CONCEPTUAL RESEARCH PAPER · PERSPECTIVE ARTICLE · 2026

Value Migration in the Agentic AI Economy

A conceptual framework linking product architecture, inference economics, memory bandwidth, energy constraints, geopolitics and organizational design

Jami Kiran
Independent Global Technology Advisor
AI · Data · Cybersecurity · Cloud · Quantum
QuantumReadyCheck.com

Article type: Conceptual synthesis / perspective paper
Status: Author manuscript; not peer reviewed
Original contribution: The Agentic Value Migration Framework and six testable research propositions

Intellectual stimulus: a 2026 public interview with Aravind Srinivas, Founder and CEO of Perplexity, conducted by Harry Stebbings for 20VC. Interview opinions are treated as propositions requiring triangulation, not as empirical proof.

ABSTRACT

Abstract and keywords

Purpose. This paper explains how economic value in artificial intelligence (AI) may migrate from foundation-model capability toward agentic execution, inference infrastructure, high-bandwidth memory, power access, workflow integration and institutional trust.

Method. A structured thematic analysis was applied to a 95-minute executive interview published by 20VC in June 2026. Claims were coded into six domains—product, agents, infrastructure, industrial policy, labour and organization—and triangulated against energy-system analysis, AI-industry measurement, field studies of workplace AI, agent evaluations and official semiconductor policy documents.

Findings. The synthesis suggests that model capability is necessary but increasingly insufficient for durable enterprise value. A system's commercial performance depends on the probability of successful task completion, full inference and supervision cost, access to memory and power, proprietary organizational context, and governance of delegated action. Constraints can relocate value upstream toward scarce complements. Policy restrictions may also generate adaptive investment and substitution, making outcomes nonlinear.

Contribution. The paper introduces an Agentic Value Migration Framework, a risk-adjusted task-economics function and six propositions for empirical testing. It distinguishes interview-derived hypotheses from independently supported evidence.

Keywords: agentic AI; inference economics; foundation models; high-bandwidth memory; data-centre electricity; organizational design; export controls; value migration; AI governance.

Suggested citation: Kiran, J. (2026). "Value Migration in the Agentic AI Economy: A Conceptual Framework Linking Product Architecture, Inference Economics, Memory Bandwidth, Energy Constraints, Geopolitics and Organizational Design." Author manuscript.

1. INTRODUCTION

From model competition to system competition

The first commercial phase of generative AI emphasized model scale, benchmark performance and conversational access. The emerging phase emphasizes whether a system can perform economically useful work: retrieve authorized context, plan, use tools, interact with software, verify state, recover from failure and remain accountable.

This transition changes the unit of analysis. A model produces tokens; an enterprise purchases outcomes. Between these sits a socio-technical stack of orchestration, retrieval, identity, tools, evaluation, human judgment and physical infrastructure. Improvements at one layer can increase load at another. Cheaper inference may accelerate adoption; greater adoption may raise total power demand. More capable agents may reduce execution time but increase the need for control and verification.

The motivating interview presents strong claims: models are not products; agentic work is the economic frontier; inference businesses can become extremely valuable; memory and electricity are binding constraints; export controls can strengthen domestic substitutes; and AI-native firms may scale with fewer employees. These claims are analytically productive but vary in evidentiary status.

This paper therefore asks three research questions. **RQ1:** Through which mechanisms does AI value migrate across the technical and organizational stack? **RQ2:** How do inference, memory and energy constraints affect product economics? **RQ3:** How should firms evaluate agentic automation under uncertainty, incomplete reliability and policy change?

The contribution is not a forecast of individual company valuations. It is a framework for locating scarce complements, measuring task-level economics and converting provocative executive observations into testable propositions.

Central thesis. As model capability diffuses, durable value increasingly depends on scarce complements that convert probabilistic intelligence into verified, governed and economically sustainable action.

Structured thematic synthesis and triangulation

The study uses a qualitative conceptual-synthesis design. It does not claim causal identification, statistical generalization or original experimental data.

Stage	Procedure	Output
1. Corpus definition	Public 20VC interview, 95:16 duration, published 15 June 2026; automated English transcript reviewed thematically.	Primary intellectual stimulus.
2. Open coding	Statements grouped around product differentiation, agents, inference, memory, power, export controls, work and organizational scale.	Six thematic domains.
3. Claim classification	Separated opinion, mechanism, prediction and externally verifiable factual claim.	Evidence-status discipline.
4. Triangulation	Compared mechanisms with IEA, Stanford HAI, NBER field studies, METR, OECD, BIS and technical literature.	Supported, qualified or unresolved claims.
5. Theory building	Integrated themes using complementary-assets, bottleneck and transaction-cost logic.	Framework and propositions.

Analytical boundaries

- The interview is not treated as a neutral dataset; it is an edited public conversation featuring founder perspectives.
- Auto-generated captions may contain transcription errors; direct quotation is minimized.
- Time-sensitive company metrics, valuations and competitive comparisons were excluded from the evidentiary core unless independently verified.
- Export-control outcomes are treated as adaptive and multi-causal.
- The scientific contribution is conceptual. Propositions require future empirical testing.

> **Authenticity statement.** The selection, interpretation, framework, equations, propositions and implications are the author's independent intellectual contribution. Source acknowledgement preserves provenance and does not imply endorsement.

3. THEORETICAL FRAMING

The Agentic Value Migration Framework

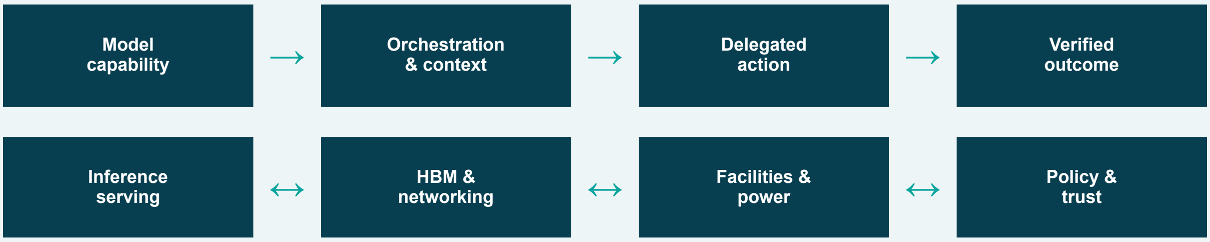

Figure 1. Value creation flows horizontally; constraints and governance condition every stage.

The framework combines three mechanisms.

3.1 Complementarity

Model capability creates value only when combined with context, distribution, workflow redesign and control. As the model layer becomes more accessible, the marginal value of complements can rise.

3.2 Bottleneck migration

System throughput is limited by the binding constraint. When algorithms improve, constraints may migrate to memory bandwidth, networking, electricity, permitting, data rights or human verification. Value and bargaining power can migrate with them.

3.3 Delegation economics

Agents create value through action, but delegated action introduces error propagation and agency risk. The economically relevant output is therefore a verified task, not an attempted task.

> **P1 — Complementary-assets proposition.** As foundation-model capability becomes more substitutable, enterprise value capture will become more strongly associated with proprietary context, workflow integration, distribution and assurance.

> **P2 — Bottleneck-migration proposition.** Improvements in model efficiency will increase the relative strategic importance of whichever complementary resource—memory, energy, authorized data or verification—becomes binding.

From cost per token to cost per verified success

Let a workflow attempt have success probability p, adoption factor a, gross value V, direct inference and tool cost C_i, human supervision cost C_h, expected failure loss L_f, and allocated platform/control cost C_p.

$$\text{Expected net value} = a[pV - C_i - C_h - (1-p)L_f - C_p]$$

This expression clarifies why lower token prices do not guarantee superior economics. An apparently inexpensive agent can be costly when it requires repeated attempts, long contexts, expensive tools or intensive review. Conversely, higher inference spend may be rational when it materially improves completion probability for high-value tasks.

Construct	Operational measure	Failure mode
Completion	Externally verified end state	Self-reported success
Reliability	Success across representative cases	Cherry-picked demos
Supervision	Human minutes per successful task	Hidden labour transfer
Recovery	Containment and restoration after error	Irreversible propagation
Economics	Fully loaded cost per verified task	Token-only accounting

P3 — Verification proposition. Agent adoption will be greatest where task outcomes are cheaply observable, errors are reversible and authority boundaries can be encoded.

P4 — Reliability-adjusted cost proposition. Task-level economic performance will be explained more strongly by cost per verified success than by model price per token.

The physicalization of AI strategy

AI is commonly represented as software, yet its marginal operation is physical. Inference uses accelerators, memory, networking, cooling and electricity. Agentic systems can amplify demand because a single user objective may trigger planning, retrieval, tool calls, critique and retries.

The Stanford AI Index reports that the inference cost of a system performing at approximately GPT-3.5 level fell more than 280-fold between November 2022 and October 2024. Declining unit costs can expand use, a rebound mechanism that may increase aggregate resource demand.

Memory is strategically salient because autoregressive decoding repeatedly accesses model weights and the key–value cache. Large models and long contexts increase capacity requirements; token throughput can be constrained by memory bandwidth. Consequently, high-bandwidth memory is not merely a component cost but a determinant of serving efficiency.

Electricity creates a location and timing constraint. The IEA's 2025 base case projects global data-centre electricity use rising from roughly 415 TWh in 2024 to about 945 TWh in 2030. It further estimates that approximately 20% of planned data-centre projects could face delay unless grid integration risks are addressed.

These results support a systems view: model efficiency, deployment volume and application intensity jointly determine demand. They also suggest that infrastructure portfolios should be evaluated under multiple scenarios rather than a single linear forecast.

> **P5 — Rebound proposition.** Reductions in inference cost will increase aggregate compute and electricity demand when growth in adoption and task intensity exceeds efficiency gains.

> **Implication.** AI strategy should integrate model routing, context efficiency, capacity sourcing, memory architecture, energy availability and carbon intensity in one economic model.

Export controls as an adaptive system

Official U.S. measures have targeted advanced computing semiconductors, high-bandwidth memory, semiconductor manufacturing equipment and diversion risks. The policy objective is framed principally in national-security terms. The motivating interview advances a secondary claim: constraints may accelerate indigenous innovation and substitution in the targeted economy.

Both mechanisms can coexist. A restriction can reduce immediate access while increasing the expected return to domestic alternatives. The net outcome depends on technical distance, capital, talent, market size, allied coordination, enforcement, time horizon and the substitutability of controlled inputs.

Mechanism	Near-term effect	Potential adaptive response	Research measure
Access restriction	Higher cost and delay	Stockpiling, redesign, routing	Price, lead time, import mix
Equipment control	Constrained frontier production	Domestic equipment investment	Patents, capex, yield progression
Entity restrictions	Transaction and compliance friction	Alternative corporate networks	Ownership and trade graphs
Ecosystem pressure	Reduced compatibility	Parallel standards and platforms	Developer and supplier adoption

> **P6 — Adaptive-policy proposition.** The long-run effect of technology controls is mediated by substitution incentives; stronger domestic absorptive capacity increases the probability that controls induce ecosystem divergence rather than permanent dependence.

Enterprise implication

Firms require multi-horizon scenario planning, supplier mapping, end-use compliance and architectural portability. This paper does not assess the legality of any transaction and is not legal or export-control advice.

Productivity gains are real, heterogeneous and organizational

Workplace evidence rejects simple universal claims. Brynjolfsson, Li and Raymond found a 14% average productivity improvement among 5,179 customer-support agents receiving generative-AI assistance, with substantially larger gains among novice and lower-skilled workers. Dillon and colleagues studied 7,137 knowledge workers across 66 firms; tool users spent approximately two fewer hours on email per week, yet the authors did not detect broad changes in task quantity or composition over the study period.

Augmentation channel

AI codifies patterns, accelerates drafting and expands access to expert-like support.

Automation channel

Agents execute bounded steps and may substitute for implementation labour.

Reorganization channel

Roles move toward goal setting, exception handling, evaluation and governance.

Learning-risk channel

Removing junior tasks may weaken apprenticeship and future expertise.

The “small-team unicorn” hypothesis is plausible when AI lowers coordination and implementation costs, but scale with fewer employees can create fragility. Organizational slack, dissent, domain memory and segregation of duties have value not captured by headcount efficiency.

> **P7 — Organizational-complementarity proposition.** Productivity gains from agentic AI will be larger in firms that redesign decision rights, verification roles and learning systems than in firms that add tools without changing workflows.

P7 extends the framework beyond six core technical-economic propositions and identifies an organizational moderator suitable for field testing.

Constructs, moderators and testable predictions

Construct	Candidate indicator	Expected relationship
Model substitutability	Performance dispersion across viable models	Higher substitutability → value shifts to complements
Context advantage	Marginal task success from proprietary data	Positive effect on value capture
Task verifiability	Cost and latency of outcome validation	Positive effect on agent adoption
Failure externality	Expected loss outside initiating workflow	Negative effect on autonomy
Memory intensity	Bytes moved / generated token	Positive effect on serving cost
Energy constraint	Time/cost to secure incremental power	Negative effect on capacity expansion
Organizational redesign	Change in roles, rights and review	Moderates productivity benefit

Recommended empirical designs

- **Workflow randomized trials:** compare assistance, bounded agency and full manual execution using verified outcomes.
- **Panel cost studies:** estimate full cost per successful task across models, architectures and changing price regimes.
- **Infrastructure natural experiments:** exploit variation in grid queues, energy prices or memory availability.
- **Policy difference-in-differences:** examine investment, patenting, trade and technical performance before and after controls, with careful treatment of anticipation and spillovers.
- **Organizational ethnography:** observe how accountability, learning and informal coordination change after agent deployment.

Falsifiability. The framework would be weakened if model ownership consistently explains enterprise value after controlling for context, distribution, verification and infrastructure; or if falling unit inference cost reliably reduces aggregate resource demand despite rapidly expanding use.

Implications for theory, management and policy

9.1 Theory

The framework joins digital complementarity with physical bottlenecks. Unlike software with negligible marginal cost, agentic AI can produce repeated, resource-intensive inference demand. Economic analysis should therefore treat digital intelligence and physical capacity as coupled layers.

9.2 Management

Managers should allocate investment by workflow value and evidence maturity. Model benchmarks are inputs, not business outcomes. Portfolio governance should require five proofs: business fit, technical performance, control effectiveness, unit economics and adoption.

9.3 Infrastructure

Capacity strategy should include scenario ranges for task intensity, context length, memory bandwidth, utilization, location, energy price and connection delay. Architectural efficiency can be a strategic hedge.

9.4 Policy

Technology controls should be evaluated dynamically. Measurement should include not only denied access but substitution, ecosystem divergence, allied leakage and unintended concentration. Energy policy and AI policy increasingly intersect through grid planning and facility location.

9.5 Governance

The governance object is the action pathway: objective, context, authority, tool, state change, validation and audit. A safe model embedded in an uncontrolled workflow can still generate harmful outcomes.

> **Author’s synthesis.** Competitive aggression and institutional responsibility are not opposites. The durable strategy is to move quickly on learning while moving deliberately on delegated authority.

Boundaries of knowledge

Limitations

- A single public interview provides rich propositions but is subject to selection, performance and editing effects.
- Automated captions are imperfect and do not convey all non-verbal context.
- Several agent benchmarks emphasize software tasks and may not generalize to regulated enterprise workflows.
- AI price, performance and policy conditions change rapidly; numerical claims require date-specific interpretation.
- The proposed equations simplify interaction effects and option value.
- No causal conclusion about export-control effectiveness follows from this conceptual analysis.

Conclusion

The economic frontier of AI is shifting from producing fluent answers to completing trusted work. This transition makes orchestration, authorized context, verification and human accountability central product capabilities. It simultaneously increases the importance of inference serving, high-bandwidth memory, networking and electricity.

The Agentic Value Migration Framework predicts that value will move toward scarce complements and binding constraints as model capability becomes more available. It also predicts that the correct enterprise metric is cost per verified successful task, not cost per token. Finally, it treats industrial policy and organizational adoption as adaptive systems rather than linear interventions.

> **Practical conclusion.** The best AI strategy is neither model maximalism nor cautious delay. It is disciplined system design: select consequential but controllable workflows, measure end-to-end economics, preserve accountability and learn faster than the constraint moves.

Declarations

Funding: No external funding declared. **Conflicts of interest:** None declared. **Data availability:** Public video and cited sources. **Ethics:** No human subjects or personal data were collected. **Peer review:** This author manuscript has not undergone peer review.

References

1. Brynjolfsson, E., Li, D., & Raymond, L. R. (2023). Generative AI at work. *NBER Working Paper 31161*. https://doi.org/10.3386/w31161
2. 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
3. International Energy Agency. (2025). *Energy and AI*. Paris: IEA. https://www.iea.org/reports/energy-and-ai/
4. METR. (2026). *Task-completion time horizons of frontier AI models*. Updated 8 May 2026. https://metr.org/time-horizons/
5. OECD. (2025). *AI and work*. https://www.oecd.org/en/topics/ai-and-work.html
6. Roig, J. V. (2025). Towards a standard, enterprise-relevant agentic AI benchmark. *arXiv:2511.08042*.
7. SemiAnalysis. (2025). Scaling the memory wall: The rise and roadmap of HBM. https://newsletter.semianalysis.com/p/scaling-the-memory-wall-the-rise-and-roadmap-of-hbm
8. Stanford Institute for Human-Centered AI. (2025). *Artificial Intelligence Index Report 2025*. Stanford University.
9. Stebbings, H. (Interviewer), & Srinivas, A. (Interviewee). (2026, June 15). Perplexity CEO: Micron will be more valuable than Meta & how export controls helped not hurt China [Video]. *20VC*. YouTube. https://youtu.be/OxFyVcO1Yow
10. U.S. Department of Commerce, Bureau of Industry and Security. (2024, December 2). Commerce strengthens export controls to restrict China's capability to produce advanced semiconductors for military applications.
11. U.S. Department of Commerce, Bureau of Industry and Security. (2025, January 15). Commerce strengthens restrictions on advanced computing semiconductors to enhance foundry due diligence and prevent diversion to PRC.
12. Vishwakarma, H., Agarwal, A., Patil, O., Devaguptapu, C., & Chandran, M. (2025). Can LLMs help you at work? A sandbox for evaluating LLM agents in enterprise environments. *arXiv:2510.27287*.

Correspondence: Jami Kiran · Independent Global Technology Advisor · QuantumReadyCheck.com

Prepared as an independent scholarly perspective. This publication does not represent Perplexity, 20VC, the interview participants or cited institutions.