ChatGPT

# The Physical Limits of Machine Intelligence (Through 2050)

## Introduction

Machine intelligence ultimately runs on physical hardware, so its capabilities and limits are constrained by fundamental math, physics, and thermodynamics. This report explores *first-principles* limits on AI computing power – focusing on hardware (energy, transistors, etc.) rather than algorithms – and compares them to the "gold standard" of human cognition. We review the historical progress of computational power (e.g. Moore's Law for transistor density and Koomey's Law for energy efficiency) and extrapolate trends toward 2050, examining whether any fundamental limits will impede AI. We also compare machine computation to the human brain's efficiency and consider thermodynamic bounds (like Landauer's principle) on computation. The goal is to answer: **from a physics and energy perspective, what (if any) limits exist to AI progress through 2050?**

## The Human Brain as a Baseline for Intelligence

The human brain provides a useful benchmark for general-purpose intelligence. It operates on about 20 watts of power – roughly the energy consumption of a dim light bulb [1] – yet enables remarkable cognitive capabilities. In computational terms, the brain's effective processing rate is hard to pin down, but estimates range widely. Recent analyses suggest that *on the order of* 10^15–10^17 operations per second might suffice to match the brain's functionality [2] (depending on what level of neural detail is simulated). For instance, one detailed report by Carlsmith (2020) finds it **more likely than not** that about 10^15 FLOPS (floating-point operations per second) is enough to perform tasks as well as a human brain, and very unlikely that more than 10^21 FLOPS would be required [2]. In other words, a computer performing ~1 quadrillion operations per second might reach human-level cognitive performance if organized effectively.

To appreciate the brain's *energy efficiency*, consider that performing ~10^15–10^17 operations on only 20 W means the brain achieves on the order of 5×10^13–5×10^15 operations per joule (i.e. FLOPS per watt). **That is vastly more efficient than today's computers.** A 2023 NIST article illustrates this vividly: *"The human brain is an amazingly energy-efficient device. In computing terms, it can perform the equivalent of an exaflop (10^18 operations per second) with just 20 watts of power"* [3]. By contrast, the cutting-edge **Frontier** supercomputer (Oak Ridge National Lab) recently achieved about an exaflop of performance but required ~20 *million* watts (20 MW) of power [3]. This means the brain was **on the order of a million times more energy-efficient** at that exaflop-scale task. Even allowing for uncertainty in brain FLOP estimates, biology clearly outclasses current silicon in joules per computation. Another comparison: the Fugaku supercomputer (~0.45 exaflops) draws around 30 MW, versus the brain's 20 W. In raw FLOP per second, today's largest machines *exceed* brain-scale (Frontier's peak ~10^18 FLOPS vs. brain ~10^15–10^17), but they consume **over a million-fold more power** to do so [3].

Why is the brain so efficient? One factor is its highly parallel, 3D architecture: ~10^11 neurons with ~10^14 synapses operate in parallel at slow speeds (on the order of tens of Hz), which avoids the high clock

frequencies and off-chip data movement that burn energy in modern chips. The brain also relies on low-power analog signaling (spikes of ~100 millivolts) and densely co-located memory and processing (synapses and neurons), whereas conventional computers expend much energy moving data between separate CPU and memory units. This suggests that **the efficiency gap is not a fundamental physics necessity**, but rather a result of current engineering paradigms. In principle, if we can design hardware more like brains (e.g. neuromorphic or analog in-memory computing), we could dramatically cut energy per operation [4] [5]. By 2050, advances in chip architecture may narrow this gap, but for now the brain remains the *benchmark* for efficient, general intelligence. We will return to how close future hardware might come to brain-like efficiency after examining current trends and limits.

## Historical Progress in Computing Power

Progress in machine intelligence has ridden on exponential improvements in semiconductor technology. **Moore's Law**, the historical observation that transistor counts on integrated circuits double roughly every 18–24 months, held for several decades and drove an exponential rise in computing performance. Although Moore's Law (in its original form) has slowed in recent years, today's chips still pack **tens of billions of transistors** in a single package, enabling petaflop-scale computations in devices as small as a graphics card. For example, NVIDIA's high-end GPUs (circa 2023) contain ~80 billion transistors and achieve on the order of 10^14 FLOPS (0.1 quadrillion) in specialized precision modes. The *cumulative* effect is that computing power available per dollar has increased by **many orders of magnitude** over the last half-century, enabling the AI breakthroughs we see today.

However, raw FLOP/s is only part of the story – energy efficiency has also improved exponentially, following a trend sometimes called **Koomey's Law** (the number of computations per joule doubles on a similar periodic cadence). Between 1946 and 2000, the energy efficiency of computing (FLOPS per joule) doubled about every 1.5 years; since 2000, the doubling slowed to ~2.7 years as CMOS scaling encountered new challenges. Even so, modern chips are *vastly* more energy-efficient than those of the past. One analysis notes that since the early days of computing, transistors have become about **a thousand times smaller, 100,000× faster, and a billion times more energy-efficient** [6]. This relentless improvement has meant that tasks once requiring room-sized supercomputers can now run on a smartphone.

A concrete measure of progress is the **Green500** (a list ranking supercomputers by energy efficiency). As of 2023, the top systems achieve around 60–70 gigaflops per watt (GFLOPS/W) on the Linpack benchmark. For example, the Flatiron Institute's *Henri* system led with ~65 GFLOPS/W in late 2023, and an experimental NVIDIA-Jülich system achieved ~72.7 GFLOPS/W. This is a remarkable gain considering that a decade prior, ~5 GFLOPS/W was a leading score. In other words, top supercomputers improved from ~5 to ~70 GFLOPS/W in about 10–12 years – roughly a 14× efficiency gain. Over a longer horizon (2008–2023), average HPC energy efficiency doubled roughly every 2.3 years. If sustained, that trend would yield another ~10×–20× efficiency increase by 2035 and perhaps ~100× by around 2050 (though extrapolation beyond known physics must be done cautiously, as we discuss next).

It's important to note that **Moore's Law is approaching physical limits**. Transistor feature sizes are now measured in just a few nanometers – only a handful of silicon atoms wide. Industry roadmaps show transistor gate lengths shrinking to ~1 nm and below by the late 2020s, entering the *"angstrom era"* [7]. In fact, researchers at IMEC (a leading semiconductor R&D center) have outlined a path to ~0.2 nm scale devices by around 2036 through novel transistor designs [8]. This involves moving from today's FinFET transistors to **Gate-All-Around (GAA)** nanosheet transistors at 2 nm, then to fork-sheet and ultimately

*nanowire or atomic-channel transistors* at scales of 5 Ångstrom (0.5 nm) or even 2 Å (0.2 nm) [8]. **Figure 1** below shows IMEC's projected roadmap through 2036, with successive new device architectures enabling continued scaling even below 1 nm.

*Figure 1: Imec's projected transistor roadmap through 2036, extending beyond the 1 nm node. The industry plans to transition from FinFET transistors (down to ~3 nm) to Gate-All-Around (GAA) nanosheet and forksheet devices at 2 nm and 0.7 nm, then to breakthrough designs like complementary FETs (CFETs) with atomic-scale (sub-0.5 nm) channels [8]. Such advances aim to continue Moore's Law-style scaling in transistor density.*

Despite these astonishing feats of engineering, **exponential scaling is slowing**. As transistors approach atomic scales, further reductions face quantum tunneling, heat dissipation, and fabrication limits [9]. Indeed, in many respects *"the physical limits to transistor scaling have been reached"* in standard silicon technology [9]. This doesn't mean progress halts, but it necessitates innovation in materials and architecture. Chipmakers are already moving to 3D integration (chip stacking and chiplet architectures) to keep increasing effective density. For example, TSMC's recent roadmap emphasizes advanced packaging that will allow **>1 trillion transistors in a single package by 2030** using multiple stacked chiplets [10]. In short, while individual transistors can't shrink much beyond 1 nm without exotic approaches, we can still *add more transistors* via larger die, multi-die modules, and 3D stacking – albeit with diminishing returns and higher complexity. By 2050, we might see specialized AI hardware using stacks of silicon (or even optical and carbon-based processors) to greatly multiply computing power within a given volume.

## Energy, Thermodynamics, and the Limits of Computation

From a first-principles physics perspective, the most fundamental limits on computation arise from **thermodynamics** (energy dissipation) and related constraints like the speed of light. Rolf Landauer's famous principle (1961) states that each irreversible bit operation (flipping or erasing a bit) must dissipate at least k·T·ln2 energy as heat, where *T* is temperature and k is Boltzmann's constant. At room temperature (≈300 K), this **Landauer limit** is about *2.8×10^-21 joules per bit operation*. In practical terms, if you had a computer that operated at the absolute thermal efficiency limit, 1 joule of energy could perform on the order of 3.5×10^20 bit flips (since 1 / 2.8e-21 ≈ 3.5e20). Even a 32-bit floating-point operation would require on the order of 32–64 bit flips, meaning the theoretical maximum is on the order of 10^19 FLOPS per joule (i.e. ~10^19 FLOPS on 1 watt for 1 second). **This is an astronomically high ceiling** – roughly 10^7 times more efficient than the human brain, and 10^8–10^9 times beyond our best current hardware in FLOPS/W.

In practice, current computing elements are still far from Landauer's limit. Modern CMOS operations dissipate orders of magnitude more energy than 10^-21 J per bit. For example, a contemporary 16-bit floating operation might use ~150 femtojoules (1.5×10^-13 J) in a well-optimized GPU, which is some 10^8 times the Landauer energy per bit. A DRAM memory access might consume ~5 pJ (5×10^-12 J) – even more costly. Overall, today's **best chips** achieve around 70 GFLOPS/W, meaning each 64-bit FLOP takes ~14×10^-12 J, or on the order of 10^-11 joules. Each such operation likely involves hundreds of bit-level transitions across logic and memory, so it's still vastly above the thermodynamic minimum. The gap *is* closing steadily as efficiency improves, but we remain many orders shy of the fundamental physics limits. One study of CMOS efficiency limits (Heim et al. 2023) estimated that an optimally engineered CMOS system might be about 200× more efficient than current state-of-the-art (H100 GPU) – in other words, there is room for perhaps two additional orders-of-magnitude improvement within the *current paradigm* before hitting a wall. Even that 200× gain would not reach the Landauer limit; it would just mean maybe ~10^4 GFLOPS/W instead of ~10^2 GFLOPS/W. Indeed, extrapolating recent trends, researchers suggest the Landauer limit

**won't be reached by irreversible computing until around 2090**. By 2050, however, we may begin to brush up against lesser (engineering) limits unless new techniques are adopted.

It's worth noting that **reversible computing** can evade Landauer's limit in principle, by avoiding bit erasures. Reversible logic (and quantum computing as a special case) can perform computations without the mandated kT ln2 heat loss per operation, theoretically allowing far lower energy per operation. The *Margolus-Levitin quantum limit* suggests a quantum system can perform operations at an energy cost of h/(4Δt), implying a lower bound of ~3×10^-34 J per op for a 1 Hz operation – vastly below the classical Landauer cost. However, practical reversible or quantum computing is very challenging and not general-purpose at large scale yet. By 2050, if reversible logic or adiabatic computing finds mainstream use, we might see incremental improvements in energy per operation, but it is unlikely to fully replace CMOS for general AI computing in that timeframe. **Quantum computers**, while powerful for certain tasks, do not obviously help with *general* AI workloads before 2050 (and introduce their own overhead and specialization). Thus, conventional thermodynamic limits will likely remain the key consideration for machine intelligence in the coming decades.

Another physical consideration is **communication and latency**. The speed of light imposes a limit on how fast information can travel across a computing system. As we pack more components into a chip or system, distributing signals and memory updates without incurring delay or synchronization issues becomes harder. A signal can travel ~30 cm in 1 ns (one clock cycle at 3.3 GHz) in a vacuum, and much less in copper or silicon. This means future exascale or greater systems might need to be physically larger (for cooling), but that introduces communication delays if the system is not carefully architected. Engineers mitigate this through multi-core and distributed processing (bringing computation closer to data), but it's a **constraint on effective, synchronized "intelligence"** – a large AI can't be a single monolithic, instantaneously communicating device beyond a certain size. By 2050, AI systems will likely consist of *modular, distributed hardware* (potentially millions of cores or neurons in a network), rather than one giant processor, to sidestep these speed-of-light and heat dissipation bottlenecks. This mirrors how the brain itself is organized (localized processing in parallel).

## Outlook Through 2050: How Far Can AI Go?

Barring unforeseen breakthroughs, it appears **no absolute physical barrier** will halt AI progress before 2050 – but progress will increasingly rely on engineering innovations to work around slowing silicon scaling. We can make some forecasts based on current trends and limits:

- **Computing Power Growth:** Even if single-core speeds plateau (they have been roughly stuck ~3–5 GHz for two decades due to power limits), aggregate AI computing will grow by using *more cores, more chips, and specialized accelerators*. By 2050, it's conceivable that we'll have routine access to **exascale** (10^18 OPS) or even *zetta*-scale (10^21 OPS) computing for AI, likely delivered by massive parallelism and 3D-integrated chips. For example, if transistor counts per chip continue to increase via 3D stacking and chiplet integration, we might see **trillion-transistor** chips by the 2030s [10], enabling single-package performance in the tens of petaflops or higher. Large AI datacenters in 2050 could contain thousands of such packages, reaching into the 10^21 FLOPS range (though software and memory bandwidth will need to keep up to utilize this).

- **Energy Efficiency:** Trends suggest a continued doubling of FLOPS/W every ~2–3 years for leading-edge systems. If this holds to 2050 (~25 years, roughly 10 doublings), that's a ~1000× efficiency

improvement over 2025 levels. A current 50 GFLOPS/W system could then be ~50,000 GFLOPS/W (50 TFLOPS/W). Notably, this is about the **same order as the brain's efficiency (~50–100 TFLOPS/W by some estimates)** [3]. In other words, by 2050, cutting-edge artificial hardware might rival the brain in ops per watt – a dramatic closing of the gap, if achieved. However, reaching 1000× efficiency gains may require moving beyond today's CMOS. Likely assistive technologies include **new materials** (e.g. carbon nanotube transistors or tunnel FETs with lower switching energy), **3D in-memory computing** (to cut data movement energy), and possibly **optical interconnects** (photons dissipate less energy over distance than electrons). Each of these could incrementally push us closer to Landauer's limit. It's also possible that by 2050 we will use **cryogenic cooling** for some AI hardware to reduce kT (making Landauer's limit lower and allowing more efficient reversible logic), though this comes with significant complexity.

- **Comparing to Human-Level Intelligence:** If ~10^15 FLOPS is indeed sufficient for human-like cognition [2], then by 2050 it is extremely likely that even pocket-sized devices will exceed that (given that even a 2023 GPU approaches 10^14 FLOPS). The frontier will have shifted to *how to harness* such hardware effectively. In terms of raw hardware, multiple "human-level" AI instances could run in parallel on a single small cluster by 2050. The limiting factor may be energy cost and efficiency. A human brain uses 20 W; to *match* that at 10^15 FLOPS in 2025 might take megawatts in a data center. By 2050, if hardware reaches ~50 TFLOPS/W, then 10^15 FLOPS would need only ~20 W – effectively on par with a brain in energy-use. Thus, **from a pure hardware standpoint, we see no fundamental barrier to machines equaling or exceeding human computational capacity** within the next few decades. The challenge will be managing the heat and energy when we aggregate millions of times human capability (for "super-intelligence" scale systems).

- **Ultimate Physical Limits:** Looking well beyond 2050, if one asks "How far can AI go" in theory, physics offers immense headroom. One rough bound, Bremermann's limit (derived from quantum physics and relativity), is ~10^50 operations per second per kilogram of matter [11]. A 1-kg "ultimate computer" running near that limit could perform ~10^50 ops/sec, an *astronomical* figure (for comparison, 10^50 ops is more than a trillion trillion times the operations performed by all computers on Earth today). The fact that such limits are so high means *physics does not fundamentally prevent far more advanced AI*. The constraints are practical: how to approach those limits with technology that doesn't melt down or become unreliable. Through 2050, we will still be many orders of magnitude below such extreme limits – likely closer to 10^18–10^21 ops/sec in the largest systems – so the ceiling will be set by engineering and economic factors rather than physics.

In summary, from a first-principles perspective **we do not see a hard wall stopping AI progress by 2050**. The laws of thermodynamics impose a minimum energy per computation, but current and near-future tech are far enough above that floor to allow continued improvements. Transistor scaling will reach atomic scales by ~2030, but creative engineering (3D integration, new transistor designs, better cooling) will extend computing growth at least into the 2030s, and likely new paradigms (optical, neuromorphic, etc.) will supplement CMOS afterward. The human brain demonstrates what is physically possible in terms of efficient, general computation at 20 W; by 2050 our machines should at least replicate that efficiency, and exceed the brain in raw speed by many orders.

## Conclusion

**Are there limits to AI, from a physics standpoint?** In the ultimate sense, there are limits – but they are extremely high, far beyond what even the year 2050 likely holds. The fundamental limits (Landauer's heat dissipation limit, light-speed communication, quantum limits) do constrain how efficient and large intelligence can grow, but our current technology hasn't yet bumped into those ceilings. Through 2050, the trajectory of AI will be determined by *engineering challenges*: Can we continue packing more computing power into chips once transistors hit a few atoms in size? Can we remove heat and supply power to massively parallel AI systems? These are serious challenges, but not insurmountable with focused R&D. Incremental advances in materials and design (e.g. new transistor types, 3D chip stacking, specialized accelerators, cooling techniques) are expected to carry us forward.

From a policy perspective, this means we should anticipate **dramatically more capable and efficient AI systems by 2050**. There may not be a single moment where "physics stops further AI progress"; instead, gains could gradually slow as we spend more effort for smaller improvements. But **no abrupt wall** is visible yet. If anything, human biological brains are proof that far greater efficiency is achievable – and our machines are steadily bridging that gap. By 2050, machines are likely to reach human-level cognition at human-level power budgets, and with many times human speed. Beyond that, the only known limits would be approached as we near the end of Moore's Law and start requiring fundamentally new computing paradigms.

In conclusion, *the limits of AI are primarily engineering and economic up to 2050, rather than absolute physical constraints*. As we approach 2050, we expect AI hardware to leverage essentially all available physical leeway – wringing out efficiency to within perhaps an order of magnitude of thermodynamic limits, and scaling up total compute closer to what physics comfortably allows (though still far from the *ultimate* limits). No known first-principles law forbids AI from matching or exceeding human brain performance; it is only a matter of time, innovation, and resources. In short, from a first-principles perspective, **the "limits" of machine intelligence through 2050 are loose and high – we have plenty of headroom, and the coming decades will likely see AI systems grow ever more powerful until practical considerations, not physics, dictate their limits**.

**Sources:**


- Carlsmith, J. (2020). *How Much Computational Power Does It Take to Match the Human Brain?* (Open Philanthropy report) [2].
- AI Impacts. *Brain Performance in FLOPS*.
- NIST (Madhavan, A., 2023). *Brain-Inspired Computing...* [3] [12].
- Green500 (Top500.org, 2023) – Energy efficiency of supercomputers.
- Heim, L. et al. (2023). *Limits to the Energy Efficiency of CMOS Microprocessors*.
- Moore's Law and semiconductor roadmaps (TSMC/Imec) [8] [10].
- Landauer's principle and limits (Cluster Computing, 2024).
- Additional references on AI hardware trends and comparisons.


---

[1] Learning from the brain to make AI more energy-efficient
https://www.humanbrainproject.eu/en/follow-hbp/news/2023/09/04/learning-brain-make-ai-more-energy-efficient/

2 How Much Computational Power Does It Take to Match the Human Brain? | Open Philanthropy
https://www.openphilanthropy.org/research/how-much-computational-power-does-it-take-to-match-the-human-brain/

3 4 5 6 12 Brain-Inspired Computing Can Help Us Create Faster, More Energy-Efficient Devices — If We Win the Race | NIST
https://www.nist.gov/blogs/taking-measure/brain-inspired-computing-can-help-us-create-faster-more-energy-efficient

7 8 Imec Presents Sub-1nm Process and Transistor Roadmap Until 2036: From Nanometers to the Angstrom Era | Tom's Hardware
https://www.tomshardware.com/news/imecs-sub-1nm-process-node-and-transistor-roadmap-until-2036-from-nanometers-to-the-angstrom-era

9 Moore's law - Wikipedia
https://en.wikipedia.org/wiki/Moore%27s_law

10 TSMC reaffirms path to 1-nm node by 2030 on track - EDN
https://www.edn.com/tsmc-reaffirms-path-to-1-nm-node-by-2030-on-track/

11 Bremermann's limit - Wikipedia
https://en.wikipedia.org/wiki/Bremermann%27s_limit