

THE THREE CLOCKS

MOORE'S ILLUSION

When Intelligence Outpaces Its Substrate

NABIL EL MAHYAOUI

Moore's Illusion

Moore's Illusion

When Intelligence Outpaces Its Substrate

Nabil El Mahyaoui

A must-read ebook for strategists, engineers, and policymakers

Copyright © 2026 Nabil El Mahyaoui.
All rights reserved.

Contents

1 The Chart That Ate the World **1**

1.1 The Three Clocks 4

2 The Silicon Clock **7**

2.1 The Death of Free Performance 9

2.2 The Economic Wall 11

2.3 The Ghost in the Fab 12

2.4 What Moore's Law Actually Left Us. . . 14

3 The Intelligence Clock **17**

3.1 The Scaling Revelation 18

3.2 The Decoupling 20

3.3 The Hunger 23

3.4 The Bitter Lesson, Revisited. 24

4 The Energy Clock 27

4.1 The Pace of Atoms 29

4.2 The Cloud Is Made of Coal 31

4.3 The Jevons Trap 33

4.4 The Twenty-Watt Benchmark. 35

5 The Infrastructure Gap 38

5.1 The Temporal Mismatch 39

5.2 The Water Nobody Mentions. 41

5.3 When Tech Companies Become Utilities 43

5.4 The Dark Silicon Metaphor, Writ Large 45

6 The Geopolitics of the Joule 48

6.1 The Taiwan Singularity 49

6.2 The Gulf States' Pivot 52

6.3 Europe's Strategic Paralysis 54

6.4 The New Colonial Calculus 57

6.5 The FLOPS-to-Joules Translation 59

7 The Vampire's Arithmetic 62

7.1 The Last Five Percent 63

7.2 The Inference Tax 65

7.3 The Rebound Loop 67

7.4 The Honest Accounting. 69

8 After the Illusion 72

8.1 The Efficiency Imperative 73

8.2 The Re-materialisation of Tech. 76

8.3 The Three Futures. 78

8.4 Redefinition 82

References **84**

1

The Chart That Ate the World

In 1965, Gordon Moore, co-founder of Fairchild Semiconductor and later Intel, published a four-page paper in *Electronics* magazine. It contained a single observation: the number of components on an integrated circuit had been doubling approximately every year, and he expected this trend to continue for at least a decade.

That was it. No grand theory. No mathematical proof. No invocation of physical law. Just a man reading a graph and drawing a line.

That line became the most consequential prediction in the history of technology. It was not a law of nature. It was something far more powerful: a *self-fulfilling prophecy backed by trillions of dollars of coordinated investment*.

For fifty years, the entire semiconductor industry—chip designers, equipment manufacturers, materials scientists, software developers—organised itself around Moore's prediction. TSMC built fabs. Intel designed architectures. ASML invented lithography machines. Qualcomm designed mobile chips. Every player knew the tempo: double the transistors, halve the cost, ship on schedule. The roadmap was published years in advance. Competitors collaborated on pre-competitive

research because everyone benefited from the shared rhythm.

The result was the most sustained exponential improvement in any technology in human history. Between 1965 and 2020, the number of transistors on a chip increased by a factor of roughly ten billion. The cost per transistor fell by a factor of fifty million. The computational power available to an average citizen in 2025 would have ranked among the world's top supercomputers in 1995.

And here is the crucial part: this exponential improvement in silicon *leaked* into our collective psychology. It became the template for how we think about progress itself.

> *"We've been living inside an exponential for so long that we've mistaken it for the shape of reality."*

We began to assume that everything would follow this curve. That energy would get cheaper. That medicine would advance at digital speed. That economic growth would compound indefinitely. That the future would always be better, faster, more abundant than the present.

This is Moore's Illusion: the cognitive bias that extrapolates one industry's extraordinary anomaly into a universal law of progress.

The illusion is now shattering, because three fundamental curves—silicon, intelligence, and energy—are diverging at speeds that make collision inevitable.

1.1 The Three Clocks

Imagine three clocks on a wall, each ticking at a radically different speed.

The first clock is **silicon**. For fifty years it ticked with metronomic regularity: double the transistors every two years. But the ticking is slowing. The intervals between ticks are stretching. The clock is running down. This is the logarithmic clock—still advancing, but with each step smaller than the last.

The second clock is **artificial intelligence**. It began ticking slowly, almost imperceptibly, in the academic labs of the 1990s. Then it accelerated. And accelerated. And accelerated. It now ticks so fast that its movements blur. AI capability doubles not every two years, not every year, but every *three to six months*. This is the hyper-exponential clock—and it is demanding more from its substrate than the substrate can provide.

The third clock is **energy**. It ticks at the pace of physics, of civil engineering, of geological

time. A power plant takes a decade to build. A battery improves by five to eight percent per year. A transmission line requires right-of-way negotiations, environmental reviews, and physical copper strung across physical land. This clock does not care about venture capital timelines or product launch dates. It ticks at the speed of atoms, not bits.

Three clocks. Three speeds. One collision.

This book is about that collision—what caused it, what it means, and what happens when an unstoppable intelligence meets an immovable physics.

2

The Silicon Clock

Let us be precise about what died and when it died, because the mythology surrounding Moore's Law has obscured the actual history to the point of uselessness.

Moore's Law, in its original formulation, was an *economic* observation: the cost-optimal number of transistors on a chip doubles every two years. It said nothing about performance. It said nothing about power consumption. It

said nothing about clock speed. It was, at its core, a statement about manufacturing yield curves and the economics of photolithography.

The performance improvements we associate with Moore's Law—the reason your phone is faster than last year's phone—came from a *different* observation entirely: Dennard Scaling, formulated by Robert Dennard at IBM in 1974. Dennard's insight was that as transistors shrink, they consume proportionally less power, which means you can pack more of them together *and* run them faster without overheating.

Dennard Scaling died in 2006.

2.1 The Death of Free Performance

The implications of Dennard's death cannot be overstated. Before 2006, software developers could write inefficient code and simply wait. The next generation of chips would run it faster, for free. Bloated enterprise software, poorly optimised games, memory-leaking web browsers—all of it was forgiven by the relentless march of clock speed. From 1985 to 2004, CPU clock speeds rose from 5 MHz to 3.8 GHz—a 760x improvement in two decades.

Then it stopped. The clock speed of a typical desktop processor in 2006 was approximately 3 GHz. The clock speed of a typical desktop processor in 2026 is approximately 5 GHz. In twenty years, we gained less than a factor of

two. Compare that to the 760x of the previous twenty years.

What happened? Physics intervened. Below the 65-nanometre process node, transistors began to leak current even when switched off. The power density—watts per square centimetre—hit a wall. Intel's engineers called it the "power wall." If they continued to shrink transistors and increase clock speed at the historical rate, the chip would have reached the power density of a nuclear reactor by 2010 and the surface of the sun by 2015.

So they stopped trying. Instead, they pivoted to parallelism: multi-core processors. Your laptop has eight cores not because eight cores are better than one fast core for most tasks, but because one fast core is *physically impossible* at the performance level we need. Multi-core is

a brilliant engineering workaround. It is also an admission of defeat.

2.2 The Economic Wall

Even after Dennard Scaling died, transistor *count* continued to increase. Moore's Law, in its narrow economic formulation, limped on. But the costs shifted dramatically.

In 2000, building a state-of-the-art chip fabrication plant cost approximately $1 billion. By 2015, the figure was $10 billion. TSMC's latest 2nm fab in Arizona is projected to cost $40 billion. Samsung's next-generation facility will exceed $30 billion.

The number of companies that can afford to play this game is shrinking. In 1990, there were over 20 companies manufactur-

ing leading-edge chips. Today, there are three: TSMC, Samsung, and Intel. By the time 1nm processes arrive, there may be one.

This is not a competitive market. It is an oligopoly racing toward monopoly, with the fixed costs of entry rising faster than the revenues they generate.

The cost per transistor, which had fallen steadily for fifty years, *stopped falling* at the 28nm node around 2014. For anything below 7nm, the cost per transistor is actually *increasing*. Moore's Law, in its original economic sense, is not just slowing. It has reversed.

2.3 The Ghost in the Fab

There is a deeper story here, one that rarely makes it into the business press: the stagger-

ing physical complexity of modern chip fabrication.

A modern semiconductor fab is the most complex manufacturing environment ever created by human beings. The air is cleaner than an operating theatre by a factor of ten thousand. The lithography machines—built exclusively by ASML in the Netherlands—use extreme ultraviolet light generated by vaporising droplets of molten tin with a laser, fifty thousand times per second, to project patterns onto silicon wafers with a precision of 13.5 nanometres.

A single EUV lithography machine costs $380 million. It weighs 180 tonnes. It contains 100,000 parts sourced from 5,000 suppliers across 30 countries. There are approximately 200 of these machines in the world. If the single ASML factory in Veldhoven were de-

stroyed, the entire trajectory of advanced computing would halt.

This is not resilience. This is fragility dressed in the language of precision engineering.

2.4 What Moore's Law Actually Left Us

The legacy of Moore's Law is not, as commonly believed, infinite exponential progress. It is something more nuanced and more dangerous: a *mental model* that assumes all hard problems yield to exponential technological improvement, given sufficient time and investment.

This mental model is embedded in venture capital pitch decks ("we're riding the exponential curve"), in government policy ("technology will solve climate change"), in corpo-

rate strategy ("we'll optimise our way out of this"), and in popular culture ("the singularity is near").

The mental model was accurate for exactly one domain—semiconductor transistor density—for exactly one fifty-year period. It was never accurate for energy. It was never accurate for biology. It was never accurate for the physical infrastructure on which digital systems depend.

And now it is no longer accurate even for semiconductors.

We are, collectively, operating with an outdated map. The territory has changed. The exponential is over. And the institutions, incentives, and assumptions that were built on its continuation are about to discover what it

means when the ground shifts beneath their feet.

3

The Intelligence Clock

Something happened in 2017 that the semiconductor industry did not anticipate and, to this day, has not fully absorbed.

A team of researchers at Google Brain published a paper titled "Attention Is All You Need." It introduced the Transformer architecture—a neural network design that, for the first time, could process sequences of

arbitrary length by attending to all positions simultaneously rather than step by step.

The paper was eight pages long. It has since become the most consequential technical document of the 21st century. Not because the Transformer is elegant (it is), or because attention is a powerful mechanism (it is), but because the Transformer did something no previous architecture had done: it *scaled*.

3.1 The Scaling Revelation

Previous neural network architectures hit diminishing returns quickly. You could make an LSTM larger, but beyond a certain point, it simply failed to learn. Convolutional networks scaled well for images but plateaued for language. The prevailing wisdom in 2016 was that neural networks were useful for nar-

row tasks—image classification, speech recognition, machine translation—but fundamentally limited in their capacity for general reasoning.

The Transformer shattered this assumption.

In 2020, OpenAI researchers Jared Kaplan and colleagues published the "Scaling Laws" paper. Its finding was simple and devastating: the performance of a Transformer model improves as a *smooth, predictable power law* of three variables—the number of parameters, the quantity of training data, and the amount of compute invested. There was no plateau in sight. No diminishing returns at any scale tested. Just a clean, relentless line climbing upward on a log-log plot.

This was not supposed to happen. Every previous technology had hit a ceiling. Every

previous scaling curve had bent. The Transformer's scaling curve, as far as anyone could measure, did not.

The AI research community collectively realised that intelligence might be—at least in part—a *commodity*. Something you could buy with money, like electricity or steel. Double the compute, get predictably better performance. Triple it, get more. The only constraint was the budget.

3.2 The Decoupling

Here is the critical insight for our three-clock framework: AI improvement has *decoupled* from hardware improvement.

During the Moore's Law era, software performance was tethered to hardware performance.

If you wanted a faster program, you waited for a faster chip. The two curves—hardware capability and software capability—moved in lockstep, synchronized by the semiconductor industry's roadmap.

AI broke this coupling in three ways.

Algorithmic efficiency. The compute required to train an ImageNet-class model fell by a factor of 44x between 2012 and 2019, even as hardware improved by only 11x over the same period. The gap was closed by better algorithms—better architectures, better optimisers, better training recipes. Pure intellectual progress, requiring no new silicon.

Data leverage. A model trained on more data, or better-curated data, outperforms a model trained on less data, regardless of model size. The "Chinchilla" paper from DeepMind

(2022) demonstrated that most existing models were dramatically over-parameterised and under-trained. By simply rebalancing compute allocation—spending more on data, less on parameters—they achieved equivalent performance at a fraction of the cost. No hardware improvement was involved. Just a better use of the hardware that already existed.

Architectural innovation. Mixture-of-Experts models, flash attention, speculative decoding, quantisation techniques—each of these innovations delivered 2–10x improvements in inference efficiency without any change to the underlying hardware. The software was getting better at *using* the hardware, faster than the hardware was getting better at *being* hardware.

The result: AI capability is now improving at a rate of roughly 4–10x per year. Hardware ca-

pability is improving at roughly 1.4x per year. The gap between these two rates is the defining tension of the current technological moment.

3.3 The Hunger

The Transformer's blessing is also its curse: it scales. Which means it is *hungry*.

The compute required to train frontier AI models is doubling every 3.4 months—a rate that makes Moore's Law look sedentary. GPT-2 (2019) required approximately 10^{20} floating-point operations. GPT-4 (2023) required approximately 10^{25}. The next generation of frontier models will require 10^{26} to 10^{27}.

Each order of magnitude increase in compute requires a corresponding increase in hardware, electricity, cooling, and data centre in-

frastructure. The AI industry is not merely consuming the surplus produced by Moore's Law. It is *exhausting* it, drawing down the accumulated reserves of cheap compute faster than the semiconductor industry can replenish them.

This creates a scissors dynamic. The AI curve is accelerating upward. The hardware curve is decelerating. The two curves, which once ran in parallel, are now diverging—the intelligence clock running faster and faster, the silicon clock slowing to a crawl, and the gap between them widening with every quarter.

3.4 The Bitter Lesson, Revisited

In 2019, the AI researcher Rich Sutton published a short essay titled "The Bitter Lesson." His argument was that, historically, the AI

methods that won in the long run were always the ones that leveraged more computation, not the ones that incorporated more human knowledge. Clever hand-crafted features lost to brute-force learning. Domain expertise lost to scale. Every time researchers tried to "be smart" about AI, they were eventually beaten by researchers who simply "threw more compute at it."

The Bitter Lesson was prescient. It correctly predicted the trajectory from narrow AI to large language models. But it contained an implicit assumption that Sutton did not examine: that compute would *always be available*. That the exponential supply of cheap transistors would continue to fuel the exponential demand for computation.

That assumption is Moore's Illusion in its purest form.

The Bitter Lesson tells us that intelligence wants more compute. The Silicon Clock tells us that more compute is becoming harder to supply. The result is not a crisis of intelligence—the algorithms will continue to improve—but a crisis of *access*. Intelligence will become stratified: abundant for those who can afford the energy and hardware to run frontier models, scarce for everyone else.

This is not a theoretical future. It is already happening. The cost of training a frontier model now exceeds $100 million. Only a handful of organisations on Earth—OpenAI, Google, Anthropic, Meta, a few Chinese labs—can afford to play. The democratisation of AI that characterised the 2015–2020 era is ending. Intelligence is re-centralising, not because of regulation or ideology, but because of physics.

4

The Energy Clock

There is a sentence that every AI enthusiast should be required to tattoo on the inside of their eyelids: *you cannot download a kilowatt-hour*.

The digital revolution was, at its core, a revolution in *dematerialisation*. Music became files. Books became bytes. Maps became data. Commerce became transactions in a database. The pattern was consistent and intoxicating: take

a physical thing, strip away its atoms, replace them with bits, and watch the marginal cost fall to zero.

This worked spectacularly well for information goods. It does not work for energy.

A lithium-ion battery is not a file. You cannot compress it. You cannot stream it. You cannot replicate it at zero marginal cost. It is a physical object containing specific chemical compounds—lithium, cobalt, nickel, manganese—that must be mined from the earth, refined in chemical plants, assembled in factories, and shipped in trucks. Every step involves atoms, not bits. Every step is governed by thermodynamics, not Moore's Law.

4.1 The Pace of Atoms

The improvement rate of energy technologies is not exponential. It is linear, and modestly so.

Lithium-ion battery energy density has improved from approximately 80 Wh/kg in 1991 to approximately 300 Wh/kg in 2025. That is a 3.7x improvement in 34 years, or roughly 4% per year. Compare this to transistor density, which improved by *billions of times* over a comparable period.

Solar panel efficiency has improved from approximately 15% to approximately 24% for commercial panels over the past two decades. That is a 1.6x improvement. Meaningful, but nowhere near exponential.

Wind turbine capacity has grown, but through physical scaling—building larger turbines—not through any fundamental improvement in the efficiency of converting kinetic energy to electricity. The Betz limit, derived from basic fluid dynamics, caps the theoretical maximum at 59.3%. We are already at 45–50% in practice. The physics allows no further dramatic improvement.

Nuclear fission is, thermodynamically, extraordinarily efficient. But the *time* required to build a nuclear power plant has, if anything, *increased* over the past three decades, due to regulatory complexity, safety requirements, and the atrophy of construction expertise. The average construction time for a new nuclear reactor in the West is 10–15 years. In some cases, notably the Vogtle plant in Georgia and Flamanville in France, it has exceeded 15 years.

The energy clock ticks at the pace of chemistry, civil engineering, and regulatory processes. It does not care about product cycles. It does not respond to demand signals with the agility of a software company pushing a patch. It operates on the timescale of *infrastructure*—the slowest, heaviest, most capital-intensive layer of civilisation.

4.2 The Cloud Is Made of Coal

One of the most successful marketing achievements of the technology industry was the invention of the word "cloud." It evokes lightness, weightlessness, abstraction—data floating ethereally in the sky, accessible from anywhere, unencumbered by the constraints of the physical world.

The cloud is not light. The cloud is *heavy*.

A hyperscale data centre consumes 100–200 megawatts of electrical power. That is the output of a small power plant, running 24 hours a day, 365 days a year, dedicated to keeping servers cool and computations running.

In 2024, data centres consumed approximately 4% of total U.S. electricity generation. By 2028, projections suggest this will exceed 8%. The International Energy Agency estimates that global data centre electricity consumption will more than double between 2024 and 2030.

This is not a marginal load. It is a *structural transformation* of the electrical grid. And it is being driven almost entirely by AI.

A single query to a large language model consumes approximately 10x the electricity of a traditional search engine query. As AI is em-

bedded into every product—every search, every email, every document, every customer service interaction—the cumulative energy demand compounds with terrifying speed.

The cloud is not made of vapour. It is made of silicon, copper, concrete, diesel generators, cooling towers, and an ever-growing appetite for electrons. Demystifying this is not pedantry. It is essential for understanding the constraint that will shape the next decade of technology.

4.3 The Jevons Trap

In 1865, the English economist William Stanley Jevons observed a paradox that bears his name: the introduction of more efficient steam engines did not reduce coal consumption. It *increased* it. Because the engines were more

efficient, they were more economical to operate, which made them more widely adopted, which drove total consumption higher even as per-unit consumption fell.

AI is walking directly into the Jevons Trap.

Every improvement in model efficiency—smaller models, better quantisation, more efficient inference engines—makes AI cheaper to deploy. Cheaper deployment means wider adoption. Wider adoption means more total compute. More total compute means more energy.

The industry celebrates each efficiency breakthrough as a step toward "sustainable AI." It is the opposite. Each efficiency breakthrough *accelerates* total energy consumption by making AI viable for use cases that were previously too expensive.

When inference costs dropped by 10x between 2023 and 2025, the response was not to use the same amount of AI at one-tenth the cost. The response was to use 50x more AI. The net effect was a 5x *increase* in total energy consumption for AI workloads.

This is not a market failure. It is rational behaviour in a market that correctly prices compute but does not price carbon, water, or grid stability. The Jevons Trap is not a trap for the company or the consumer. It is a trap for the *planet*.

4.4 The Twenty-Watt Benchmark

The human brain operates on approximately 20 watts. Twenty watts. The power consumption of a dim light bulb.

On those 20 watts, the brain performs feats of perception, reasoning, creativity, emotional processing, motor coordination, and self-awareness that no artificial system can yet match across all dimensions simultaneously.

A frontier AI model, performing a subset of these tasks—primarily language and reasoning—consumes *millions* of watts during training and thousands of watts during inference. The gap between biological intelligence and artificial intelligence, measured in watts per cognitive operation, is roughly six orders of magnitude.

This is not merely an engineering problem. It is a *thermodynamic indictment* of our current approach to AI. We have built an intelligence that is, watt for watt, a million times less efficient than the one evolution produced. And we are

attempting to scale it up by brute-forcing our way through the energy constraint.

The brain did not achieve its efficiency by accident. It achieved it through four billion years of evolutionary pressure, operating under the merciless constraint of caloric scarcity. Every synapse, every neuron, every neural circuit was optimised not for maximum performance, but for maximum performance *per joule*.

The AI industry has never faced this constraint. Until now.

5

The Infrastructure Gap

We have now established the three clocks and the divergence between them. This chapter examines what happens at the point of collision: the physical infrastructure that must mediate between AI's insatiable demand and energy's stubborn limits.

The gap is not metaphorical. It is measured in megawatts, in months of construction time, in millions of litres of cooling water, and in kilo-

metres of transmission cable that do not yet exist.

5.1 The Temporal Mismatch

The most dangerous number in the current AI landscape is not a parameter count or a benchmark score. It is a *ratio of timescales*.

Time to train a frontier AI model: 3–6 months.
Time to build a data centre: 18–24 months.
Time to secure a power purchase agreement: 12–36 months.
Time to build a solar farm: 2–4 years.
Time to build a nuclear reactor: 10–15 years.
Time to upgrade a national electrical grid: 15–30 years.

AI operates on the timescale of software. Its infrastructure operates on the timescale of *civil*

engineering. The gap between these timescales is not closing. It is widening, because AI development is accelerating while infrastructure construction is not.

This creates a structural deficit. The AI industry is placing orders for electricity that the energy industry cannot deliver on the required timeline. Microsoft, Google, Amazon, and Meta are collectively planning data centre expansions that will require an estimated 40–60 additional gigawatts of power capacity by 2030. For reference, the entire installed generation capacity of a country like the Netherlands is approximately 40 GW.

The hyperscalers are, in effect, requesting the energy output of a mid-sized European country, delivered within five years, dedicated exclusively to computation. This is not an engineering challenge. It is a *civilisational negotia-*

tion between the speed of bits and the inertia of atoms.

5.2 The Water Nobody Mentions

Energy dominates the infrastructure discussion, but water is the silent constraint.

Modern data centres use evaporative cooling systems that consume vast quantities of water. A single large data centre—the kind operated by Google, Microsoft, or Amazon—consumes approximately 3–5 million litres of water per day. In a year, that is 1–2 billion litres. Per facility.

Microsoft's water consumption increased by 34% between 2021 and 2022, a period that coincided precisely with the scaling of its AI in-

frastructure. Google's water consumption increased by 20% over the same period.

These data centres are not randomly distributed. They are concentrated in regions with cheap electricity and fibre connectivity—often the same regions facing water stress. Northern Virginia, which hosts the largest concentration of data centres in the world, draws from the Potomac River watershed. Phoenix, Arizona—home to a growing cluster of data centres including TSMC's new fab—is in a desert that has been in drought conditions for over two decades.

The irony is architectural: we are building the infrastructure of artificial intelligence in places where the natural resource it most depends on, after electricity, is running out.

5.3 When Tech Companies Become Utilities

The infrastructure gap has forced a remarkable transformation in the identity of technology companies. Microsoft, Google, and Amazon are no longer merely software and services companies. They are becoming *energy companies*.

Microsoft signed a 20-year power purchase agreement to restart the Three Mile Island nuclear reactor—the same facility whose partial meltdown in 1979 effectively killed the American nuclear industry for a generation. Google has invested in small modular reactor (SMR) technology. Amazon has acquired data centre campuses that come with their own on-site power generation.

These are not peripheral investments. They are existential commitments. The hyperscalers have recognised that their ability to serve AI workloads is gated not by their software capabilities, not by their talent pool, not by their product design, but by *how many watts they can secure*.

The competitive advantage in AI is migrating from the model layer to the infrastructure layer. The company that can deliver the most compute per dollar will win—and that computation is ultimately denominated in joules.

This has profound implications for market structure. Building a nuclear reactor is not a startup activity. Securing gigawatt-scale power capacity is not something a Series B company can do. The infrastructure requirements of frontier AI are creating *natural monopoly conditions*—not because of network

effects or regulatory capture, but because of the sheer capital intensity of the physical substrate.

The AI industry is recapitulating the history of the electrical utility industry, but on a compressed timeline. Just as the early 20th century saw the consolidation of hundreds of small electricity providers into a handful of regulated monopolies, the early 21st century is witnessing the consolidation of AI infrastructure into a handful of hyperscalers with the capital and the energy access to operate at scale.

5.4 The Dark Silicon Metaphor, Writ Large

In semiconductor design, there is a concept called "dark silicon": the transistors on a chip that must remain powered off at any given moment because the chip cannot dissipate the

heat that would be generated if all transistors were active simultaneously. In modern processors, as much as 50–80% of the chip is "dark" at any given time.

This metaphor scales to civilisation.

We are building computational capacity that we cannot fully power. The data centres are designed, the servers are ordered, the AI models are waiting to run. But the electricity is not there. The grid interconnection is pending. The power plant is still under construction. The transmission line has not been approved.

We have built a civilisation of dark silicon—infrastructure that exists on paper, in purchase orders, and in corporate planning documents, but that cannot be activated because the energy substrate has not kept pace.

The AI industry's response has been to move faster on the infrastructure it *can* control (data centres, networking, cooling) and to lobby furiously for acceleration of the infrastructure it *cannot* control (grid interconnection, permitting, power generation). The tension between these two timescales is the defining industrial policy challenge of the decade.

6

The Geopolitics of the Joule

The 20th century was shaped by the geopolitics of oil. Wars were fought over it. Alliances were forged around it. The map of global power was, in large part, a map of petroleum reserves and the shipping lanes that connected them to markets.

The 21st century will be shaped by the geopolitics of *compute-per-joule*: the intersection of chip fabrication capacity, energy surplus, and

the AI capability they jointly enable. The nations that control this intersection will dominate. Everyone else will be tenants in someone else's cognitive infrastructure.

6.1 The Taiwan Singularity

TSMC—the Taiwan Semiconductor Manufacturing Company—produces approximately 90% of the world's most advanced semiconductors. Not 90% of all chips; 90% of the chips that matter most: the 5nm, 3nm, and soon 2nm processors that power every frontier AI model, every high-end smartphone, every advanced military system.

This concentration is not the result of market failure. It is the result of decades of deliberate industrial policy by Taiwan, combined with the extraordinary difficulty of replicating

TSMC's manufacturing expertise. The company's process technology, its yield optimisation, its supply chain relationships, and its institutional knowledge represent a moat that no competitor has been able to cross despite tens of billions of dollars in investment.

The strategic implications are staggering.

A Chinese military action against Taiwan—whether invasion, blockade, or even targeted strikes against fab facilities—would not merely disrupt the chip supply chain. It would *crash* the entire computational trajectory of the Western world. Every AI training run would slow. Every new product launch would stall. Every defence system that depends on advanced processors would face an indefinite supply interruption.

TSMC's fabs in Hsinchu are, by any rational strategic calculus, the most valuable industrial facilities on Earth. Their destruction or capture would represent a greater economic shock than the 1973 oil embargo, the 2008 financial crisis, and the COVID-19 pandemic combined.

This is why the United States has invested $52 billion through the CHIPS Act to subsidise semiconductor fabrication on American soil. It is why TSMC is building fabs in Arizona and Japan. It is why Samsung is expanding in Texas. The strategic logic is simple: *de-risk the supply of the substrate on which intelligence depends*.

But de-risking takes time. TSMC's Arizona fab, originally announced in 2020, is not expected to produce leading-edge chips until 2028 at the earliest. In the meantime, the Tai-

wan Singularity persists: a single island, a single company, a single point of failure for the future of artificial intelligence.

6.2 The Gulf States' Pivot

While the West agonises over semiconductor supply chains, a different strategic realignment is unfolding in the Persian Gulf.

Saudi Arabia, the United Arab Emirates, and Qatar are executing the most ambitious economic pivot in modern history: the transition from petrostates to *compute-states*. The logic is elegant and, in its way, inevitable.

These nations possess three strategic assets that are perfectly aligned with the AI era: abundant cheap energy (both fossil and solar), sovereign wealth funds with trillions in

deployable capital, and geographic positions at the crossroads of Asia, Europe, and Africa.

The UAE's investment in AI is not incidental. It is *strategic doctrine*. Abu Dhabi's Technology Innovation Institute built Falcon, one of the world's largest open-source language models. Saudi Arabia is building NEOM with AI infrastructure at its core. Qatar is positioning itself as a neutral data haven, offering sovereignty guarantees that neither American nor Chinese cloud providers can match.

The irony is exquisite. The same fossil fuels that powered the industrial age—the hydrocarbons that Moore's Illusion told us were becoming irrelevant—may power the AI age. Not through factories and automobiles, but through data centres that convert natural gas into computation.

The petrostates understood something that the technology industry was slow to grasp: in a world where intelligence is a commodity, the scarce resource is not the algorithm. It is the *energy to run the algorithm*. And they have more of it than anyone.

6.3 Europe's Strategic Paralysis

The European Union finds itself in a position of remarkable strategic impotence, and the diagnosis is unflattering.

Europe has no frontier AI lab. No European company is training models at the scale of GPT-4 or Gemini. DeepMind, the continent's most significant AI achievement, was acquired by Google in 2014 and is now an American subsidiary operating under American corporate governance.

Europe has no leading-edge chip fabrication. ASML, the Dutch lithography monopoly, is a critical supplier—but it supplies *equipment*, not chips. The chips themselves are made in Taiwan, South Korea, and increasingly the United States. Europe is essential to the supply chain but absent from the product.

Europe has an energy deficit. Germany's decision to abandon nuclear power, combined with the continent's dependence on Russian natural gas (now severely curtailed), has left the EU with some of the highest electricity prices in the developed world. A European data centre pays 2–3x more per kilowatt-hour than an American one. This is not a regulatory problem. It is a *physics* problem that makes large-scale AI infrastructure economically uncompetitive.

What Europe does have is regulation. The AI Act, the Digital Markets Act, the Digital Services Act, the GDPR—an impressive apparatus of normative governance. But regulation without production capacity is not sovereignty. It is commentary.

The harsh strategic reality is this: Europe regulates AI it does not build, running on chips it does not fabricate, powered by energy it does not have in surplus. This is not a position of influence. It is a position of *dependency*—disguised, for now, by the regulatory leverage that comes from being a large consumer market, but increasingly irrelevant as the strategic center of gravity shifts to those who control the physical substrate.

6.4 The New Colonial Calculus

For the Global South, the geopolitics of compute presents a binary future.

Path one: leapfrog. Just as African nations skipped landlines and went directly to mobile telephony, developing nations could skip the SaaS/cloud era and go directly to AI-native infrastructure. Build data centres powered by abundant solar energy (the Sahara receives more solar radiation in six hours than humanity consumes in a year). Develop local AI capabilities tailored to local languages, local markets, local needs. The absence of legacy infrastructure becomes an advantage.

Path two: cognitive colonialism. Developing nations become consumers of intelligence produced elsewhere—dependent on American cloud providers, Chinese platforms, or

Gulf State data centres for access to AI capabilities. Every query, every inference, every decision support system runs through foreign infrastructure, subject to foreign terms of service, foreign pricing, and foreign geopolitical interests.

This is not hyperbole. If compute becomes as essential as electricity—and every indicator suggests it will—then nations without sovereign compute infrastructure become *cognitively dependent* in the same way that nations without refinery capacity are dependent on imported fuel.

The colonial era was defined by the extraction of physical resources: minerals, crops, labour. The AI era may be defined by the extraction of *cognitive resources*: data, attention, decision-making authority. The mechanism is different. The structural dynamic is identical.

6.5 The FLOPS-to-Joules Translation

The geopolitical community has not yet adapted its analytical frameworks to the AI era. Strategy is still discussed in terms of GDP, military spending, population, and natural resources. These metrics remain important, but they are increasingly incomplete.

The missing metric is *sovereign compute capacity*: the total AI inference and training capability that a nation can deploy using its own hardware, powered by its own energy, under its own legal and operational control.

This metric is a function of three variables:

1. **Chip access.** Can the nation procure or fabricate the processors needed for AI workloads?

2. **Energy surplus.** Does the nation generate more electricity than its current demand, with capacity available for AI infrastructure?

3. **Skilled labour.** Does the nation have the engineers capable of building and operating AI systems at scale?

By this measure, the United States is dominant but vulnerable (chip access dependent on Taiwan, energy grid under strain). China is capable but constrained (chip access limited by export controls, energy grid coal-dependent). The Gulf States are ascendant (energy surplus, capital surplus, talent importable). Europe is declining (no chips, no energy surplus, talent emigrating). The Global South is at risk.

The geopolitics of the 21st century will not be determined by who has the most oil, the

largest army, or the highest GDP. It will be determined by who can convert the most joules into the most intelligence, at the lowest cost, under sovereign control.

7

The Vampire's Arithmetic

We arrive at the most uncomfortable calculation in the AI landscape: the relationship between energy input and cognitive output, and the diminishing returns that govern it.

The metaphor of the vampire is deliberate. A vampire consumes exponentially more than it produces. It thrives in darkness—in the opacity of data centre energy bills, in the careful omission of carbon figures from AI product

launches, in the industry's collective refusal to publish comprehensive energy audits. And like all vampires, it is sustained by a host that does not fully understand the nature of the bargain.

7.1 The Last Five Percent

In many practical domains, AI has already achieved 90–95% of expert-level performance. A large language model can draft a legal contract that is 93% as accurate as one produced by a junior associate. It can diagnose a skin lesion with 94% of a dermatologist's accuracy. It can write functional code that passes 90% of test cases on the first attempt.

The instinct of the AI industry is to close the remaining gap. To push from 93% to 98%, from 94% to 99%. To achieve "superhuman" perfor-

mance, or at least parity, across every dimension.

But the cost curve is not linear. It is *exponential in the wrong direction*.

Going from 80% to 90% accuracy might require a 10x increase in model size and training compute. Going from 90% to 95% requires another 10x. Going from 95% to 99% requires yet another 10x. Each incremental percentage point of improvement costs multiplicatively more than the last.

This is the regime of diminishing returns, and it has profound economic consequences.

A model that is 95% as good as a human expert and costs $0.01 per query is economically devastating to the human expert. A model that is 99% as good but costs $1.00 per query is economically irrelevant. The marginal value of

the last 4% is swamped by the marginal cost of achieving it.

The industry's fixation on frontier performance—on benchmark leaderboards, on ever-larger models, on the relentless pursuit of the last few percentage points—is therefore not just scientifically questionable. It is *economically irrational*, at least from the perspective of energy efficiency.

7.2 The Inference Tax

Training gets the headlines. Inference pays the bills.

The energy cost of training a model is finite: it happens once (or a few times, with subsequent fine-tuning). The energy cost of *inference*—of actually running the model to pro-

duce outputs—is recurring and scales with usage.

By 2025, inference already accounts for approximately 60–70% of total AI-related energy consumption, and this share is growing rapidly. As AI is embedded into every search query, every email draft, every code suggestion, every customer service interaction, the cumulative inference load compounds.

Consider the scale. Google processes approximately 8.5 billion search queries per day. If each query triggers an AI inference costing 10x the energy of a traditional search, the daily energy overhead is equivalent to powering a city of 500,000 people. And search is just one application among thousands.

This is the inference tax: a permanent, growing energy surcharge on every digital interac-

tion, levied not by a government but by the thermodynamic requirements of running neural networks at global scale.

The tax is currently hidden in the cloud providers' operating budgets. It will not remain hidden. As energy costs rise and grid capacity tightens, the inference tax will be passed through to consumers—either directly, through higher prices for AI-powered services, or indirectly, through higher electricity bills for everyone sharing the grid.

7.3 The Rebound Loop

The Jevons Paradox, discussed in the previous chapter, operates at the level of aggregate demand. But there is a more insidious feedback loop operating at the level of *model development itself*.

Larger models require more energy to train. Once trained, they generate more valuable outputs. More valuable outputs generate more revenue. More revenue funds the training of even larger models. Which require even more energy.

This is the rebound loop: a positive feedback cycle in which each generation of AI models is both the product and the cause of increased energy consumption.

The loop is accelerated by competition. If OpenAI trains a model at 10^{25} FLOPS and achieves a performance breakthrough, Google and Anthropic cannot afford to *not* train a comparable model. The competitive dynamics of the AI industry ensure that any efficiency gained at the model level is immediately reinvested in scale, not in conservation.

There is no natural brake on this loop. The market rewards performance, not efficiency. Investors fund the most capable model, not the most energy-efficient one. Customers choose the best output, not the greenest process. The rebound loop will continue until it hits an external constraint: either the price of energy rises enough to change the economics, or the physical capacity of the grid imposes a hard ceiling on total computation.

7.4 The Honest Accounting

The AI industry's environmental disclosures are, to put it charitably, incomplete.

Most companies report Scope 1 and Scope 2 emissions: the direct emissions from their operations and the emissions from the electricity

they purchase. These figures are non-trivial but manageable.

What they do not report comprehensively is the full lifecycle energy cost of the AI supply chain: the energy required to manufacture the GPUs (semiconductor fabrication is extraordinarily energy-intensive), the energy required to build the data centres, the energy required to manufacture and install the cooling systems, the embodied energy in the rare earth minerals that go into every server.

A comprehensive lifecycle analysis of a single frontier AI training run would include: the energy to mine the lithium, cobalt, and rare earths; the energy to refine them; the energy to manufacture the chips; the energy to assemble the servers; the energy to build the data centre; the energy to run the training; the energy to

cool the facility; and the energy to eventually decommission and recycle the hardware.

No company has published this full accounting. The reason is not technical. The data exists. The reason is that the number would be so large, and so unflattering, that it would fundamentally alter the public narrative around AI as a "clean" technology.

AI is not a clean technology. It is an energy technology—one of the most energy-intensive ever devised by human civilisation, rivalling aluminium smelting and cryptocurrency mining in its appetite for electrons.

Acknowledging this is not defeatism. It is a prerequisite for honest problem-solving.

8

After the Illusion

We have mapped the collision. We have identified the three clocks and their divergent tempos. We have traced the geopolitical fault lines, the infrastructure gaps, the thermodynamic constraints, and the feedback loops that make the current trajectory unsustainable.

Now comes the harder question: what do we do?

The temptation is to conclude with either techno-optimism ("fusion will save us") or techno-pessimism ("we're doomed"). Both are evasions. The reality demands something more difficult: strategic clarity about a set of nested problems that admit no single solution.

8.1 The Efficiency Imperative

If we cannot have infinite energy, we must have radical efficiency. This is not a platitude. It is a restatement of the fundamental constraint in the form of a design principle.

The brain runs on 20 watts. A frontier AI model runs on 20 megawatts. The gap is six orders of magnitude. Even if we close it by only three orders of magnitude—a thousand-fold improvement—the energy constraints we face today would largely dissolve.

How do we get there?

Sparse computation. The human brain does not activate all 86 billion neurons for every thought. It activates a tiny fraction—the specific circuits relevant to the task at hand. Current AI models activate *every* parameter for *every* token. Mixture-of-Experts architectures are a first step toward sparsity, routing each input to only a subset of the model's parameters. The gains so far are modest (2–4x). The theoretical potential is enormous (100–1000x).

Neuromorphic hardware. Traditional processors—CPUs and GPUs—are designed for *precise arithmetic*. They compute exact floating-point results to 16 or 32 decimal places. The brain does not work this way. It operates on approximate, noisy, asynchronous signals. Neuromorphic chips—Intel's Loihi, IBM's NorthPole, various aca-

demic prototypes—attempt to replicate this approach in silicon, achieving 10–100x better energy efficiency for certain workloads. The challenge is software: we do not yet have programming paradigms that can fully exploit neuromorphic hardware.

Optical and photonic computing. Light travels faster and generates less heat than electricity. Optical computing, which performs matrix multiplications using beams of light rather than electronic circuits, offers the tantalising possibility of near-zero-energy inference. Several startups—Lightmatter, Luminous Computing—are pursuing this approach. The technology is immature but the physics is sound.

In-memory computing. The energy cost of moving data between memory and processor (the "von Neumann bottleneck") accounts

for a majority of total system energy consumption. In-memory computing architectures, which perform computation *within* the memory chips themselves, eliminate this data movement and can reduce energy consumption by 10–100x for inference workloads.

None of these approaches is ready for production at scale today. All of them are plausible within a decade. The question is whether the industry will invest in them with the same urgency it currently devotes to training ever-larger models.

8.2 The Re-materialisation of Tech

For thirty years, the technology industry defined itself by its immateriality. Software companies had no factories. Cloud companies had no inventory. The tech sector was "asset-

light"—a designation that Silicon Valley wore as a badge of honour, distinguishing itself from the heavy industries of the 20th century.

AI reverses this completely.

The most valuable AI companies of the next decade will own *physical assets*: data centres, power plants, chip fabrication facilities, cooling infrastructure, land. They will look less like software companies and more like utilities, or mining companies, or vertically integrated industrial conglomerates.

This is not a regression. It is a recognition that intelligence, however immaterial it appears, is ultimately *physical*. It is computed on physical machines, powered by physical energy, cooled by physical water, connected by physical cables. The abstraction layer that we called "the cloud" was always a thin veneer over an enor-

mous physical infrastructure. AI is stripping that veneer away.

The implications for the technology workforce are profound. The engineer of the AI era must understand not just algorithms but *systems*: power delivery, thermal management, semiconductor physics, grid interconnection, chemical engineering (for cooling), and civil engineering (for construction). The full-stack developer of 2030 reads Python and circuit diagrams.

8.3 The Three Futures

We can now sketch three plausible trajectories for the collision we have described. They are not mutually exclusive—elements of all three may coexist—but they represent distinct

equilibria toward which the system might converge.

Future 1: Energy abundance. Fusion energy, or a dramatic acceleration of fission deployment, removes the energy constraint. AI scales without limit. The three clocks re-synchronize, with energy finally matching the pace of intelligence. Moore's Illusion becomes retroactively prophetic: progress really was exponential all along; it just needed the right energy source.

This future is possible but not probable in the near term. Fusion remains 10–20 years from commercial viability (a timeline that has itself been stubbornly constant for decades). Fission deployment could be accelerated with sufficient political will, but "sufficient political will" is the most unreliable variable in any strategic plan.

Future 2: The plateau. Energy constraints cap AI development at a level substantially below its theoretical potential. Frontier models stabilise in size and capability. The industry pivots from scaling to *optimisation*—making existing models more efficient, more specialised, more accessible. The exponential dream ends, but civilisation adapts. AI becomes a powerful but bounded tool, like electricity itself: ubiquitous, essential, and taken for granted, but not infinite.

This future is moderate and plausible. It would disappoint the prophets of the singularity but relieve the catastrophists. It would produce enormous value—personalised education, accelerated drug discovery, automated infrastructure management—without the existential risks associated with unbounded superintelligence.

Future 3: Stratification. Energy constraints do not halt AI development globally, but they stratify it. A handful of nations and corporations with sufficient energy surplus and capital continue to push the frontier. Everyone else falls behind. AI access becomes a new axis of global inequality, as deterministic as access to clean water or electricity.

In this future, the digital divide is replaced by the *cognitive divide*: not who has a smartphone, but who has access to frontier intelligence. The geopolitical consequences are severe. Nations on the wrong side of the divide become permanent consumers of intelligence produced elsewhere, with all the dependency and vulnerability that implies.

This is the most likely near-term future, and the one that demands the most urgent strategic attention.

8.4 Redefinition

The deepest consequence of Moore's Illusion is not economic or geopolitical. It is *philosophical*.

For half a century, we defined progress as "more": more transistors, more compute, more data, more capability, more growth. This definition was sustained by the extraordinary anomaly of Moore's Law, which made "more" reliably available at ever-decreasing cost.

The post-Moore era demands a different definition: progress as "less." Less energy per inference. Less data per training run. Less hardware per capability. Less waste per outcome. Elegance, not brute force. Sufficiency, not excess.

This is not austerity. It is sophistication. The most advanced civilisations in human history were not the ones that consumed the most resources. They were the ones that extracted the most value from the fewest resources. The Roman aqueduct. The Japanese tea ceremony. The Swiss watch. Mastery is the art of doing more with less.

The AI industry has never practiced this art. It has always had access to Moore's abundant surplus, and it has always spent that surplus lavishly. The end of the surplus is not a tragedy. It is an invitation to engineering discipline that the field has never had to exercise.

The illusion is over. The real work begins.

References & Further Reading

Moore, G. E. (1965). "Cramming More Components onto Integrated Circuits." *Electronics*, 38(8).
The four-page paper that launched a fifty-year prophecy.

Kaplan, J. et al. (2020). "Scaling Laws for Neural Language Models." *arXiv:2001.08361*.
The empirical demonstration that Transformer performance scales as a smooth power law of compute.

Jevons, W. S. (1865). *The Coal Question*. Macmillan.

The original statement of the efficiency-consumption paradox that haunts the AI energy debate.

Smil, V. (2017). *Energy and Civilization: A History*. MIT Press.
The definitive account of how energy transitions actually unfold—slowly, unevenly, and painfully.

Miller, C. (2022). *Chip War: The Fight for the World's Most Critical Technology*. Scribner.
Essential history of the semiconductor industry and TSMC's rise to strategic dominance.

Sutton, R. (2019). "The Bitter Lesson." Personal blog.
The argument that compute, not cleverness, wins in AI—and its hidden assumption about supply.

Hoffmann, J. et al. (2022). "Training Compute-Optimal Large Language Models." *arXiv:2203.15556.*

The "Chinchilla" paper that proved most models were trained wrong—too large, too little data.

IEA (2024). *Electricity 2024: Analysis and Forecast to 2026*. International Energy Agency.
Data on global electricity consumption trends, including the accelerating demand from data centres.

Bostrom, N. (2014). *Superintelligence: Paths, Dangers, Strategies*. Oxford University Press.
The intelligence explosion hypothesis, viewed through the lens of energy constraints.

About the Author

Nabil El Mahyaoui

https://nabil-el-mahyaoui.com/

https://www.linkedin.com/in/elmahyaoui