International Scientific Journal of Engineering and Management (ISJEM) ISSN: 2583-6129
Volume: 04 Issue: 11 | Nov – 2025 DOI: 10.55041/ISJEM05177
An International Scholarly || Multidisciplinary || Open Access || Indexing in all major Database & Metadata

Architectural Shift Necessitated by Voracious Energy Demands of Generative Artificial Intelligence (GAI) and High-Performance Computing

[1]M L Sharma, [2]Sunil Kumar, [3]Ajay Kumar Garg, [4]Mahir Pandey, [5]Parth Shukla

[1,2,3]Faculty, Maharaja Agrasen Institute of Technology, Delhi

[4,5]Research Scholar, Maharaja Agrasen Institute of Technology, Delhi
[1]madansharma.20@gmail.com, [2]sunilkumar@mait.ac.in, [3]ajaygargiitr@gmail.com,
[4]mahirpandey2024@gmail.com, [5]parthshuklaji69@gmail.com

Abstract

The computational substrate of the 21st century is undergoing a radical phase transition. The deterministic certainty that defined the era of Moore's Law—where performance gains were achieved through the reliable shrinking of transistors without a penalty in power density—has irrevocably collapsed. As the semiconductor industry confronts the breakdown of Dennard scaling and the physical limits of lithography, a new paradigm has emerged: Approximate Computing (AC). This architectural shift, necessitated by the voracious energy demands of generative artificial intelligence and high-performance computing, deliberately trades bit-level precision for gains in energy efficiency and throughput. However, this transition from exactitude to approximation is not merely a technical optimization; it is a profound reordering of the sociotechnical contract between human operators and machine agents.

This report, "Accountable Approximation," provides an exhaustive analysis of the implications of this shift. By synthesizing data from energy audits, hardware security research, legal theory, and the geometry of neural loss landscapes, we demonstrate that the introduction of stochastic error into the hardware layer possesses significant, yet largely unexamined, agency. We explore how quantization noise—the arithmetic distortion introduced by reducing numerical precision—interacts with the high-dimensional geometry of deep learning models to disproportionately erode the representation of minority data, effectively embedding bias into the silicon itself. Furthermore, we examine the security paradox where the "fog of error" sanctioned by approximation creates a camouflage for hardware Trojans, rendering traditional redundancy-based detection methods obsolete.

Synthesizing the latest findings from the International Energy Agency (IEA), Google's 2025 environmental reports, and cutting-edge research into "Fair-GPTQ" algorithms, this report argues that the sustainability of the AI revolution hinges on our ability to govern this new "technological unconscious." We propose a framework of Accountable Approximation that demands transparency in error budgets, rigorous auditing of the bias-variance trade-off in hardware, and a modernization of liability laws to address the non-deterministic nature of future computing systems. The era of the perfect machine is over; the era of the accountable machine must begin.

Keywords

Approximate computing, Thermodynamic computing, Hessian Spectrum Analysis, Large language models, Fair-GPTQ, Post-Moore's Law computing, Gradient Normal Disparity

1. Introduction: The End of the Deterministic Era

1.1 The Physical Limits of the Digital Age

For over fifty years, the global digital economy was underwritten by a predictable contract with physics, colloquially known as Moore's Law. This observation, made by Gordon Moore in 1965, posited that the number of transistors on a microchip would double approximately every two years. It was a prophecy of exponential growth that held true for decades, driving the revolution in personal computing, the internet, and mobile connectivity.[1] However, the engine beneath Moore's Law was not merely transistor density; it was a scaling behavior described by Robert Dennard in 1974. Dennard scaling stated that as transistors became smaller, their power density stayed constant, allowing engineers to increase clock speeds and performance without increasing the power budget of the chip.[3]

The physics of this era could be summarized by the proportionality of dynamic power consumption (P) to capacitance (C), voltage (V), and frequency (f):

$$P \approx CV^2 f$$

Under ideal scaling, if transistor dimensions were reduced by a factor S (where $S > 1$), capacitance C would reduce by $1/S$ and voltage V could be reduced by $1/S$ to maintain constant electric field. This allowed frequency f to increase by S while keeping the power density ($\frac{P}{\text{Area}}$) constant.[3]

That contract has been broken. Around 2006, Dennard scaling collapsed. As transistors shrank below 90 nanometers, leakage currents and thermal effects made it impossible to power all transistors on a chip simultaneously without exceeding the thermal design power—a phenomenon that gave rise to the era of "Dark Silicon." Today, we face the imminent flattening of Moore's Law itself, with consensus estimates suggesting the economic and physical viability of strictly lithographic scaling will cease to be the primary driver of performance by roughly 2025.[5] The breakdown is so severe that industry leaders like NVIDIA CEO Jensen Huang have declared Moore's Law "dead," prompting fierce rebuttals from counterparts at Intel, illustrating the deep anxiety pervading the hardware sector.[1]

This physical stalling occurs at the precise moment that humanity's demand for computation has entered a hyper-exponential phase. The rise of Large Language Models (LLMs) and generative AI has created a computational workload that doubles not every two years, but every few months. The training of a single state-of-the-art model requires energy expenditure comparable to the annual consumption of a small town, and the subsequent inference phase—the daily use of these models—threatens to consume terawatt-hours on a national scale.[7]

1.2 The Turn to Approximation

Faced with an immovable thermal ceiling and an unstoppable demand for intelligence, computer architects have been forced to abandon the ideal of exactness. The industry is pivoting toward Approximate Computing (AC). The core premise of AC is simple yet radical: many modern applications, particularly in AI, media processing, and data mining, are inherently error-resilient. A neural network does not need to know that a weight is exactly 0.123456789; it functions perfectly well—and often faster—if it assumes the weight is 0.12.

By relaxing the requirement for strict Boolean correctness, engineers can achieve massive gains in efficiency. Techniques such as voltage over-scaling (running chips at lower voltages than is safe for perfect accuracy), truncation (chopping off the least significant bits of a calculation), and aggressive quantization (representing numbers with 4 bits instead of 32) allow for the reclamation of performance lost to the death of Dennard scaling.[9] This shift is evident in the architecture of modern GPUs, such as NVIDIA's Blackwell, which achieves its generational leaps in efficiency largely through the support of lower-precision number formats designed specifically for AI inference.[11]

1.3 The Thesis of Accountable Approximation

However, the transition to approximate computing is not a value-neutral engineering decision. It introduces a probabilistic element into the heart of the digital stack. When a processor is allowed to "guess," the errors it produces are not random, benign noise. They are structured artifacts of the hardware design that interact with software and data in complex, non-linear ways.

This report posits that approximation grants hardware a form of "material agency".[12] The silicon itself begins to make decisions about what information is preserved and what is discarded. If these decisions are left unexamined, they can amplify social biases, create unmonitorable security vulnerabilities, and create a "liability gap" where neither the software developer nor the hardware manufacturer can be held responsible for system failures.

"Accountable Approximation" is a proposed framework for navigating this new reality. It argues that error-tolerant systems must be designed with the same rigor applied to safety-critical systems. We must understand the geometry of the errors we introduce, ensuring they do not disproportionately impact vulnerable populations. We must secure the "fog of error" to prevent malicious actors from hiding within it. And we must update our legal frameworks to assign responsibility in a world where computation is no longer guaranteed to be correct.

---

2. The Thermodynamic Floor: Energy, Inference, and the Cost of Precision

2.1 The Escalating Energy Appetite of AI

To understand the inevitability of approximation, one must first confront the thermodynamic reality of modern AI. The energy consumption of the information technology sector is no longer a rounding error in global electricity usage; it is becoming a primary driver of grid demand. The International Energy Agency (IEA) projects that electricity demand from data centers, driven largely by AI, could double between 2022 and 2026, growing from 460 terawatt-hours (TWh) to over 1,000 TWh—roughly equivalent to the entire electricity consumption of Japan.[14]

This growth is bifurcated into two distinct phases: training and inference. While the public imagination focuses on the massive energy cost of *training* a model—a discrete event that is undeniably energy-intensive—it is the *inference* phase that poses the long-term sustainability challenge.

2.1.1 The Training Burden

The training of a foundation model is an industrial-scale energy event. Estimates for the training of GPT-3 place its consumption around 1,287 MWh, emitting over 550 metric tons of carbon dioxide equivalent (CO_2e).[8] To contextualize this, 1,287 MWh is enough energy to power approximately 120 average US households for a full year.[17]

However, GPT-3 is now considered a legacy model. Its successor, GPT-4, is estimated to have consumed between 51,773 and 62,319 MWh during training—a staggering 40-fold increase.[18] This exponential rise in training costs follows a trend where the compute required for cutting-edge AI doubles every 3.4 months, far outstripping the historical 2-year doubling time of Moore's Law.[19] This divergence between the demand for compute and the efficiency of the underlying hardware creates an unsustainable trajectory that can only be flattened by radically improving the efficiency of the computation itself—hence, the move to approximation.

2.1.2 The Inference Tsunami

While training is intense, it happens once. Inference happens billions of times a day. Inference refers to the process of the model generating a response to a user prompt. Every time a user asks ChatGPT a question, generates an image with Midjourney, or receives a recommendation from a streaming algorithm, inference is occurring.

Recent data suggests that inference already accounts for the majority of AI's lifecycle energy footprint, potentially up to 90%.[8] The energy cost per query varies wildly depending on the modality. A simple text-based query might consume around 0.047 kWh, while generating a single image can consume nearly 3 kWh—roughly the same amount of energy as fully charging a smartphone.[20] When aggregated across hundreds of millions of daily active users, the energy demand becomes colossal. If every Google search performed today were transitioned to a generative AI interaction, the energy consumption would rival that of the entire nation of Ireland.[19]

Google's 2025 environmental reporting provides a granular look at this challenge. They report that the median energy consumption for a "Gemini Apps" text prompt is approximately 0.24 Watt-hours (Wh). Google frames this optimistically, equating it to "watching television for less than nine seconds".[21] While this individual unit cost seems low, the aggregate scale is the defining factor. Trillions of "nine-second TV spots" amount to a massive continuous load on the grid. Furthermore, these figures often obscure the "embodied carbon"—the energy used to manufacture the chips themselves—which is a significant portion of the total footprint.[23]

2.2 Water: The Hidden Resource

The thermodynamic cost of precision is not paid in electricity alone; it is also paid in water. Data centers generate immense heat, and removing that heat requires industrial-scale cooling systems that often rely on evaporative cooling. The training of GPT-3 is estimated to have consumed over 700,000 liters of clean freshwater for cooling, enough to fill two-thirds of an Olympic swimming pool.[8]

As chip density increases to compensate for the slowing of Moore's Law, the heat density of the hardware rises, necessitating even more aggressive cooling. Google reported replenishing 4.5 billion gallons of water in 2024 to offset this consumption, aiming for "water positivity," but the local impact on drought-stricken regions where data centers are often located remains a critical sociotechnical tension.[24] Approximate computing offers a direct remediation here: by reducing the precision of calculations, the switching activity of the transistors decreases, generating less heat and directly reducing the water intensity of the compute.[7]

2.3 The Role of Approximation in Sustainability

The industry's response to this energy and water crisis has been a decisive move toward specialized, approximate hardware. The efficiency gains reported by major players are inextricably linked to this shift. NVIDIA's Blackwell platform, for instance, claims a 25x improvement in energy efficiency for LLM inference compared to previous generations.[11] This leap is not due to a magical breakthrough in transistor physics; it is largely due to the adoption of 4-bit floating-point arithmetic (FP4) and other reduced-precision formats.

Similarly, Google cites a 30x improvement in the power efficiency of its Tensor Processing Units (TPUs).[24] These gains are achieved by stripping away the "unnecessary" precision of 32-bit or 64-bit computing. The reasoning is that the statistical noise of a neural network allows it to absorb the errors of lower precision without breaking. However, as we will explore in subsequent sections, this "absorption" of error is not uniform, and the efficiency gained here is purchased with a currency of transparency and fairness.

International Scientific Journal of Engineering and Management (ISJEM) ISSN: 2583-6129
Volume: 04 Issue: 11 | Nov – 2025 DOI: 10.55041/ISJEM05177
An International Scholarly || Multidisciplinary || Open Access || Indexing in all major Database & Metadata

3. The Geometry of Loss: Sharp Minima and the Physics of Quantization

3.1 The Mathematical Definition of Quantization

To understand why approximation is risky, we must look at the mathematics of quantization. In digital signal processing, quantization is the mapping of a continuous range of values to a finite set of discrete levels.[25] When applied to a neural network, it involves taking the "weights"—the billions of parameters learned during training, usually represented as high-precision floating-point numbers—and snapping them to the nearest value on a coarse grid.

For a uniform quantization scheme with step size Δ, a real-valued weight x is mapped to a quantized value x_q:

$$x_q = \Delta \cdot \text{round}\left(\frac{x}{\Delta}\right)$$

This process introduces "quantization noise" (ϵ), defined as the difference between the original signal and the quantized representation: $\epsilon = x - x_q$. In audio, this noise sounds like a hiss or distortion overlying the music.[26] In a neural network, this noise distorts the model's "understanding" of the world.

3.2 Loss Landscapes and Hessian Spectra

The impact of this noise depends entirely on the geometry of the model's "loss landscape." The loss landscape is a visualization of how the model's error (loss, L) changes as its parameters ($\mathbf{w}$) change. It is a high-dimensional terrain of hills (high error) and valleys (low error). The goal of training is to find the deepest valley—the global minimum.[27]

We can approximate the loss function around a minimum $\mathbf{w}^*$ using a second-order Taylor expansion:

$$L(\mathbf{w}^* + \mathbf{d}) \approx L(\mathbf{w}^*) + \nabla L(\mathbf{w}^*)^T \mathbf{d} + \frac{1}{2}\mathbf{d}^T \mathbf{H} \mathbf{d}$$

Here, $\mathbf{d}$ represents the perturbation vector caused by quantization noise. Since $\mathbf{w}^*$ is a minimum, the gradient $\nabla L(\mathbf{w}^*)$ is zero. The increase in loss is therefore determined primarily by $\mathbf{H}$, the Hessian matrix of second-order derivatives ($\mathbf{H}_{ij} = \frac{\partial^2 L}{\partial w_i \partial w_j}$).[57]

- **Flat Minima:** Some valleys are wide and flat. In these regions, the eigenvalues (λ) of the Hessian $\mathbf{H}$ are small. If the model's parameters are in a flat minimum, pushing them slightly to the left or right (as quantization does) results in a negligible increase in loss ($\mathbf{d}^T \mathbf{H} \mathbf{d} \approx 0$). These models are robust.[29]
- **Sharp Minima:** Other valleys are narrow and steep—like a ravine. These regions are characterized by large maximum eigenvalues ($\lambda_{max} \gg 0$) in the Hessian spectrum. If the model is balanced on the razor's edge of a sharp minimum, the term $\frac{1}{2}\mathbf{d}^T \mathbf{H} \mathbf{d}$ becomes large, sending the error skyrocketing. These models are brittle.[31]

3.3 The Volume Hypothesis and Generalization

This geometric understanding links approximation to the fundamental theory of learning. The "Volume Hypothesis" suggests that flat minima occupy a larger volume in the parameter space and therefore represent solutions that generalize better to new data.[27] Conversely, sharp minima often represent "overfitted" solutions that have memorized the training data but fail on the test data.

Crucially, quantization acts as a filter. It is physically impossible for a quantized model to reside in a sharp minimum because the "grid" of allowable values is too coarse to resolve the bottom of a narrow ravine. Therefore, approximate computing forces models to find flat minima. While this can theoretically improve generalization (by acting as a regularizer), it introduces a dangerous instability during the conversion process. If a model trained in high precision settles into a sharp minimum, attempting to quantize it for efficient inference will result in catastrophic accuracy degradation unless complex "fine-tuning" or "Hessian-aware" techniques are employed.[29]

This is where the "accountability" gap opens. If a deployer uses a cheap, "post-training quantization" method that ignores the Hessian spectrum, they may degrade the model's performance in subtle, non-uniform ways. The model might still work for the "average" query (the flat part of the manifold) but fail catastrophically for "edge cases" (the sharp parts), which often correspond to minority data or complex reasoning tasks.

---

4. The Mechanics of Approximation: How Hardware Guesses

4.1 Beyond Software: Hardware-Level Approximation

While quantization is often handled in software, true approximate computing modifies the hardware circuits themselves. To squeeze the last drops of efficiency out of the post-Dennard silicon, engineers are redesigning the fundamental logic gates that perform arithmetic.

One common technique is the use of **Approximate Adders**. A standard adder (like a Ripple Carry Adder) calculates the sum of two numbers and propagates the "carry" bit all the way from the least significant bit to the most significant bit. This propagation takes time and energy. An approximate adder, such as the Lower-Part-OR Adder (LOA), simply ignores the carry propagation for the lower bits, using a faster, cheaper OR gate instead.[33]

For an approximate adder, the output sum S_{approx} deviates from the exact sum S_{exact} by an error distance (ED):

$$ED = |S_{approx} - S_{exact}|$$

To evaluate the quality of these circuits, designers rely on the Mean Error Distance (MED) across all inputs N:

$$MED = \frac{1}{N} \sum_{i=1}^{N} |S_{approx}^{(i)} - S_{exact}^{(i)}|$$

This means that for certain input combinations, the hardware will literally calculate the wrong answer. $5 - 5$ might equal 10, but $5,000,005 + 5$ might equal $5,000,009$ due to a precision drop in the lower bits. This is not a bug; it is a design feature intended to save energy.

4.2 Voltage Over-Scaling and Timing Errors

Another technique is **Voltage Over-Scaling (VOS)**. Digital circuits require a certain voltage to switch their transistors fast enough to meet the system clock. If you lower the voltage, you save quadratic amounts of energy ($Power \propto Voltage^2$), but the transistors switch slower. Eventually, they switch too slowly to finish the calculation before the clock cycles, resulting in a "timing error".[34]

In an exact system, a timing error is a fatal crash. In an approximate system, it is treated as noise. The system is designed to accept that some percentage of operations will fail to complete, effectively truncating the calculation. This turns the processor into a stochastic machine: the output depends not just on the inputs, but on the physical temperature of the chip,

the minute variations in voltage from the power supply, and the specific "path delay" of the numbers being added.

4.3 Quantization Formats: The Battle for Bits

To manage this chaos, the industry has developed specialized data formats. We are moving away from the standard 32-bit Floating Point (FP32) toward formats like **BF16** (Brain Float 16), **FP8**, and even **INT4** (4-bit Integer).

Newer techniques like **AWQ (Activation-aware Weight Quantization)** and **GGUF** focus on identifying the "salient" weights—the 1% of parameters that are most critical for the model's accuracy—and keeping them in high precision, while crushing the rest of the model down to low precision.[35] This creates a "mixed-precision" architecture where the hardware dynamically adjusts its exactitude based on the importance of the data it is processing.

However, who decides what is "salient"? As we will see in the next section, the algorithms that determine which bits to keep and which to discard are often blind to sociological concepts of importance, leading to the emergence of "algorithmic bias in hardware."

---

5. Algorithmic Bias in Hardware: The Sociotechnical Agency of Silicon

5.1 The Myth of Neutral Compression

There is a pervasive assumption in engineering that compression is content-neutral—that shrinking a file or a model removes "redundancy" without altering "meaning." In the context of AI quantization, this assumption is demonstrably false. Recent research has revealed that quantization does not degrade model performance uniformly; it disproportionately impacts the model's ability to process information related to underrepresented groups.

This phenomenon stems from the statistical nature of quantization methods like **GPTQ (Generative Pre-trained Transformer Quantization)**. These algorithms effectively optimize for the "average" case. They try to minimize the error across the entire dataset. Since the dataset is dominated by majority representations (e.g., Western cultural norms, English language syntax, white male faces), the quantization algorithm prioritizes preserving the weights that encode these majority features. The weights that encode "outlier" features—often corresponding to minority groups or rare linguistic patterns—are deemed statistically less significant and are the first to be "rounded away".[36]

5.2 Case Study: Fair-GPTQ and the quantified Bias

The "Fair-GPTQ" study provides empirical evidence of this "hardware gentrification." Researchers found that when standard quantization was applied to Large Language Models (reducing them to 4-bit or 2-bit precision), the "perplexity" (a measure of confusion) on minority dialects increased significantly more than on standard English. Furthermore, bias metrics regarding gender, race, and religion worsened. The model became more stereotypical because the nuanced, high-dimensional representations required to understand context and avoid stereotypes were flattened by the quantization grid.[36]

To address this, the researchers introduced **Fair-GPTQ**, a modification to the quantization algorithm that includes a fairness-aware regularization term. Standard GPTQ minimizes the squared error of the weights. Fair-GPTQ modifies this

objective function to explicitly penalize bias:

$$\mathbf{W}c = \arg\min \mathbf{W}' \left(||\mathbf{W}\mathbf{X} - \mathbf{W}'\mathbf{X}||_2^2 + \alpha ||\mathbf{W}'(\mathbf{X}_0 - \mathbf{X}_1)||_2^2 \right)$$

Here, $\mathbf{W}$ is the full-precision weight matrix, and $\mathbf{W}'$ is the quantized matrix. The terms $\mathbf{X}_0$ and $\mathbf{X}_1$ represent stereotyped and anti-stereotyped inputs, respectively. The parameter α controls the penalty for bias. This equation forces the algorithm to find a quantized weight configuration $\#\mathbf{W}'$ that not only preserves accuracy (the first term) but also minimizes the difference in how the model treats different demographic groups (the second term).[37]

This creates a profound realization: we can no longer separate the hardware efficiency from the social outcome. The decision to use a standard quantization method versus a fairness-aware one is an ethical choice. If a deployer chooses the standard method to save 1% more energy or memory, they are actively choosing to degrade the experience for minority users. The "material agency" of the chip is thus realized—the hardware configuration itself dictates the fairness of the automated decision.[12]

5.3 The Technological Unconscious

This leads to the concept of the **Technological Unconscious**. Just as the human unconscious processes information below the level of awareness, the hardware layers of an AI system process data below the level of the software "consciousness." The "technological unconscious" of an approximate chip is "lossy." It is constantly forgetting information to save energy.

If this forgetting is not audited, the hardware becomes a silent oppressor. A medical diagnostic AI might perform brilliantly in the lab (on high-precision hardware) but fail in a rural clinic where it is deployed on a low-power, heavily quantized mobile chip. The failure is not in the software code, but in the *translation* of that code into the "technological unconscious" of the approximate hardware.[38]

---

6. The Security Frontier: Hardware Trojans and the Fog of Error

6.1 The Vulnerability of Imprecision

Security in computing has traditionally relied on the concept of "golden reference." You compare the output of your chip to the known correct output. If they differ, you have a problem—either a defect or a hack.

In approximate computing, there is no golden reference. The output is *expected* to be wrong occasionally. This creates a massive vulnerability. It generates a "fog of error" in which malicious actors can operate with impunity. This is the domain of **Hardware Trojans in Approximate Circuits**.[34]

6.2 The Mechanism of the Attack

A Hardware Trojan is a malicious modification to a circuit, inserted by an untrusted foundry or a rogue designer. In an exact circuit, Trojans are hard to hide because their activation usually causes a noticeable error. In an approximate circuit, the adversary can design the Trojan to affect only the Least Significant Bits (LSBs)—the bits that are already noisy.

For example, an adversary could insert a Trojan into an approximate adder used in a neural network. The Trojan is designed to trigger only when a specific "key" pattern appears on the input bus. When triggered, it flips a bit in the output. Because the application (the neural net) is designed to tolerate LSB errors, this flip is indistinguishable from normal quantization noise to the system's error-checking logic. However, to the adversary, that flipped bit could be a leaked cryptographic key or a "backdoor" that forces the neural network to misclassify a specific input (e.g., ignoring a stop sign).[33]

The attacker specifically targets "rare nodes" for the trigger to avoid detection during standard testing. The probability π of a signal line switching can be modeled, and if it falls below a threshold (γ_{th}), it is a candidate for a Trojan trigger:

$$\pi(1-\pi) < \gamma_{th}$$

This mathematical obscurity ensures the Trojan remains dormant during most tests, only activating under precise, malicious conditions.

6.3 The Failure of Traditional Defenses

Standard defense mechanisms like **Triple Modular Redundancy (TMR)**—running the calculation three times and voting on the result—fail completely in this context.

1. **Common Mode Failure:** If the approximation logic is deterministic (e.g., an LOA adder), all three redundant copies will produce the same "wrong" answer, validating the error.
2. **Statistical Divergence:** If the approximation is non-deterministic (e.g., voltage over-scaling), the three copies might legitimately produce three different answers. A voting system cannot distinguish between three honest approximate answers and one malicious Trojan answer.[40]

Research shows that **Majority Voting (MV)** techniques are invalid when trusted chips are assembled with approximate components from multiple vendors.[40] The inherent diversity of error profiles makes it impossible to establish a baseline of trust.

6.4 The Need for Statistical Security

Securing approximate hardware requires a paradigm shift from "logical security" to "statistical security." We cannot check for correctness; we must check for *distributional anomalies*. Security monitors must be embedded on the chip to track the statistical properties of the errors (mean, variance, skewness) and raise an alarm if the error distribution shifts in a way that suggests a Trojan activation.[34]

To accurately model these expected error distributions, simple Gaussian models often fail because approximation errors can be multimodal (having multiple peaks). Advanced **Gaussian Mixture Models (GMM)** are required to characterize the "fingerprint" of the approximation noise:

$$p(e) = \sum_{k=1}^{K} \phi_k \mathcal{N}(e|\mu_k, \Sigma_k)$$

By continuously monitoring whether the observed errors fit this complex GMM distribution, the system can potentially detect the statistical anomaly introduced by a Trojan.[52] However, this introduces a cruel irony: the overhead of these statistical monitors consumes energy, eating into the very efficiency gains that motivated the approximation in the first place. We are thus left with a trade-off: we can have ultra-efficient approximate computing, or we can have secure computing, but combining them requires complex, expensive, and currently immature technologies.

7. Legal and Liability Architectures for the Probabilistic Age

7.1 The Liability Gap

The introduction of probabilistic hardware fundamentally breaks existing product liability frameworks. Current law, such as the EU Product Liability Directive, is predicated on the notion of "defect." A product is defective if it does not provide the safety that a person is entitled to expect.

But what is the "expected safety" of a chip designed to be 98% accurate? If a self-driving car crashes because its approximate perception chip missed an object, is the chip "defective"? The manufacturer will argue that the chip performed exactly as specified within its error budget. They will claim the failure was a statistical inevitability, akin to a "force majeure," rather than a design flaw.[42]

This creates a **Liability Gap**. The software developer blames the hardware for the error. The hardware manufacturer blames the software developer for not making their model robust enough to handle the noise. The victim is left with no recourse.

7.2 B2B vs. Consumer Protections

The legal landscape is further complicated by the distinction between Business-to-Business (B2B) and Business-to-Consumer (B2C) contexts. In B2B contracts, liability is often limited by extensive warranty disclaimers. A cloud provider (like AWS or Google Cloud) selling "Spot Instances" or "Approximate Compute Instances" will likely include clauses stating that the user accepts the risk of calculation errors in exchange for the lower price.

However, when that computation affects a consumer—for example, a bank using that cloud instance to deny a loan—consumer protection laws (like the UK Consumer Protection Act) may kick in.[44] The consumer cannot sign away their right to fair treatment. If the approximation caused a discriminatory outcome (via the bias mechanisms discussed in Section 5), the company deploying the AI could be liable, even if they were unaware of the hardware-level distortion.

7.3 The "Service vs. Product" Debate

A critical legal ambiguity is whether AI and its underlying compute constitute a "product" or a "service." If AI is a service, strict liability rules often do not apply; the claimant must prove negligence (i.e., that the provider failed to take reasonable care). Proving negligence in the design of a probabilistic approximate circuit is incredibly difficult for a plaintiff who lacks access to the proprietary "error budget" documents of the chip manufacturer.[42]

7.4 Proposed Legal Standards: The "Rebuttable Presumption"

To address this, legal scholars and the European Commission are proposing a **"Rebuttable Presumption of Defectiveness"** for complex AI systems. This would shift the burden of proof. Instead of the victim having to prove the chip was defective, the manufacturer would have to prove that their approximation techniques *did not* cause the accident.[42]

Under an "Accountable Approximation" framework, this would require manufacturers to maintain immutable logs of the "quantization noise levels" during critical decisions—a "black box" for the chip itself. If they cannot produce these logs to prove the hardware was operating within safe statistical bounds, they are presumed liable.

8. Future Trajectories and Implications

8.1 Carbon Lock-in and the Infrastructure of AI

As we build out the infrastructure for the AI age—data centers, cooling systems, dedicated power plants—we risk creating a **Carbon Lock-in**. By optimizing our entire digital ecosystem around the massive energy consumption of generative AI, we entrench these technologies. We are pouring concrete and silicon for a future that *requires* high-energy compute.[46]

Approximate computing is a double-edged sword here. On one hand, it lowers the carbon intensity of each operation. On the other hand, Jevons Paradox suggests that as compute becomes cheaper and more efficient, we will simply consume more of it. The efficiency gains from approximation might simply fuel larger models and more ubiquitous inference, leading to a net increase in total environmental impact.[47]

8.2 Neurotechnology and the Ultimate Approximation

The principles of accountable approximation will soon extend beyond the data center to the human body. The rise of Brain-Computer Interfaces (BCIs) and neurotechnology relies on implantable chips that must operate under extreme thermal and power constraints (you cannot heat up brain tissue). These chips will *necessarily* rely on heavy approximation and compressive sensing.[48]

If we do not solve the problems of bias and security in approximate hardware now, we will be implanting these vulnerabilities directly into the human nervous system. A "quantization error" in a BCI could mean a loss of agency over one's own limbs, or a misinterpretation of neural intent. The "material agency" of the chip becomes indistinguishable from the biological agency of the human.

8.3 Conclusion: The Imperative of Auditability

The era of the "black box" must end. As we transition to a probabilistic computing paradigm, we must adopt a new standard of **Accountable Approximation**. This requires:

1. **Transparency:** Error budgets and quantization policies must be public and auditable.
2. **Fairness:** Hardware must be tested for demographic bias, not just signal-to-noise ratio.
3. **Forensics:** Systems must be able to reconstruct the "state of the error" after a failure.

Only by illuminating the "technological unconscious" of our hardware can we ensure that the post-Moore's Law world is sustainable, secure, and just.

References

- **Energy & Moore's Law:**.[1]
- **Quantization & Geometry:**.[25]
- **Hardware Mechanics & Security:**.[11]
- **Ethics, Law & Society:**.[12]

Works cited

1. Moore's law - Wikipedia, accessed November 16, 2025, https://en.wikipedia.org/wiki/Moore%27s_law
2. The Death of Moore's Law: What it means and what might fill the gap going forward, accessed November 16, 2025, https://cap.csail.mit.edu/death-moores-law-what-it-means-and-what-might-fill-gap-going-forward
3. Dennard scaling - Wikipedia, accessed November 16, 2025, https://en.wikipedia.org/wiki/Dennard_scaling
4. The Ripple Effects of Robert H. Dennard, Inventor of DRAM and Dennard Scaling - News, accessed November 16, 2025, https://www.allaboutcircuits.com/news/robert-h-dennard-inventor-dram-dennard-scaling/
5. The future of computing beyond Moore's Law - OSTI.GOV, accessed November 16, 2025, https://www.osti.gov/servlets/purl/1619164
6. The future of computing beyond Moore's Law | Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences, accessed November 16, 2025, https://royalsocietypublishing.org/doi/10.1098/rsta.2019.0061
7. Energy and AI - NET, accessed November 16, 2025, https://iea.blob.core.windows.net/assets/601eaec9-ba91-4623-819b-4ded331ec9e8/EnergyandAI.pdf
8. How Hungry is AI? Benchmarking Energy, Water, and Carbon Footprint of LLM Inference, accessed November 16, 2025, https://arxiv.org/html/2505.09598v1
9. Transistors on the Edge: The Quest for Energy-Efficient Computing, accessed November 16, 2025, https://harvardtechnologyreview.com/2024/04/22/transistors-on-the-edge-the-quest-for-energy-efficient-computing/
10. 'Approximate computing' improves efficiency, saves energy - Purdue University, accessed November 16, 2025, https://www.purdue.edu/newsroom/releases/2013/Q4/approximate-computing-improves-efficiency,-saves-energy.html
11. NVIDIA Sustainability Report Fiscal Year 2025, accessed November 16, 2025, https://images.nvidia.com/aem-dam/Solutions/documents/NVIDIA-Sustainability-Report-Fiscal-Year-2025.pdf
12. Ethics of artificial intelligence - Wikipedia, accessed November 16, 2025, https://en.wikipedia.org/wiki/Ethics_of_artificial_intelligence
13. Full article: On the technological unconscious: thinking the (a)signifying production of subjects and bodies with sonographic imaging, accessed November 16, 2025, https://www.tandfonline.com/doi/full/10.1080/14649365.2022.2083665
14. Explained: Generative AI's environmental impact | MIT News, accessed November 16, 2025, https://news.mit.edu/2025/explained-generative-ai-environmental-impact-0117
15. AI for a Greener Future: Its Power is in Our Hands | NVIDIA Technical Blog, accessed November 16, 2025, https://developer.nvidia.com/blog/ai-for-a-greener-future-its-power-is-in-our-hands/
16. How Hungry is AI? Benchmarking Energy, Water, and Carbon Footprint of LLM Inference, accessed November 16, 2025, https://arxiv.org/html/2505.09598v2
17. GPT-4 & Llama Carbon Footprint Analysis - AI Energy Consumption Calculator, accessed November 16, 2025, https://aienergycalculator.com/ai-energy-consumption-calculator-gpt-llama/
18. Artificial Intelligence, Real Consequences: Confronting AI's Growing Energy Appetite, accessed November 16, 2025, https://www.extremenetworks.com/resources/blogs/confronting-ai-growing-energy-appetite-part-1
19. How AI Uses Energy - Third Way, accessed November 16, 2025, https://www.thirdway.org/memo/how-ai-uses-energy
20. The Energy Impact of LLMs — Where Are We in 2025? | by Nathan Bailey - Medium, accessed

November 16, 2025, https://nathanbaileyw.medium.com/the-carbon-impact-of-llms-where-are-we-in-2025-6b6551e193ac
21. Our approach to energy innovation and AI's environmental footprint - Google Blog, accessed November 16, 2025, https://blog.google/outreach-initiatives/sustainability/google-ai-energy-efficiency/
22. Measuring the environmental impact of delivering AI at Google Scale, accessed November 16, 2025, https://services.google.com/fh/files/misc/measuring_the_environmental_impact_of_delivering_ai_at_google_scale.pdf
23. We finally know officially how much energy and water a ChatGPT query uses - Reddit, accessed November 16, 2025, https://www.reddit.com/r/energy/comments/1l9372g/we_finally_know_officially_how_much_energy_and/
24. 2025 Environmental Report - Google Sustainability, accessed November 16, 2025, https://sustainability.google/google-2025-environmental-report/
25. Quantization (signal processing) - Wikipedia, accessed November 16, 2025, https://en.wikipedia.org/wiki/Quantization_(signal_processing)
26. How does Quantization Noise Sound? - DSPIllustrations.com, accessed November 16, 2025, https://dspillustrations.com/pages/posts/misc/how-does-quantization-noise-sound.html
27. Sharp Minima Can Generalize: A Loss Landscape Perspective On Data - arXiv, accessed November 16, 2025, https://arxiv.org/html/2511.04808v1
28. The Generalization Mystery: Sharp vs Flat Minima - inFERENCe, accessed November 16, 2025, https://www.inference.vc/sharp-vs-flat-minima-are-still-a-mystery-to-me/
29. HAWQ: Hessian AWare Quantization of Neural Networks With Mixed-Precision - University of California, Berkeley, accessed November 16, 2025, https://www.stat.berkeley.edu/~mmahoney/pubs/HAWQ_ICCV_2019_paper.pdf
30. Thoughts on Loss Landscapes and why Deep Learning works - LessWrong, accessed November 16, 2025, https://www.lesswrong.com/posts/szXa8QgxjMypabJgN/thoughts-on-loss-landscapes-and-why-deep-learning-works
31. Do Sharpness-Based Optimizers Improve Generalization in Medical Image Analysis? - NIH, accessed November 16, 2025, https://pmc.ncbi.nlm.nih.gov/articles/PMC12121992/
32. Brief Review — Visualizing the Loss Landscape of Neural Nets | by Sik-Ho Tsang | Medium, accessed November 16, 2025, https://sh-tsang.medium.com/brief-review-visualizing-the-loss-landscape-of-neural-nets-dd93cb261afc
33. Design and analysis of hardware Trojans in approximate circuits - ResearchGate, accessed November 16, 2025, https://www.researchgate.net/publication/357075126_Design_and_analysis_of_hardware_Trojans_in_approximate_circuits
34. A Study of Trojan Attacks on Approximate Adders and Multipliers - IEEE Xplore, accessed November 16, 2025, https://ieeexplore.ieee.org/document/10483477/
35. A Comprehensive Analysis of Post-Training Quantization Strategies for Large Language Models: GPTQ, AWQ, and GGUF | Uplatz Blog, accessed November 16, 2025, https://uplatz.com/blog/a-comprehensive-analysis-of-post-training-quantization-strategies-for-large-language-models-gptq-awq-and-gguf/
36. Fair-GPTQ: Bias-Aware Quantization for Large Language Models - arXiv, accessed November 16, 2025, https://arxiv.org/pdf/2509.15206?
37. [Literature Review] Fair-GPTQ: Bias-Aware Quantization for Large Language Models, accessed November 16, 2025, https://www.themoonlight.io/en/review/fair-gptq-bias-aware-quantization-for-large-language-models
38. Consciousness and Unconsciousness of Artificial Intelligence - Future Human Image, accessed November 16, 2025, https://www.fhijournal.org/wp-content/uploads/2019/04/FHI_11_Piletsky.pdf
39. Complex Cognitive Systems and Their Unconscious. Related Inspired Conjectures for Artificial Intelligence - MDPI, accessed November 16, 2025, https://www.mdpi.com/1999-5903/12/12/213
40. A novel method against hardware trojans in approximate circuits - Queen's University Belfast, accessed November 16, 2025, https://pure.qub.ac.uk/en/publications/a-novel-method-against-hardware-trojans-in-

approximate-circuits/
41. Design and analysis of hardware Trojans in approximate circuits - Queen's University Belfast, accessed November 16, 2025, https://pure.qub.ac.uk/files/352933761/Electronics_Letters_2021_Dou_Design_and_analysis_of_hardware_Trojans_in_approximate_circuits.pdf
42. AI liability – who is accountable when artificial intelligence malfunctions? - Taylor Wessing, accessed November 16, 2025, https://www.taylorwessing.com/en/insights-and-events/insights/2025/01/ai-liability-who-is-accountable-when-artificial-intelligence-malfunctions
43. In support of "no-fault" civil liability rules for artificial intelligence - PMC - PubMed Central, accessed November 16, 2025, https://pmc.ncbi.nlm.nih.gov/articles/PMC7851658/
44. Legal Liability for AI-Driven Decisions – When AI Gets It Wrong, Who Can You Turn To?, accessed November 16, 2025, https://www.hfw.com/insights/legal-liability-for-ai-driven-decisions-when-ai-gets-it-wrong-who-can-you-turn-to/
45. Man vs Machine: Legal liability in Artificial Intelligence contracts and the challenges that can arise | DLA Piper, accessed November 16, 2025, https://www.dlapiper.com/insights/publications/2021/10/man-vs-machine-legal-liability-artificial-intelligence-contracts
46. Our New Artificial Intelligence Infrastructure: Becoming Locked into an Unsustainable Future, accessed November 16, 2025, https://www.mdpi.com/2071-1050/14/8/4829
47. How much energy does ChatGPT use? - Epoch AI, accessed November 16, 2025, https://epoch.ai/gradient-updates/how-much-energy-does-chatgpt-use
48. The Future Is Now: Wrestling with Ethics, Policy and Brain-Computer Interfaces, accessed November 16, 2025, https://news.ncsu.edu/2023/04/ethics-brain-computer-interfaces/
49. Ethical considerations for the use of brain–computer interfaces for cognitive enhancement - PMC - NIH, accessed November 16, 2025, https://pmc.ncbi.nlm.nih.gov/articles/PMC11542783/
50. What Is Quantization? | How It Works & Applications - MATLAB & Simulink - MathWorks, accessed November 16, 2025, https://www.mathworks.com/discovery/quantization.html
51. [2509.15206] Fair-GPTQ: Bias-Aware Quantization for Large Language Models - arXiv, accessed November 16, 2025, https://arxiv.org/abs/2509.15206
52. Gaussian Mixture Error Estimation for Approximate Circuits, accessed November 16, 2025, https://past.date-conference.com/proceedings-archive/2017/pdf/0467.pdf
53. What is a Gaussian Mixture Model? - IBM, accessed November 16, 2025, https://www.ibm.com/think/topics/gaussian-mixture-model
54. Survey on Approximate Computing and Its Intrinsic Fault Tolerance - MDPI, accessed November 16, 2025, https://www.mdpi.com/2079-9292/9/4/557
55. Approximate computing: An emerging paradigm for energy-efficient design - ResearchGate, accessed November 16, 2025, https://www.researchgate.net/publication/261378723_Approximate_computing_An_emerging_paradigm_for_energy-efficient_design
56. Towards a Standard for Identifying and Managing Bias in Artificial Intelligence - NIST Technical Series Publications, accessed November 16, 2025, https://nvlpubs.nist.gov/nistpubs/SpecialPublications/NIST.SP.1270.pdf
57. An Investigation into Neural Net Optimization via Hessian Eigenvalue Density - Proceedings of Machine Learning Research, accessed November 16, 2025, https://proceedings.mlr.press/v97/ghorbani19b/ghorbani19b.pdf