ESP-JETA
ESP Journal of Engineering & Technology Advancements
ISSN: 2583-2646 / Volume 6 Issue 2 April to June 2026
Paper Id: JETA-V6I2P120/ Page No: 175-187

Original Article

Cognition Economy: Compute, Infrastructure, and the Future of Synthetic Intelligence

Jackson Andrew Srivathsan
Data & AI Leader, Researcher, Chennai, India

Received Date: 20 May 2026 Revised Date: 25 May 2026 Accepted Date: 29 May 2026

Abstract: Human civilization largely solved information scarcity through the rise of the internet, cloud computing, and global digital connectivity. Information scarcity has been mostly overcome by the advent of the internet, cloud computing, and worldwide interconnectedness. This type of information, which could only be found in libraries or institutions, can now be searched and shared across the globe. With AHI, however, it adds another layer of scarcity: the resources necessary for doing cognition on such a large scale, including the energy, organization, and economics of cognitive processes. In this study, the concept of AHI will be used to refer to contemporary AI systems based on the accelerated organization and recombination of language, knowledge systems, etc., from humans.

This paper defines the idea of Cognitive Scarcity, positing that the coming constraint is not simply information access, but access to augmented reasoning capacity, scalable cognition engines, and sustainable cognitive infrastructures. The paper discusses cognition as strategic infrastructure, as well as its dependence on the GPU hardware, data center operations, bandwidth requirements, cooling technologies, orchestration methods, and massive energy infrastructure. Some key ideas include cognitive elites, cognition-per-watt, distributed intelligence, and the nascent "architecture wars" between simple scaling and cognition-based designs. Modern AI systems have capabilities which extend beyond reasoning and cognition – these include sensing, generation, automation, and multimodal interaction. However, while they play a role in the context of the discussion, these concepts lie out of the main scope of this paper.

Keywords: *Cognitive Scarcity, Accelerated Human Intelligence, Artificial Cognition, AI Infrastructure, Distributed Intelligence, Cognitive Economy, AI Sustainability, Cognitive Elites, Amplified Cognition.*

I. INTRODUCTION TO COGNITIVE SCARCITY

Human civilization is about to undergo a transition which will possibly be as revolutionary as the industrial revolution, the age of electricity, and even the emergence of the Internet. The development of humanity was limited by lack of information for centuries. Knowledge could only be accessed based on geographic location, institutional affiliation, social class, and professional expertise. The arrival of the Internet changed this situation dramatically.

The development of search engines, cloud computing, mobility solutions, and international digital connections led to creation of an economy based on information retrieval, exchange, and coordination. Knowledge could now be accessed virtually instantaneously from anywhere in the world.

Nevertheless, whereas humans have succeeded to overcome the problem of information scarcity, there appears another level of scarcity that is quickly surfacing.

This is because contemporary intelligent machines are starting to do more than simply retrieve information and instead are beginning to actively engage in reasoning, synthesis, orchestration, decision making, programming, optimization, and cognitive processing.

As the digital economy came into its own, information became known as “the new oil.” But as far as the cognition economy is concerned, the information available is closer to being crude oil rather than refined fuel. What matters now is not simply the data available but the refinement and synthesis that comes from combining information to create cognition and reason. Cognition itself becomes the refined output of modern computing capabilities.

Introduction of the idea of Cognitive Scarcity – The notion that while in the past the bottleneck was access to information, in the future the bottleneck will be access to scalable cognition. This, in conjunction with the continued development of frontier cognition technologies, means that cognitive competition will be a relative rather than absolute race.

In this paper, the phrase “Accelerated Human Intelligence” or AHI will refer to the modern AI paradigm, which relies mainly on the acceleration, coordination, and recombination of human language, structures of reasoning, knowledge bases, and computing infrastructure in order to realize its capabilities. Modern AI systems are no longer considered alien cognitive

constructs completely separate from the human mind but rather advanced extensions of human cognition through infrastructural means.

Unlike standard information systems, amplified cognition systems require significant infrastructure such:

- GPUs,
- specialized semiconductors,
- datacenters,
- orchestration systems,
- memory bandwidth,
- cooling infrastructure,
  and large-scale energy consumption.

As frontier cognition systems continue scaling, cognition itself increasingly becomes an infrastructural resource constrained by economics, compute availability, energy, and industrial capacity.

The significance of this transition is already visible globally. Organizations such as OpenAI, Anthropic, Google, Microsoft, and xAI are rapidly expanding datacenter infrastructure to support increasingly compute-intensive cognition systems, while NVIDIA has emerged as one of the most strategically important infrastructure providers of the cognition economy.

This paper argues that cognition itself may increasingly become a strategic economic and industrial resource. The paper further explores how amplified cognition systems are reshaping:

- infrastructure,
- economics,
- productivity,
- energy demand,
- architecture design,
- cognitive inequality,
  and the future distribution of civilizational power.

The paper introduces concepts including:

- cognitive elites,
- cognition-per-watt,
- distributed intelligence,
- local versus frontier cognition,
  and the emerging architecture war between brute-force scaling and efficient cognition systems.

The paper further argues that future civilization may increasingly revolve not around access to information alone, but around sustainable cognition architectures capable of amplifying intelligence efficiently at scale.

While modern AI systems possess capabilities extending beyond reasoning and cognition, including perception, generation, automation, robotics, and multimodal interaction, these domains remain outside the primary scope of this paper. References to such capabilities may appear where contextually relevant, but the central focus of this paper remains the infrastructure, economics, scalability, and sustainability of amplified cognition systems.

Humanity solved information scarcity through the internet. Accelerated Human Intelligence introduces a new scarcity layer: the cost of synthetic thought itself.

II. FROM THE INFORMATION ECONOMY TO THE COGNITION ECONOMY

The transition of information access to cognition enhancement constitutes a paradigmatic shift for civilizations. The source of competitive advantage within the information economy was based on the velocity of accessing, distributing, and communicating information. The source of competitive advantage within the cognition economy is based on the velocity of synthesizing, interpreting, orchestrating, and transforming information into action by means of cognition enhancement.

The cognition economy, thus, is different from the information economy. While conventional information systems mainly retrieve information, accelerated human intelligence (AHI) systems are becoming more and more involved in the process of reasoning itself.

The modern AHI system is more adept at reasoning, creating software, coordinating processes, conducting analysis, creating strategic decisions, automating cognitive functions, speeding up human problem solving, debugging systems, summarizing scientific literature, architecting systems, optimizing processes, and taking part in strategic decision making.

The advent of systems like GitHub Copilot, Claude Code, OpenAI Codex, and Gemini-driven developer platforms by Google illustrates this shift. This is because the system is no longer about just providing application layer & data.The system takes an active part in cognitive execution.

A single person working with advanced cognition systems is able today to accomplish the work that would normally require bigger engineering teams. The result is an imbalance between productivity and a novel type of inequality: “cognitive inequality,” wherein those who have better cognition systems will be able to do things faster and learn and innovate better than others.

In other words, the move from information access to cognition amplification is not simply a matter of technology. It marks the rise of a new economy – one driven by enhanced reasoning.

With respect to the information economy, searches obtain knowledge, human cognition interprets results, and information access defines advantage. With regard to the cognition economy, AHI increasingly reasons, engages in execution, and enhances cognitive throughput.

With continued advancements in frontier cognition technology, competition within the realm of scalable cognition increasingly evolves into a relative race in relation to advancing intelligence baselines, rather than the acquisition of technological parity.

Unlike traditional information distribution systems, AI or synthetic cognition requires substantial infrastructure, including GPUs, specialized semiconductors, memory bandwidth, orchestration systems, datacenters, cooling infrastructure, and large-scale energy consumption. As frontier AHI systems continue scaling, cognition itself increasingly becomes an infrastructural resource.

The significance of this shift is already visible in real-world AI infrastructure expansion. NVIDIA’s rapid rise as a dominant AI hardware supplier demonstrates how synthetic cognition increasingly depends on semiconductor manufacturing and compute concentration. Simultaneously, organizations such as OpenAI, Anthropic, Google, Microsoft, and xAI continue rapidly expanding datacenter infrastructure to support increasingly compute-intensive cognition systems.

Humanity solved information scarcity through the internet. Accelerated Human Intelligence introduces a new scarcity layer: the cost of synthetic thought itself.

III. THE COST OF SYNTHETIC THOUGHT

One of the largest misconceptions surrounding artificial intelligence is the assumption that synthetic cognition scales similarly to information distribution.

In the era of the internet, information duplication was relatively cheap. Information once created would have been duplicated, distributed, cached, and transferred all over the world without any additional cost. The synthetic thinking approach is fundamentally different when compared to the economics of information reproduction. While information is static, synthetic thinking involves computations on every move.

As a result, synthetic cognition remains computationally expensive. Modern frontier AI systems depend heavily on industrial-scale infrastructure, including:

- specialized GPUs,
- large-scale memory systems,
- distributed orchestration frameworks,
- inference infrastructure,
- high-speed networking,
- cooling systems,

and large-scale energy infrastructure.

Unlike traditional search systems that primarily retrieve indexed information, synthetic cognition continuously consumes computational resources during operation. Large language models actively utilize:

- compute cycles,
- memory bandwidth,
- networking throughput,
- storage access,

and electrical power.

The operational cost of synthetic cognition increases significantly as:

- context windows expand,
- reasoning depth increases,
- multimodal systems evolve,
- agent orchestration complexity grows,

and real-time inference requirements increase.

The cost of synthetic thought is therefore not purely software-driven. It is industrial.

Frontier AI systems such as GPT, Claude, Gemini, Grok, and other large-scale cognition platforms demonstrate the scale of this infrastructure dependency. Industry reports and infrastructure disclosures increasingly point toward substantial operational expenses associated with:

- inference operations,
- training workloads,
- hardware acquisition,
- semiconductor supply,
- networking infrastructure,

and datacenter expansion.

The publicized restrictions imposed by Anthropic due to GPU limitations are additional proof of how synthetic intelligence suffers from physical limits of chip manufacturing capacity and availability of computing infrastructure. Unlike biological intelligence, which works using 20 watts of energy as per the understanding so far, the next-generation AI could possibly need millions of watts of energy for processing tasks.

This creates an important distinction:Human cognition evolved through extreme energy efficiency. Synthetic cognition currently scales through industrial expansion. As AI adoption accelerates globally, cognition itself increasingly becomes constrained by:

- economics,
- infrastructure availability,
- energy generation,
- semiconductor manufacturing capacity,

and sustainability.

Traditional datacenter KPIs primarily measure computational efficiency shown in Table I. However, the cognition economy increasingly requires new metrics capable of measuring scalable reasoning efficiency itself. Existing industry metrics include:

Table 1: Traditional Infrastructure Kpis For Ai Datacenters

Traditional Infrastructure KPI	**Purpose**
PUE (Power Usage Effectiveness)	Measures datacenter energy efficiency
GPU Utilization	Measures hardware efficiency
FLOPS/Watt	Measures compute efficiency
Tokens per Second	Measures inference throughput
Inference Latency	Measures response performance
Cost per Inference	Measures operational cost
Memory Bandwidth	Measures data transfer capability

While these metrics effectively measure computational infrastructure performance, they do not fully capture the efficiency of scalable cognition systems. As synthetic cognition becomes increasingly industrialized, cognition-centric metrics may become equally important.

One proposed metric is Watts-per-Cognition (WpC), representing the amount of energy consumed to generate a unit of usable cognitive output (Equation 1).

$$WpC = \frac{E_c}{C_o} \quad \square\square\square$$

Where:

- WpC = Watts-per-Cognition
- C_o = Cognitive Output
- E_c = Energy Consumption

Similarly, USDollar (DpC) may become increasingly important as organizations attempt to optimize scalable cognition economically (Equation 2).

$$DpC = \frac{C_t}{C_o} \quad (2)$$

Where:

- DpC = USDollar
- C_o = Cognitive Output
- C_t = Total Operational Cost

In addition to WpC and DpC, more comprehensive cognition and economic metrics may become more prominent as the industrialization of synthetic cognition advances. Organizations may start to measure their operations based on efficiencies related to the cognitive process itself, with metrics including Humans-per-Cognition (HpC), or the number of humans necessary for operational support for each cognitive task, Rent-per-Cognition (RpC), or the cost associated with physical infrastructure in relation to cognition creation, and Revenue-per-Cognition (RpC), or the value created by synthetic cognitive processes. In the same way that industrial economies came to be measured using metrics such as the cost of producing each unit, economies involving synthetic cognition may come to be measured using metrics directly related to cognition production.

As synthetic cognition starts to become part of operations within organizations, there is the possibility that businesses start to compare labor productivity and synthetic cognition productivity. Historically, organizations have used labor intensity measures to gauge their performance, including the number of people working per million dollars generated in revenue. Within the cognition economy, it is possible that this measure evolves into Cognitions-per-Million-Dollars, where the measure of synthetic cognitive outputs produced to create or help produce one million dollars in revenue will be gauged. It should be noted that these metrics do not mean that humans are going to be replaced. The organization would be able to compare the efficiency of operating in the human-only mode, AI-supported human mode, and cognition-driven operations.

Thus, the problem that will occur in the future will not be the availability of information or computer processing capabilities per se. Instead, it might become the ability to produce scalable cognition.

IV. COMPUTE SOVEREIGNTY AND THE GEOPOLITICS OF COGNITION

As synthetic cognition comes to depend more on industrial infrastructure, compute capacity might come to be seen as a strategic geopolitical resource. For traditional industrial economies, dependency on energy exporting countries for maintaining manufacturing, transportation, and continuity was key. Similarly, for the cognition economy, dependency on cognition-exporting nations able to produce scalable reasoning, inference, and cognitive orchestration capacity might prove crucial.

The development of artificial intelligence has thus evolved from being a matter of technology alone into being an infrastructure challenge. The ongoing global race to dominate semiconductors, GPUs, datacenters, cloud computing facilities, networking infrastructure, and energy production is in reality one of controlling scalable cognition. State-of-the-art cognition systems rely on advanced semiconductor fabrication facilities, hyperscale computing centers, highly specialized highperformance GPUs, energy production at scale, networking technologies, and industrial cooling technologies.

The growth of AI departments in companies like Google and Microsoft and the emergence of organizations dedicated specifically to AI such as OpenAI, Anthropic, xAI, and even the dominance of NVIDIA in the realm of compute infrastructure reflect this shift. The dominance of NVIDIA in the sphere of AI compute infrastructure represents how synthetic cognition is becoming increasingly dependent on concentration of semiconductors and access to compute power. Graphics Processing Units, which were initially designed for graphic rendering, emerged as critical to machine learning due to their parallel processing power.

Such a concentration will generate a new kind of dependency. Companies and countries will be dependent upon cognition infrastructure providers for capabilities such as inference, AI orchestration, cloud cognition, large scale reasoning systems, and

synthetic cognitive execution. Economies have traditionally had to import energy to operate industrial machines. Cognition economies too may soon import scalable synthetic cognition to operate systems.

This paper defines cognition export as the large-scale delivery of synthetic reasoning, inference capability, cognitive orchestration, and AI-assisted decision execution across organizational or national boundaries. Unlike traditional software export, cognition export represents ongoing synthetic cognitive execution delivered continuously through infrastructure-intensive systems. The global economy may therefore gradually transition from human cognition export toward scalable synthetic cognition export.

Countries like India have historically made their mark in the international community through mass export of human cognitive labor, ranging from software engineering to analytics, consulting, support, and outsourcing. With the advent of scalable synthetic cognitive capabilities that would automate part of this cognitive labor, there is every likelihood that the process of cognition export will transform itself in many respects. It doesn't imply that human cognitive labor has no further scope, but that the human contribution could very well involve cognition orchestration and governance among others.

The countries which have a large number of engineers, an indigenous AI education system, infrastructure for implementation, and scalable organizations would be able to continue their strategic advantages within the cognition economy. While, simultaneously, the nations that seek to adopt a sovereign AI strategy for themselves can be seen reducing their reliance on other nations to provide them with cognition services.

This introduces geopolitical implications extending beyond technology itself. Restrictions involving semiconductor exports, GPU access, cloud infrastructure availability, and energy allocation may increasingly influence global cognitive competitiveness. The cognition economy may therefore reshape traditional geopolitical dynamics. Historically, industrial power depended heavily on access to land, energy, manufacturing, transportation, and information systems. In the cognition economy, strategic influence may increasingly depend on access to scalable synthetic cognition infrastructure and cognitive execution capability.

Nevertheless, the cognition economy may also work to reduce traditional entry barriers. Countries and corporations, in particular, can now afford to utilize scalable cognition capabilities due to the cloud, distributed cognition systems, and AI natively designed operations. Competitiveness would be gained by companies that could quickly align their organizational structures, educational systems, and operations to synthetic cognition adoption.

The future geopolitical order does not necessarily have to be determined simply by a country's population, manufacturing capability, or economic size. In past times, a large population was a great advantage as the productivity of nations was heavily dependent on labor efficiency. However, this might not be applicable for the future with the emergence of the cognition economy, where merely having a larger population might not necessarily provide an advantage in competition. With more effective cognitive capabilities being achieved through cognitive augmentation technology and AI-enabled operations, countries with smaller populations who are capable of utilizing cognitive technology effectively might be better able to compete than larger populations without the same technology.

Ultimately, the future of synthetic cognition may depend not only on algorithms, but on compute sovereignty, semiconductor manufacturing capacity, cognition infrastructure ownership, energy availability, distributed cognitive ecosystems, and strategic control over scalable synthetic reasoning systems.

V. SYNTHETIC COGNITION AS ACCELERATED HUMAN COGNITION

Cognition in a synthetic form is often presented as mysterious, foreign, or alien to organic forms of intelligence. In reality, artificial intelligence largely exists in the form of speeded up, mechanized versions of human intelligence. Disruptive capability of synthetic cognition does not consist in producing new forms of foreign intelligence, but rather in accelerating and orchestrating human intellectual achievements.

Large language models are trained on:

- human language,
- human reasoning structures,
- human software,
- human knowledge,
- and human-generated patterns.

In this regard, synthetic cognition is not a product that evolves without mankind's participation. Instead, it constitutes enhanced human cognition utilizing computing power on a scale that can be achieved by computers. AI acquires knowledge based on human cognitive products that have been created by humanity and synthesizes them beyond human capabilities.

The apparent "intelligence" of synthetic cognition therefore emerges primarily from:

- speed,
- scalability,
- memory accessibility,
- parallelization,
- and continuous operation.

This creates what may be described as cognitive compression. Tasks that may historically require humans days, weeks, or even years of sequential reading, comparison, synthesis, and reasoning can increasingly be processed within minutes through industrial-scale cognition systems. The disruption created by synthetic cognition may therefore stem less from the existence of intelligence itself and more from the compression of human cognitive timescales.

The functioning of human cognition is sequential in nature and is limited by biological parameters like fatigue, attention spans, memory, and time. Artificial cognition, however, can concurrently process immense spaces of knowledge, function without biological fatigue, quickly integrate information from different fields, and parallelize reasoning loads through distributed computing systems.

AI-assisted scientific systems such as AlphaFold demonstrate this acceleration clearly. Protein structure prediction, historically requiring substantial experimental effort and lengthy research cycles, became dramatically accelerated through synthetic cognition systems capable of large-scale pattern analysis. Similarly, enterprise AI copilots increasingly accelerate software development, analytics, workflow automation, documentation generation, operational support, and strategic analysis.

Importantly, synthetic cognition should not necessarily be interpreted as replacing human cognition entirely. Rather, it increasingly functions as cognitive amplification infrastructure capable of extending human cognitive throughput. A single individual operating alongside scalable synthetic cognition systems may increasingly perform workloads that historically required substantially larger teams, longer timelines, or specialized analytical structures.

Human civilization has been able to industrialize their physical effort using machines and manufacturing technologies. Synthetic cognition can allow the cognition economy to industrialize certain parts of reasoning, synthesis, analysis, and execution using synthetic cognition technology. Therefore, the importance of artificial intelligence could be in its ability to industrialize accumulated human cognition rather than create new forms of intelligence.

While AI could substitute some human intellectual functions, the overall human contribution to the functioning of civilization cannot be substituted. Human synthetic intelligence can never exist independently of human cognitive abilities based on human knowledge and language.

VI. COGNITIVE INEQUALITY AND THE NEW COGNITIVE CLASSES

The emergence of artificial intelligence brings with it the threat of civilization-level cognitive disparity. Information-based economic inequality was characterized by disparities between access to education, internet access, computing resources, and information itself. But cognitive economic inequality could potentially be marked by a disparity between those who have access to cognitive enhancement through sophisticated artificial intelligence solutions and those who can effectively utilize their cognitive enhancement.

The transition from an information-based economic model of inequality into a cognitive economic model is a fundamental change. For example, during the period of internet dominance, having access to information gave one a competitive edge. In the cognition economic era, two people could both have equal access to information, but one person could still have superior cognitive abilities due to cognitive augmentation enabled by the use of artificial intelligence technology and AI-driven workflows.

Those individuals and firms who have access to highly sophisticated synthetic cognition systems will perform better than their competitors in terms of execution, rate of learning, development of software programs, analysis, operations, adaptability, and cognitive processing. Productivity acceleration is an asymmetric phenomenon seen in enterprise AI copilots. Individuals can accomplish tasks that otherwise could not be performed by them due to the involvement of more people, more time, and more resources.

As synthetic cognition increasingly finds itself integrated within the operations of various systems, new cognitive categories could slowly come to light. Among those who possess strong cognitive capabilities are people or nations that are able to efficiently incorporate synthetic cognition in such things as education systems, business operations, software programming, analysis, decision-making, and strategic implementation.

The cognitively poor, by contrast, may not necessarily lack internet access or economic resources. Instead, they may possess limited ability to operationalize scalable cognition systems effectively. In the cognition economy, cognitive disadvantage

may increasingly emerge from limited AI literacy, weak orchestration capability, low cognition integration, or slow adaptation toward AI-native operational models.

This creates an important distinction between information access and cognition amplification. Historically, literacy gaps created inequality. Later, internet access and digital literacy created new forms of economic divergence. The cognition economy may similarly introduce a new divide centered around cognition orchestration capability itself.

Cognitive inequalities can arise at various social strata all at once. On an individual basis, cognitive differences due to AI literacy, cognition amplification ability, and learning acceleration will come to define competitiveness. In terms of labor forces, AI-driven productivity inequality will have a huge impact on differences in productivity among employees who do similar tasks. From an organizational perspective, firms that have better orchestration ability and cognition integration capabilities will dominate strategically. From a national perspective, nations with more advanced AI infrastructure, education systems, implementation environments, and cognition architecture will prevail.

Importantly, future cognitive inequality may not emerge solely from ownership of massive datacenters or frontier-scale infrastructure. Efficiency may disrupt scale. History repeatedly demonstrates that efficient architectures often challenge brute-force systems. Smaller organizations or nations possessing highly optimized cognition systems, distributed intelligence architectures, and strong AI-native operational ecosystems may potentially outperform significantly larger competitors operating with weaker cognition amplification capability.

As synthetic cognition continues evolving, competitive advantage may increasingly depend not only on intelligence itself, but on the ability to amplifing intelligence operationally at scale. The cognition economy may therefore create a new societal divide not solely based on wealth, labor, or information access, but on the ability to orchestrate, operationalize, and amplifying cognition itself.

VII. FRONTIER COGNITION VS LOCAL COGNITION AND THE NEED FOR DISTRIBUTED INTELLIGENCE

The current ecosystem of artificial intelligence is dominated by frontier scale cognitive architectures that run on hyperscale data centres. Examples of such include GPT, Claude, Gemini, and Grok. These exhibit highly advanced cognitive reasoning abilities, but these come at the cost of infrastructure, power usage, orchestration complexities, and computational concentration. But not all cognitive tasks need frontier scale cognitive architectures.

Many real-world cognitive tasks involve lightweight reasoning, localized inference, contextual assistance, search augmentation, summarization, autocomplete functionality, workflow support, or operational guidance that can increasingly be handled efficiently by smaller cognition systems operating with significantly lower infrastructure requirements. This introduces the growing importance of distributed intelligence architectures combining:

- frontier cognition,
- local cognition,
- edge inference,
- selective escalation,
- and hybrid orchestration.

In small-scale ecosystems of Gemma, Phi, Mistral, Llama, and other lightweight models, it is now becoming apparent that efficient cognition systems do not have to be run on large computing power and can easily be performed on local devices, which require less computational resources. It is clear, for example, from the performance of Mistral 7B in reasoning capabilities.

These trends indicate that in the future of artificial cognition, there will be no need for a central, frontier-scale model responsible for performing any type of cognitive function. Instead, the future cognition ecosystems may increasingly tend towards orchestration systems where different cognition models with diverse performance and reasoning depth co-exist and co-work.

AI systems in their modern form tend to implement the above-described architecture more and more frequently. Advanced AI solutions like GPT, Claude, or Gemini already implement different modes of operations optimized for various criteria such as performance, reasoning depth, or efficiency. Web browsers, business solutions, productivity software, operating systems, and edge devices, in turn, are increasingly leveraging lightweight local models in operational flows. Examples include integrations into Chrome browser, AI operating systems, embedded copilots, and local inference systems, showing how cognition gradually starts to become distributed across the entire digital environment.

With AI solutions continuing to proliferate, one could expect that in the future, most software platforms would have multiple cognition layers running simultaneously within them. In this scenario, lightweight local models could process fast operations, context-awareness, personalized experiences, autocomplete functionality, light reasoning, and inference, whereas computationally expensive frontiers of cognition could be utilized only as needed.

The human brain participates not only in conscious reasoning, analysis, planning, abstraction, and decision-making, but also continuously regulates numerous unconscious operational functions including breathing regulation, motor coordination, sensory interpretation, and autonomous body management. Simultaneously, the cardiovascular system supports biological stability while emotional states significantly influence cognition, motivation, behavior, and decision-making. The gastrointestinal system similarly performs specialized biological functions while increasingly being associated with subconscious pattern recognition, intuition, stress responses, and gut-driven behavioral influence through complex neurochemical interactions.

Human intelligence therefore does not operate as a single monolithic cognition engine allocating maximum reasoning power to every activity equally. Repetitive tasks increasingly become procedural and automated through habit formation and muscle memory, while more complex situations selectively activate deeper analytical, emotional, strategic, or abstract cognitive processing.

Biology's intelligence can thus be seen as more of an orchestrated system of subsystems that continually interact dynamically, according to operational context, energy efficiencies, environmental stimulus, emotion, and cognitive need. Future synthetic cognitive systems could develop similarly, whereby many different cognitive models would concurrently operate at various levels of reasoning depth and speeds, energy limitations, and other contextual considerations instead of using only centralized cognitive processes at the scale of the frontier for everything.

Not all cognition requires centralized computation. Future synthetic cognition architectures may increasingly evolve toward:

- distributed inference,
- local specialization,
- edge cognition,
- cognition routing,
- and selective escalation into frontier systems only when necessary.

This transition may become increasingly important for sustainability, accessibility, WpC optimization, infrastructure efficiency, latency reduction, operational scalability, and global cognitive distribution. The long-term future of synthetic cognition may therefore depend not only on building increasingly powerful frontier systems, but on building efficient distributed cognition ecosystems capable of allocating cognitive resources dynamically based on real-world operational requirements.

VIII. THE ARCHITECTURE WAR: EFFICIENCY VS SCALE

One of the key contests in the cognition economy could be the design contest between brute force scaling and effective cognition architectures. The current generation of frontier AI is characterized by the ability to achieve capability expansion using brute force scaling, meaning bigger parameter sizes, bigger training data, more GPUs in clusters, more data centers, and increased energy use significantly boosted the advancement of synthetic cognitive systems in the last decade.

The approach has been remarkably successful in enhancing cognitive functions such as reasoning, multimodal abilities, language, and large-scale cognition orchestration. Frontier cognition systems like GPT, Claude, Gemini, and Grok exemplify the capabilities that arise from scaling cognition systems with compute power and infrastructure consolidation.

But history has proven many times that forceful dominance is often only temporary, not permanent. Efficiency architecture usually ends up disrupting bigger and resource-heavy architecture after some period. History proves this repeatedly in industrial context. Evolution of steam engine has led to more efficient combusting system, which then further evolved into even more efficient electric system. The same has happened repeatedly in computing architecture's evolutionary history where efficiency, not scale, dominated.

Computing architecture's evolution is a perfect example of this phenomenon. RISC architecture's evolution disrupted CISC assumptions of complexity by proving the point of simplicity and efficient instruction set execution. Then again, ARM architecture disrupted x86 architecture through better energy efficiency, thermal optimization, and lightweight scalability for mobile computing applications. Efficient distributed system architecture was able to challenge central architecture because of reduction in infrastructure costs and increased efficiency.

Current frontier cognition systems remain heavily dependent on large-scale infrastructure expansion. However, future synthetic cognition architectures may increasingly prioritize efficiency, specialization, orchestration optimization, and selective cognition allocation rather than continuous brute-force scaling alone. Emerging approaches including sparse architectures, Mixture-of-Experts systems, SSA, JEPA, distributed inference systems, model distillation, lightweight cognition models, and biologically inspired orchestration architectures may significantly reduce the infrastructure cost of synthetic thought.

Importantly, not all cognitive workloads require frontier-scale reasoning systems. Many operational tasks such as summarization, contextual assistance, workflow routing, autocomplete functionality, lightweight analytics, search

augmentation, and localized inference may increasingly be handled efficiently through smaller cognition systems operating at substantially lower infrastructure cost. This creates growing importance for selective cognition escalation, where lightweight models handle simpler cognitive tasks while more complex reasoning workloads are escalated into deeper frontier cognition systems only when necessary.

Modern ecosystems of AI already show evidence of this shift. Cutting-edge platforms not only enable more and more modes that are designed to optimize either reasoning, performance, or inference efficiency. However, simultaneously, smaller model ecosystems such as Gemma, Phi, Mistral, and Llama prove capable of efficient cognitive functions beyond the realms of massive data centres.

This mirrors biological intelligence itself. Human cognition does not allocate maximum reasoning intensity to every task equally. Repetitive activities increasingly become optimized through procedural cognition, habit formation, subconscious processing, and specialized biological orchestration, while deeper analytical reasoning activates selectively only when necessary. Future synthetic cognition systems may increasingly evolve similarly through distributed cognition architectures dynamically allocating reasoning depth based on contextual complexity, operational necessity, latency requirements, and energy efficiency constraints.

The long-term competition within the cognition economy may therefore extend beyond model intelligence alone. Increasingly, competitive advantage may depend on delivering scalable cognition sustainably under real-world infrastructure, operational, and energy constraints. The future war may not be over the smartest AI, but over the most sustainable cognition architecture.

Future cognitive dominance may increasingly depend on:

- Watts-per-Cognition (WpC),
- USDollar (DpC),
- distributed efficiency,
- selective cognition escalation,
- sustainable orchestration,

and infrastructure-efficient cognition allocation.

The long-term winners of the cognition economy may therefore not necessarily be the organizations possessing the largest models or largest data centres alone, but those capable of delivering scalable synthetic cognition most efficiently, sustainably, and economically at industrial scale.

IX. COGNITIVE ELITES AND THE FUTURE OF CIVILIZATION

The cognition economy could radically change the fabric of civilizations. Since ancient times, civilizations have been characterized by their control over critical resources and infrastructure. Agricultural civilizations were based on control of land and agriculture. Industrial civilizations evolved as a result of their ability to manufacture, transport, and provide mechanized labor. Subsequently, civilizations came to rely on energy infrastructure, semiconductors, telecommunications, and information networks.

There might be an additional great transformation that comes with the cognition economy. The reason behind this is that there will come a time when strategic advantage will not just come from physical resources, industries, and information, but cognition itself.

This paper introduces the concept of cognitive elites: individuals, organizations, enterprises, or nations possessing disproportionate cognitive amplification capability through superior cognition architectures, orchestration systems, distributed intelligence ecosystems, infrastructure efficiency, and AI-native operational models. Their advantage may increasingly emerge not simply from ownership of compute infrastructure alone, but from the ability to transform scalable synthetic cognition into operational, economic, scientific, military, organizational, and strategic acceleration.

Historically, power concentration often emerged through ownership of scarce strategic infrastructure. In the cognition economy, scalable cognition systems themselves may increasingly become strategic infrastructure. Organizations capable of integrating synthetic cognition deeply into decision-making, analytics, software development, education systems, operations, governance, research, engineering, and enterprise execution may increasingly operate at significantly higher adaptive velocity than slower-moving competitors.

It is also important to mention that future cognitive dominance may not be confined only to those organizations that have the biggest data centers or biggest models on the frontier. Historically, efficient architectures are known to disrupt inefficient large-scale systems. Computing history provides evidence of this trend. For example, efficient ARM architectures disrupted the established paradigm of x86 architectures due to their efficiency, portability, and lightness. Smartphones also

disrupted computing not in terms of exceeding desktop computing capacity, but in terms of distribution of computing to billions of users efficiently.

The same pattern could be increasingly prevalent for synthetic cognition as well. Small, optimally designed cognition systems based on distributed orchestration architectures could increasingly become competitors to the massive scale of centralized cognition systems. The open-source cognition ecosystem would speed up this process by creating an environment where more countries could engage in the development, deployment, customization, and optimization of synthetic cognition. With the increasing distribution of cognition systems, cognitive power itself could slowly shift from being dominated by a few centralized infrastructure players.

In this respect, the rising trend of sovereign AI projects is already indicating that governments know how critical cognition infrastructure is to national security and sovereignty. Countries have increasingly become focused on attaining sovereignty over their semiconductor supplies, computational power, cognition systems, AI R&D capabilities, and operations.

Cognitive elite emergence can likewise impact the structure of workforces and organizations. Traditionally, economies have relied greatly on workforce sizes. But in the cognition economy, the success of small amplified workforces may be achieved over more substantial traditional organizations because of cognition acceleration, AI-enabled implementation, and scalable orchestration capacity. The future competitive environment would consequently be dominated by organizations that could effectively integrate scalability into cognition amplification structures and human adaptability.

It might even lead to a fundamental change in the hierarchy of civilization. Traditional civilizations have fought each other in terms of land, manufacturing industries, military, transportation, energy, and information systems. Future economies may add one more level – scalable synthetic cognition.

The future hierarchy of civilization may therefore depend not solely on physical resources, population scale, industrial output, or access to information alone, but increasingly on the ability to amplify, orchestrate, distribute, and operationalize cognition itself at planetary scale.

X. CONCLUSION: THE FUTURE OF SYNTHETIC THOUGHT

Human civilization largely solved information scarcity through the rise of the internet, cloud computing, global connectivity, and digital infrastructure. Information that once required specialized institutions, physical libraries, or privileged access became globally searchable and distributable at unprecedented scale. However, the rise of artificial intelligence introduces a fundamentally different challenge. While information replication became relatively inexpensive, scalable synthetic cognition remains heavily constrained by compute infrastructure, semiconductor manufacturing, energy availability, orchestration complexity, operational economics, and sustainability requirements.

This paper introduced the concept of Cognitive Scarcity and argued that civilization is increasingly transitioning from an information economy toward a cognition economy. In this emerging economic structure, competitive advantage may increasingly depend not solely on access to information itself, but on the ability to amplify, operationalize, orchestrate, and distribute cognition efficiently at scale.

In contrast with traditional software applications, synthetic cognition is rapidly becoming part of industrial infrastructure. Modern cognition systems require substantial investment in data centres, GPU hardware, networking technology, automation, semiconductor supply chains, and energy infrastructures. In an environment where synthetic cognition is increasingly integrated within the operation of enterprises, governance frameworks, engineering processes, analysis, education, science, and strategic planning, cognition itself may become an economic and industrial production layer that can be measured.

This new cognition economy may simultaneously reconfigure labor dynamics, organizational structures, geopolitics, infrastructure planning, and civilizational competition. Scalable cognition systems will increasingly determine productivity, speed of execution, speed of learning, efficiency, and adaptation capacity both at the individual level, organizationally, and nation-state wise. Cognitive inequality will become an increasing reality while also presenting opportunities for cognition augmentation through distributed intelligent and AI-native infrastructure.

On the other hand, biological intelligence shows amazing efficiency in adaptation. Human cognition evolved through distributed orchestration, local specialization, selective allocation of reasoning, sub-conscious optimization, emotion-driven momentum, process automation, and unbelievable energy efficiency. Human intelligence does not apply full cognition intensity to all tasks equally. Cognition instead adapts dynamically depending on context complexity, necessity of action, and biological limits of energy efficiency.

Synthetic cognition systems in the future could increasingly follow similar evolutionary path. Future cognitive systems will no longer depend entirely on central cognition systems at a frontier scale for performing all tasks. Distributed intelligence systems could evolve through frontier cognition, local cognition, inference at edges, lightweight cognitive processes, selective cognition intensification, sparse cognitive networks, and orchestration of cognition systems. Evolution of future synthetic cognition would thus be determined not only by size of models but more importantly by cognition efficiency and orchestration.

Ultimately, civilization itself may increasingly reorganize around cognition infrastructure and cognition amplification capability. Historically, civilizations competed through control over land, industry, transportation, manufacturing, energy, and information systems. The cognition economy may introduce a new strategic layer centered around scalable synthetic cognition itself.

The future competition of the cognition economy will thus likely evolve beyond simply seeking the biggest models or biggest datacenters. The future battle may not be about who builds the smartest AI but about whose architecture is more sustainable. Large-scale cognition systems relying on industrial-scale infrastructure may remain competitive with their advantages in large-scale reasoning, multimodal cognition synthesis, computational science, and cognitive capability generalization. But historical precedent suggests that highly efficient architectures can overcome even bigger and more resource-intense systems in time. A well-optimized cognition system running efficiently on small, simple hardware could outperform even more infrastructure-heavy systems in specialized application domains through its efficiency, scalability, accessibility, deployability, and optimal use of resources.

The evolution of synthetic cognition in the future will therefore no longer be decided only by compute scaling, but more by considerations of scale, efficiency, distribution, and orchestration. Future limits might no longer be in information access but in cognition generation, orchestration, distribution, and operation.

XI. REFERENCES

[1] Acemoglu, D., & Restrepo, P. (2018). Artificial intelligence, automation and work (NBER Working Paper No. 24196). National Bureau of Economic Research. https://www.nber.org/papers/w24196

[2] Amodei, D., Hernandez, D., Sastry, G., Clark, J., Brockman, G., & Sutskever, I. (2018). AI and compute. OpenAI. https://openai.com/research/ai-and-compute

[3] Anthropic. (2024). Anthropic research. https://www.anthropic.com/research

[4] Bremmer, I. (2022). The power of crisis: How three threats—and our response—will change the world. Simon & Schuster.

[5] Brynjolfsson, E., & McAfee, A. (2014). The second machine age: Work, progress, and prosperity in a time of brilliant technologies. W. W. Norton & Company.

[6] Brynjolfsson, E., Li, D., & Raymond, L. R. (2023). Generative AI at work (NBER Working Paper No. 31161). National Bureau of Economic Research. https://doi.org/10.3386/w31161

[7] Castells, M. (1996). The rise of the network society. Blackwell Publishers.

[8] Chase, H. (2022). LangChain (Version 0.1) [Computer software]. https://github.com/langchain-ai/langchain

[9] Clark, A. (2003). Natural-born cyborgs: Minds, technologies, and the future of human intelligence. Oxford University Press.

[10] Cotra, A. (2020). Forecasting TAI with biological anchors. Open Philanthropy.

[11] Damasio, A. (1994). Descartes' error: Emotion, reason, and the human brain. Putnam Publishing.

[12] De Vries, A. (2023). The growing energy footprint of artificial intelligence. Joule, 7(10), 2191–2194. https://doi.org/10.1016/j.joule.2023.09.001

[13] Doyon, J., Penhune, V., & Ungerleider, L. G. (2003). Distinct contribution of the cortico-striatal and cortico-cerebellar systems to motor skill learning. Neuropsychologia, 41(3), 252–262. https://doi.org/10.1016/S0028-3932(02)00158-6

[14] Eloundou, T., Manning, S., Mishkin, P., & Rock, D. (2023). GPTs are GPTs: An early look at the labor market impact potential of large language models. arXiv. https://doi.org/10.48550/arXiv.2303.10130

[15] Garcia-Martin, E., Rodrigues, C. F., Riley, G., & Grahn, H. (2019). Estimation of energy consumption in machine learning: A systematic review. Sustainable Computing: Informatics and Systems, 22, 1–16. https://doi.org/10.1016/j.suscom.2019.01.018

[16] GitHub. (2023). Research: Quantifying GitHub Copilot's impact on developer productivity and happiness. GitHub Research. https://github.blog/news-insights/research/research-quantifying-github-copilots-impact-on-developer-productivity-and-happiness/

[17] Goldfarb, A., Taska, B., & Teodoridis, F. (2023). Could machine learning be a general purpose technology? A comparison of historical and current trends. Research Policy, 52(6), 104773. https://doi.org/10.1016/j.respol.2023.104773

[18] Google DeepMind. (2021). Highly accurate protein structure prediction with AlphaFold. Nature, 596(7873), 583–589. https://doi.org/10.1038/s41586-021-03819-2

[19] Google DeepMind. (2024). Gemini. https://deepmind.google/technologies/gemini/

[20] Hennessy, J. L., & Patterson, D. A. (2017). Computer architecture: A quantitative approach (6th ed.). Morgan Kaufmann.

[21] International Energy Agency. (2024). Electricity 2024: Analysis and forecast to 2026. https://www.iea.org/reports/electricity-2024

[22] Kandel, E. R., Schwartz, J. H., Jessell, T. M., Siegelbaum, S. A., & Hudspeth, A. J. (2013). Principles of neural science (5th ed.). McGraw-Hill Education.

[23] LeCun, Y. (2022). A path towards autonomous machine intelligence. OpenReview. https://openreview.net/forum?id=BZ5a1r-kVsf

[24] Lohn, A., & Musser, M. (2022). Mainsprings of AI: The compute simulation game. Center for Security and Emerging Technology. https://doi.org/10.51593/20210031
[25] Maguire, E. A., Gadian, D. G., Johnsrude, I. S., Good, C. D., Ashburner, J., Frackowiak, R. S. J., & Frith, C. D. (2000). Navigation-related structural change in the hippocampi of taxi drivers. Proceedings of the National Academy of Sciences, 97(8), 4398–4403. https://doi.org/10.1073/pnas.070039597
[26] McKinsey & Company. (2023). The economic potential of generative AI: The next productivity frontier. McKinsey Global Institute. https://www.mckinsey.com/capabilities/quantumblack/our-insights/the-economic-potential-of-generative-ai-the-next-productivity-frontier
[27] Meta AI. (2024). Llama. https://ai.meta.com/llama/
[28] Microsoft Research. (2024). Phi small language models technical report. https://www.microsoft.com/en-us/research/project/phi-small-language-models/
[29] Miller, C. (2022). Chip war: The fight for the world's most critical technology. Scribner.
[30] Minsky, M. (1986). The society of mind. Simon & Schuster.
[31] Mistral AI. (2024). Mistral AI. https://mistral.ai
[32] NVIDIA Corporation. (2024). NVIDIA annual report 2024. NVIDIA Investor Relations. https://investor.nvidia.com
[33] OpenAI. (2023). GPT-4 technical report. https://cdn.openai.com/papers/gpt-4.pdf
[34] Patterson, D., Gonzalez, J., Le, Q., Liang, C., Munguia, L.-M., Rothchild, D., So, D. R., Texier, M., & Dean, J. (2021). Carbon emissions and large neural network training. arXiv. https://doi.org/10.48550/arXiv.2104.10350
[35] Rifkin, J. (2011). The third industrial revolution: How lateral power is transforming energy, the economy, and the world. Palgrave Macmillan.
[36] Rosenblum, S., & Karniel, A. (2013). Learning, skill acquisition, and motor control: Neurophysiological perspectives. Frontiers in Human Neuroscience, 7, 1–5. https://doi.org/10.3389/fnhum.2013.00001
[37] Satyanarayanan, M. (2017). The emergence of edge computing. Computer, 50(1), 30–39. https://doi.org/10.1109/MC.2017.9
[38] Schumpeter, J. A. (1942). Capitalism, socialism and democracy. Harper & Brothers.
[39] Sevilla, J., Heim, L., Ho, A., Besiroglu, T., Hobbhahn, M., & Villalobos, P. (2022). Compute trends across three eras of machine learning. Proceedings of the International Joint Conference on Neural Networks (IJCNN), 1–8. https://doi.org/10.1109/IJCNN55064.2022.9892644
[40] Shi, W., Cao, J., Zhang, Q., Li, Y., & Xu, L. (2016). Edge computing: Vision and challenges. IEEE Internet of Things Journal, 3(5), 637–646. https://doi.org/10.1109/JIOT.2016.2579198
[41] Stanford Institute for Human-Centered Artificial Intelligence. (2024). AI Index report 2024. https://aiindex.stanford.edu/report/
[42] Strubell, E., Ganesh, A., & McCallum, A. (2019). Energy and policy considerations for deep learning in NLP. Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, 3645–3650. https://doi.org/10.18653/v1/P19-1355
[43] Susskind, D. (2020). A world without work: Technology, automation, and how we should respond. Metropolitan Books.
[44] The ARM Group. (2023). ARM architecture reference manual. ARM Holdings. https://www.arm.com
[45] Thompson, N. C., Greenewald, K., Lee, K., & Manso, G. F. (2020). The computational limits of deep learning. arXiv. https://doi.org/10.48550/arXiv.2007.05558
[46] Toffler, A. (1980). The third wave. Bantam Books.
[47] TSMC. (2024). TSMC corporate overview. https://www.tsmc.com
[48] United States National Institute of Standards and Technology. (2024). CHIPS for America. https://www.nist.gov/chips
[49] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. Advances in Neural Information Processing Systems, 30. https://papers.nips.cc/paper/7181-attention-is-all-you-need.pdf
[50] Wu, Q., Bansal, G., Zhang, J., Wu, Y., Li, B., Zhu, E., Luo, L., Zhang, T., Cheng, X., Liu, H., Fettiplace, M., Ma, R., Zhou, S., Song, Y., Wang, M., & Wang, C. (2023). AutoGen: Enabling next-generation LLM applications via multi-agent conversation. arXiv. https://doi.org/10.48550/arXiv.2308.08155
[51] xAI. (2024). xAI. https://x.ai
[52] Zheng, L., Chiang, W.-L., Sheng, Y., Zhuang, S., Wu, Z., Zhuang, J. E., & Stoica, I. (2024). LMSYS-Chat-1M: A large-scale real-world LLM conversation dataset. International Conference on Learning Representations (ICLR).