

The AI Value Chain Is Moving

From models and answers to agents, inference, memory, power—and measurable enterprise outcomes.

The model is an ingredient. The product is trusted work completed.

Jami Kiran

Trusted Global Technology Advisor

AI · Data · Cybersecurity · Cloud · Quantum

Independent analysis inspired by the 20VC conversation with Perplexity CEO Aravind Srinivas. No affiliation or endorsement is implied.

The next AI advantage will be built across an entire system

AI competition is shifting from who owns the most impressive model to who can convert intelligence into reliable action at sustainable cost.

What is changing

- Answers are becoming abundant; completed work remains scarce.
- Model capability is diffusing while orchestration, proprietary context, trust and distribution differentiate.
- Inference—not only training—is becoming a recurring operational economy.
- Memory, networking, power and permitting increasingly shape digital strategy.

What leaders must do

- Start from workflows and customer outcomes, not model fascination.
- Measure unit economics per successful task.
- Design human accountability into every consequential process.
- Treat compute, energy, data and policy as one strategic portfolio.

Jami Kiran’s governing principle

Be ambitious in direction, disciplined in evidence and uncompromising on trust. Attack the opportunity—not the controls that keep the enterprise safe.

Source-to-strategy translation

Interview insight	Executive interpretation	Decision consequence
"The model is not the product."	Capability alone does not create adoption or value.	Fund workflow integration, context, UX and assurance.
Agents are the frontier.	Value moves from information retrieval to delegated execution.	Measure completion, reliability, recovery and supervision.
Power is a bottleneck.	AI strategy now has a physical infrastructure boundary.	Include energy and capacity scenarios in portfolio planning.

Where durable value can accumulate

1. Models Reasoning and generation	2. Orchestration Routing, tools and memory	3. Context Enterprise data and permissions	4. Action Workflow execution	5. Assurance Evaluation and governance
6. Inference Tokens, latency and serving	7. Silicon Accelerators and HBM	8. Facilities Networks, cooling, power	9. Distribution Habit and customer access	10. Outcomes Revenue, cost, risk, speed

The strategic test

For each layer ask: Is it differentiating, constraining, regulated, scarce, or better sourced from a partner? The answer determines where to own, where to orchestrate and where to buy.

Own Proprietary context, critical workflow design, risk policy, customer relationship and outcome data.	Orchestrate Multiple models, tools, retrieval systems, identity, observability and human review.	Buy Commodity foundation capability and infrastructure where scale economics dominate.
---	--	--

Strategic warning: A spectacular demo can conceal weak unit economics. A cheap model can become expensive when retries, long context, supervision and failed actions are counted.

Move from “AI that speaks” to “AI that completes”

An agentic system observes context, plans a sequence, uses tools, checks state and either completes the objective or escalates. The commercial unit is therefore not a token or response—it is a **verified successful task**.

Measure	Why it matters	Executive threshold
Task completion rate	Tests whether the workflow actually finishes.	Defined by use-case risk; never averaged across unlike tasks.
Cost per successful task	Includes retries, tools, inference and review.	Below risk-adjusted human or current-process cost.
Exception and recovery rate	Reveals brittleness hidden by happy-path demos.	Known, monitored and improving.
Human supervision minutes	Prevents “automation” that merely relocates labour.	Falling without degradation in safety or quality.
Cycle-time reduction	Captures customer and working-capital value.	Material at end-to-end process level.
Policy violation rate	Tests access, compliance and behavioural controls.	Near zero for consequential actions.

Reliability ladder

Assist → human executes | **Recommend** → human approves | **Act within bounds** → human monitors | **Autonomous** → system acts and is audited.

Advance only when evidence, reversibility and accountability justify the next level.

Research cautions against assuming broad autonomy: enterprise-agent benchmarks continue to show steep performance declines as tasks become longer, cross-application and permission-sensitive. The winning architecture is bounded autonomy with observable state—not blind delegation.

Inference, memory and power become board-level variables

Inference compounds

Training creates a model; inference operates it every time a customer or agent asks it to reason. Agentic workflows multiply calls through planning, tool use, validation and retries.

The memory wall

During token generation, model weights and growing context must move rapidly between high-bandwidth memory and processors. This makes memory capacity and bandwidth economically consequential—not a secondary component.

Power closes the loop

Servers need electricity, cooling, grid connections and permits. The IEA projects global data-centre electricity consumption to more than double to about 945 TWh by 2030 in its 2025 base case.

Unit-economics equation

Cost / successful task

= model calls
 + token and tool cost
 + compute and memory
 + data retrieval
 + human review
 + failures and retries
 + security and compliance
 + allocated platform cost

Optimize the system, not one line item.

Capacity decisions by horizon

0–12 months

Routing, caching, smaller models, prompt/context efficiency, observability.

1–3 years

Reserved capacity, multi-provider resilience, data locality, energy-aware scheduling.

3–7 years

Facility location, grid partnership, sovereign infrastructure and supply-chain strategy.

Build the business case around outcome evidence

A credible AI case connects an observable workflow change to financial value, then discounts it for adoption, reliability and risk.

Annual risk-adjusted value

(Revenue + Cost + Risk + Speed) × Adoption × Reliability

Net value = risk-adjusted value – operating cost – change cost – control cost

ROI = net value ÷ total investment | Payback = initial investment ÷ monthly net benefit

Value lever	Evidence	Example metric
Revenue	Controlled uplift or attributable conversion	Incremental gross margin
Cost	Hours actually removed or capacity redeployed	Cost per completed case
Risk	Reduced frequency or severity, not fear-based estimates	Expected loss avoided
Speed	End-to-end cycle improvement	Days to decision / cash
Learning	Reusable data, evaluations and process knowledge	Time to launch next use case

No false precision. Publish base, downside and upside scenarios. State which numbers are observed, benchmarked or assumed. Assign an owner to every assumption.

Choose workflows where value and controllability intersect

Dimension	Question	Score (1–5)
Economic materiality	Is the volume, margin, delay or risk meaningful?	—
Process clarity	Can success, failure and escalation be defined?	—
Data readiness	Is authorized, current context accessible?	—
Technical fit	Does language/reasoning/tool use address the bottleneck?	—
Verifiability	Can outputs and actions be checked cheaply?	—
Reversibility	Can errors be contained or rolled back?	—
Adoption feasibility	Will users change the workflow?	—
Strategic learning	Does it build reusable capabilities?	—

Good early candidates

- High-volume research and synthesis
- Draft–review workflows
- Service triage with clear escalation
- Developer and analyst assistance
- Internal operations with reversible actions

Poor early candidates

- Ambiguous ownership
- Irreversible high-impact actions
- Unobservable success
- Broken processes disguised as AI opportunities
- No lawful access to required data

A 90-day path from conviction to controlled value

Days 0–15 Map value pools, constraints and baseline metrics.	Days 16–30 Select two workflows; define controls and evaluation set.	Days 31–60 Build bounded pilots; instrument cost and quality.	Days 61–75 Run with users; test failure, recovery and adoption.	Days 76–90 Decide: stop, redesign, expand or scale.
--	--	---	---	---

Minimum cross-functional team

Business owner Owns outcome and adoption.	Product/process lead Redesigns the workflow.	AI/platform lead Owns architecture and cost.
Data steward Owns context quality and rights.	Risk/security Sets boundaries and assurance.	Finance partner Validates benefits and spend.

Five gates

Business fit	Technical proof	Control proof	Economic proof	Adoption proof
--------------	-----------------	---------------	----------------	----------------

Scale only when all five gates pass. Technical success without adoption or economics is not enterprise success.

Export controls change incentives—but outcomes are adaptive

The interview argues that restrictions can catalyse domestic substitution and accelerate local innovation. This is a plausible strategic mechanism, not a settled causal conclusion. Firms should avoid simplistic “controls work” versus “controls fail” narratives.

First-order effect

Restricted access raises cost, delay and uncertainty for targeted technologies.

Second-order response

Investment moves toward substitutes, stockpiles, redesign, domestic capacity and alternative supply.

Third-order consequence

Standards, ecosystems, trade relationships and innovation paths may diverge.

Enterprise response

Scenario-plan across policy, geography, supplier concentration and data sovereignty.

Resilience checklist

- Map dependency on accelerators, HBM, networking, cloud regions and energy infrastructure.
- Identify single points of failure and contractual concentration.
- Test export classification, end-use, data-transfer and sanctions obligations with qualified counsel.
- Maintain portable evaluation sets and architecture options to reduce lock-in.
- Separate public claims from verified operating assumptions.

Leadership wisdom: Regulation is not merely friction; it reshapes the field. Strategy must anticipate how every participant adapts.

AI changes the shape of work before it determines the number of jobs

Evidence is heterogeneous. A field study of 5,179 customer-support agents reported an average 14% productivity gain, concentrated among less-experienced workers. A later experiment across 66 firms and 7,137 knowledge workers found time savings—about two fewer hours of email per week among users—but no detectable broad change in task composition during the study window.

Near-term redesign

- Shift from production to verification
- Compress research and drafting cycles
- Codify expert practice for wider access
- Create agent supervision and evaluation roles

Leadership obligations

- Measure workload, quality and wellbeing
- Train judgment, not only prompting
- Preserve learning paths for junior talent
- Share gains and explain workforce choices

The smaller-team hypothesis

AI may allow future firms to reach significant scale with fewer employees, but headcount is not the objective function. The better objective is **value per responsible decision**. Thin organizations can become fragile if institutional knowledge, challenge and accountability disappear.

Do not automate away the apprenticeship system that creates tomorrow's experts.

Twelve questions leaders should answer now

1. Which customer or operating outcomes could materially improve?
2. Where is work constrained by analysis, coordination or delay?
3. Which actions are safe to delegate—and under what limits?
4. What is the baseline cost, quality and cycle time?
5. What constitutes a verified successful task?
6. What proprietary context creates advantage?
7. How will identity and permissions follow the agent?
8. What are the full inference and supervision economics?
9. What memory, compute, energy or supplier constraints matter?
10. What happens under a policy or provider shock?
11. How will roles, learning and accountability change?
12. What evidence must exist before the next investment gate?

The enterprise that learns fastest—without losing trust—will outperform the enterprise that merely spends fastest.

Jami Kiran's executive position

AI is not a software purchase. It is a redesign of how information becomes judgment, how judgment becomes action and how action becomes accountable value. Leaders should pursue it with the confidence of an attacker, the discipline of an engineer and the responsibility of a trustee.

Evidence base

This white paper is an independent synthesis. Statements attributed to interview participants are perspectives, not verified facts unless separately supported. Company valuations and revenue figures mentioned in the video are time-sensitive and are not treated here as audited evidence.

1. 20VC with Harry Stebbings (2026). "Perplexity CEO: Micron Will Be More Valuable Than Meta & How Export Controls Helped Not Hurt China." YouTube, 15 June 2026. <https://youtu.be/OxFyVcO1Yow>
2. International Energy Agency (2025). *Energy and AI*. <https://www.iea.org/reports/energy-and-ai/>
3. Stanford Institute for Human-Centered AI (2025). *AI Index Report 2025*. <https://hai.stanford.edu/ai-index/2025-ai-index-report>
4. Brynjolfsson, E., Li, D., & Raymond, L. R. (2023). "Generative AI at Work." NBER Working Paper 31161. <https://doi.org/10.3386/w31161>
5. Dillon, E. W., Jaffe, S., Immorlica, N., & Stanton, C. T. (2025). "Shifting Work Patterns with Generative AI." NBER Working Paper 33795. <https://doi.org/10.3386/w33795>
6. METR (2026). "Task-Completion Time Horizons of Frontier AI Models." <https://metr.org/time-horizons/>
7. U.S. Bureau of Industry and Security (2024–2025). Official advanced-computing and semiconductor export-control releases. <https://www.bis.gov/>
8. OECD (2025). "AI and Work." <https://www.oecd.org/en/topics/ai-and-work.html>

Authorship and authenticity

Author: Jami Kiran, Trusted Global Technology Advisor.

Original contribution: executive synthesis, value-migration model, decision framework, commercial equations, portfolio scorecard and 90-day operating model.

Disclaimer: Educational and strategic material only; not investment, legal, employment or export-control advice.