intensive

Insight: Necessary for performance, but not a source of durable advantage.

The largest value capture occurs not in chips or models, but in infrastructure.

4.4 Massive Value Capture Layer

(hyperscalers)

Hyperscalers convert massive CAPEX into recurring revenue. They own the infrastructure everyone else depends on, creating structural lock-in and continuous value capture. Their position between chipmakers and model developers gives them systemic influence.

· they own the data centers and GPU clusters
· everyone else rents capacity
· switching costs are extremely high
· utilisation increases margins
· control over energy, networking, orchestration

Insight: Infrastructure converts CAPEX into continuous value capture.

Next, Model companies. They attract attention, but their position in the value chain is less stable than it appears.

4.5 Visible but Unstable Layer

(Models)

Models are the most visible part of the AI ecosystem and possibly the most misunderstood. They attract attention, investment, and media coverage, but their position in the value chain is structurally fragile.

Model performance improves rapidly, competitors release alternatives within months, and open-source models compress differentiation. As a result, the model layer faces constant pressure from both commoditization and substitution.

· high visibility, but limited defensibility

· rapid competition across closed and open-source ecosystems

· short innovation cycles that erode long-term advantage

· dependence on upstream compute and downstream distribution

Insight: Models matter, but they do not control the system. Their value is real but unstable because they sit between stronger layers: compute infrastructure above them and distribution interfaces below them.

Control increasingly shifts to the interfaces through which AI is accessed.

4.6 Access & Distribution Layer

(APIs, platforms)

APIs and platforms determine how AI is accessed and integrated. They create structural lock-in because once organisations build on a specific interface, switching becomes costly. This layer quietly shapes monetization and dependency.

· APIs define usage patterns
· platforms control identity, data flow, billing
· switching requires rewriting systems
· distribution becomes a leverage point

Insight: Control over access is control over usage.

Finally, several layers capture value indirectly by shaping everything else.

4.7 Hidden Dominant Layers

(talent, capital, standards)

These layers do not generate revenue directly but determine the entire system's direction. They influence what gets built, who can build it, and under which rules. They shape the boundaries of innovation and the distribution of power.

· talent concentration drives breakthroughs
· capital allocation determines what scales
· standards define interoperability and compliance

Insight: These layers govern the system from above.

Across the entire system, a clear pattern emerges. Value and control concentrate in three types of positions.

5. The Key Pattern

Across the entire system, value and control do not distribute evenly. They concentrate in a few structural positions that shape how the whole ecosystem behaves:

1. **Bottlenecks** (ASML, TSMC, NVIDIA)
2. **Control points** (hyperscalers)
3. **Interfaces** (APIs, platforms)

These positions are not accidental and emerge from the physics, economics, and dependencies described in the previous sections.

1. Bottlenecks (ASML, TSMC, NVIDIA) These actors control capabilities that cannot easily be substituted. Their constraints define the pace, cost, and feasibility of AI scaling. When bottlenecks move, the entire system moves.

2. Control points (hyperscalers) Hyperscalers sit between chips and models, converting infrastructure into recurring revenue and dependency. They determine who gets access to compute, under which conditions, and at what scale.

3. Interfaces (APIs, platforms) Interfaces shape how AI is consumed. They create lock-in, define integration patterns, and quietly determine which capabilities become widely adopted.

Together, these three positions form the architecture of power in AI. They explain why some actors capture disproportionate value, why others remain dependent, and why leadership must understand the system behind the technology. Because decisions made at these points ripple across the entire ecosystem.

These structural realities have direct implications for leadership, governance, and strategic decision-making.

6. Implications for Leadership & Governance

AI's industrial nature reshapes strategic decision-making. Leaders must understand how supply-chain constraints, infrastructure dependencies, and control layers translate into organisational risk and responsibility.

· compute becomes a strategic asset

· energy becomes a limiting factor

· hyperscalers become systemic actors

· regulation becomes a supply-chain constraint

· responsibility becomes harder to locate

· AI strategy becomes industrial strategy

Taken together, these insights redefine how leaders should think about AI: not as software, but as an industrial system with strategic consequences.

6.1 You are dependent on capacity you do not control

Access to AI depends on:

- compute availability
- hyperscaler infrastructure
- energy and location constraints

These are controlled by external actors.

Consequence:
Your AI strategy is partly determined outside your organisation.

In practice:

- timelines shift
- costs move
- capabilities appear or disappear

6.2 The system can change without you noticing

Models, infrastructure, and data pipelines evolve continuously:

- model updates
- API changes
- infrastructure reallocation

Performance may remain stable ... until it doesn’t.

Consequence:
Stability is not guaranteed, even when nothing changes on your side.

In practice:

- outputs drift
- behaviour shifts
- decisions are affected without a visible trigger

6.3 Scaling is constrained by physics, not ambition

AI does not scale on demand:

- compute is finite
- energy is limited
- deployment takes time

Consequence:
Not everything that works at small scale can be deployed at large scale.

In practice:

- pilots succeed, rollout stalls
- costs rise non-linearly
- capacity becomes a strategic constraint

6.4 Critical parts of the system sit outside your accountability perimeter

You rely on:

- external infrastructure
- external models
- external interfaces

But decisions based on these systems remain yours.

Consequence:
There is a gap between where the system is controlled and where responsibility sits.

In practice:

- you cannot fully explain how an output was produced
- you cannot fully control how the system evolves
- but you remain accountable for the decision taken

Closing Statement

AI is often framed as a software revolution, but its foundations are **industrial**. It depends on **materials, chips, energy, data centers, talent, and geopolitical alignment**. These layers introduce **constraints, bottlenecks, and dependencies** that shape what organisations can realistically build, deploy, and sustain.

For leaders, this changes the nature of decision-making. You operate in a system where **capacity is external**, **stability is uncertain**, **scaling is constrained by physics**, and **responsibility sits inside your organisation even when control does not**. This is the practical reality of governing AI in an environment defined by **dependency and asymmetry**.

Understanding the industrial system behind AI does not eliminate these tensions. It **clarifies** them. It gives leaders the ability to judge **what is feasible**, to anticipate **where risks originate**, and to take **responsibility within the limits of what can be controlled**.

AI strategy is now industrial strategy. And the organisations that recognise this (and build their governance on **reality rather than assumptions**) will navigate the next decade with **clarity, resilience, and authority**.

Igor Allinckx

AI & Governance

Lucerne, May 2026

Note

This paper is part of the AI & Governance Insights series

www.allinckx.com/insights