# When the Music Stops: The Architecture, Fragility, and Human Cost of the AI Boom


## Abstract

The dominant narrative in artificial intelligence development rests on a single thesis: that scaling large language models by increasing their parameters, data, and compute will reliably yield increasingly intelligent systems, ultimately approaching or achieving artificial general intelligence. Hundreds of billions of dollars have been invested in this assumption. This paper examines that thesis from three perspectives: technical, financial, and distributional. It surveys a growing body of credible critique from researchers, including Yann LeCun, Gary Marcus, Rodney Brooks, and François Chollet, who have identified fundamental architectural limitations in the current scaling paradigm. It analyzes the financial structure of the AI investment boom, identifying concentration risks, debt dependencies, and capex-to-revenue gaps that are historically consistent with asset bubbles. It also traces the distributional consequences of market corrections, showing that costs fall disproportionately on retirement accounts, regional economies, displaced workers, and democratic institutions. The paper concludes by proposing Augmented Human Intelligence (AHI) as a directional alternative: modular, biologically inspired architectures designed to enhance human judgment rather than replace it. It is transparent about the limitations of its own analysis and the open questions that remain.


## 1. Introduction

In classical music, the size of the orchestra does not determine the quality of the performance. A chamber ensemble of twelve can produce something transcendent. A symphony of a hundred poorly conducted instruments makes noise. The difference is never the number of instruments. It is always the architecture: the score, the conductor, the acoustics of the hall, and the discipline of the musicians.

This paper argues that the artificial intelligence industry is learning this lesson the hard way.

Over the past several years, the dominant narrative in AI development has centered on a single thesis: scale is the path to intelligence. More parameters, more data, and more compute will lead to general intelligence. Hundreds of billions of dollars have been invested on this assumption. Valuations have been set, infrastructure has been built, and debt has been issued, all predicated on a chain of beliefs that begins with scaling and ends with transformative returns.

This paper examines the chain, link by link. The core contention is this: if scaling alone does not yield general intelligence, then the largest capital deployment cycle in modern

technology history is built on a contested premise, and the costs of mispricing that premise will not fall on those who made the bet.

It draws on and extends three essays published in my Substack series *Architecture & Attention* (Maconochie, 2026a; 2026b; 2026c), which together explored the technical pressures on the scaling thesis, the structural fragility of the AI investment boom, and the distributional question of who bears the cost when market corrections occur. Here, I unify those arguments into a single, more rigorously sourced analysis and, critically, address what the essays did not: the limitations of my own thesis, the evidence that could falsify it, and the prior work on which it builds.

Three claims structure the argument:

First, a growing and credible body of technical critique, from researchers including Yann LeCun, Gary Marcus, Rodney Brooks, and François Chollet, among others, has identified fundamental architectural limitations in the current scaling paradigm. These critiques converge on a shared diagnosis, even as they diverge on prescription. This convergence matters.

Second, the financial structure underwriting the AI boom exhibits concentration risks, debt dependencies, and capex-to-revenue gaps that are historically consistent with asset bubbles, while acknowledging that historical analogy has real limits and that several plausible developments could change the picture.

Third, when market corrections occur in technology sectors, the costs are not borne symmetrically. They fall disproportionately on retirement accounts, regional economies, displaced workers, and democratic institutions, constituencies with the least influence over the investment decisions that created the exposure.

The paper concludes by connecting these threads to an alternative framework I have been developing: Augmented Human Intelligence (AHI), which argues that the path forward lies not in building bigger models but in designing architectures that enhance human judgment rather than attempt to replace it (Maconochie, 2025a; 2025b).

A note on intellectual honesty. I am not a disinterested observer. My perspective has been shaped by twenty-five years in technology consulting and by a personal displacement that redirected my career toward the research and writing that produced this paper. I discuss that experience in detail in Section 4.3, because I believe transparency about vantage point is a prerequisite for the kind of analysis that follows.

## 2. The Scaling Thesis Under Pressure

The dominant AI development paradigm rests on what might be called the scaling hypothesis: that increasing model size, training data, and computational resources will reliably yield more capable systems, and that this trajectory, if sustained, leads to artificial

general intelligence. This hypothesis has driven hundreds of billions of dollars in investment and shaped the strategic priorities of every major technology company.

It is not without basis. The scaling laws documented by Kaplan et al. (2020) demonstrated predictable relationships between model size, dataset size, compute budget, and performance. Each generation of large language models has exhibited genuinely impressive capabilities that its predecessor lacked. The progress is real, and any honest analysis must begin by acknowledging it.

But a growing body of credible critique suggests that the relationship between scale and capability is neither linear nor unlimited and is insufficient to reach general intelligence. What follows is a survey of four prominent voices in this critique, chosen not because they agree on solutions but because their diagnoses converge in ways that are difficult to dismiss.

## 2.1 Yann LeCun: World Models, Not Word Models

LeCun, Executive Chairman of Advanced Machine Intelligence Labs and a Turing Award laureate (formerly Meta's Chief AI Scientist for over a decade), has been among the most prominent critics of the scaling-to-AGI thesis. His central argument is that LLMs operate entirely within language and lack the grounded world models that biological intelligence relies on. "LLMs are useful, but they are an off-ramp on the road to human-level AI," he has stated publicly (LeCun, 2024). His proposed alternative, Joint Embedding Predictive Architectures (JEPA), aims to build systems that learn representations of the world through sensory experience rather than text prediction (LeCun, 2022). In late 2025, LeCun left Meta to found AMI Labs specifically to pursue this vision, a move that underscores the depth of his conviction that the current paradigm is insufficient (MIT Technology Review, 2026).

As recently as February 2026, LeCun reinforced this position at the AI Impact Summit, describing AI's most significant near-term role as "an amplifier for human intelligence" rather than a replacement for it, and cautioning against anthropomorphizing LLMs or mistaking language manipulation for genuine understanding (LeCun, 2026).

The implication is architectural, not incremental. LeCun is not arguing that LLMs need to be bigger. He is arguing that they need to be different.

## 2.2 Gary Marcus: The Reasoning Gap

Marcus, a cognitive scientist and persistent critic of deep learning orthodoxy, focuses on a different failure mode: the gap between pattern matching and genuine reasoning. His argument, developed across multiple books and publications (Marcus, 2020; Marcus & Davis, 2019), is that LLMs excel at statistical interpolation but cannot perform the causal, compositional reasoning that even young children demonstrate reliably. Hallucinations, in his framing, are not a bug to be fixed but a symptom of a fundamental architectural limitation.

Marcus advocates for hybrid architectures that integrate neural networks with symbolic reasoning systems (Marcus & Davis, 2019), an approach with roots in classical AI that has gained renewed interest as the limits of pure scaling become more apparent. His critique is not that deep learning is useless but that it is incomplete, and that completing it requires architectural innovation rather than further scaling.

## 2.3 Rodney Brooks: The Competence Illusion

Brooks, co-founder of iRobot and former director of MIT's Computer Science and Artificial Intelligence Laboratory, brings an embodied robotics perspective to the critique. His contribution is less about what AI cannot do and more about how humans systematically misjudge what it can do. "When a human sees an AI system perform a task, they immediately generalize it to things that are similar and make an estimate of the competence of the AI system. And they are usually very over-optimistic" (Brooks, 2024).

This matters for the scaling thesis because much of the investment case rests on extrapolation: if GPT-4 can do X, then GPT-5 will surely do X+Y. Brooks argues that this extrapolation is a cognitive bias, not an engineering forecast. He has maintained a public timeline of AI predictions, consistently demonstrating that expert and popular projections overshoot actual capability delivery (Brooks, 2018, updated annually).

## 2.4 François Chollet: Memorization Is Not Intelligence

Chollet, creator of the Keras deep learning framework and researcher at Google, has made perhaps the most precise technical contribution to the scaling critique. His ARC (Abstraction and Reasoning Corpus) benchmark (Chollet, 2019) was designed to test fluid intelligence, the ability to adapt to genuinely novel problems, as distinct from crystallized intelligence, the application of memorized patterns. On ARC, GPT-4o achieves approximately 5% accuracy. Humans achieve 84% (ARC Prize Foundation, 2024).

This gap is not a matter of scale. It reflects a structural difference in how LLMs process information versus how humans reason. "Memorization is useful, but intelligence is something else" (Chollet, 2024). Scaling current architectures will improve crystallized performance (retrieval, pattern application) while leaving the gap in fluid intelligence largely untouched.

## 2.5 The Convergence

These four critics approach the problem from different disciplines and propose different solutions. LeCun wants embodied world models. Marcus wants symbolic-neural hybrids. Brooks wants constrained domains with realistic expectations. Chollet wants benchmarks for program synthesis and abstraction.

But their diagnoses converge on four points that, taken together, constitute a serious challenge to the scaling hypothesis:

First, that scaling is encountering diminishing returns. Performance gains per unit of additional compute are flattening, particularly for tasks that require reasoning rather than retrieval.

Second, that current benchmarks mask fundamental limitations. High scores on contaminated or narrow evaluations do not equate to general intelligence, and the benchmark ecosystem has not kept pace with the claims being made about model capability (Kiela et al., 2021).

Third, that grounding and causal reasoning are architectural requirements, not emergent properties. No amount of text prediction will spontaneously produce an understanding of how the physical or social world actually works (Bender & Koller, 2020; Mitchell, 2023).

Fourth, that architectural innovation, not incremental scaling, is the most promising path forward. The disagreement is about which architecture, not about whether the current one is sufficient.

It is worth noting what this critique does not claim. It does not claim that LLMs are useless. It does not claim that scaling has produced no value. And it does not claim that the critics' proposed alternatives will succeed. What it claims is that the specific bet underpinning hundreds of billions of dollars in investment, that scaling current architectures leads to AGI and the transformative returns that AGI would justify, is contested by serious, credentialed researchers with substantive technical arguments.

This matters because the financial structure built on top of the scaling thesis assumes it holds. If this diagnosis is even partially correct, the financial superstructure built on "scaling leads to AGI leads to transformative ROI" is mispriced, which makes the fragility examined in Section 3 less a market curiosity than a systemic risk.

## 3. The Structural Fragility of the AI Boom

The technical critique outlined in Section 2 would be primarily of academic interest if it were not underwriting one of the largest capital deployment cycles in modern technology history. But the scaling hypothesis is not merely a research program. It is an investment thesis, and hundreds of billions of dollars have been committed on the assumption that it holds.

This section examines the financial structure of the AI boom, not to predict a crash, but to identify the specific points of structural fragility that would amplify the consequences if the scaling thesis proves insufficient.

### 3.1 The Investment Thesis

The logic chain is straightforward. Scaling works. Scaling leads to AGI, or at a minimum to transformative productivity gains. Those gains justify enormous upfront infrastructure

investment. The winners of this race will capture outsized returns, creating winner-take-all dynamics that reward early, aggressive spending.

Each link in this chain has been used to justify the next. OpenAI's valuation assumes the company will capture a significant share of all future knowledge work. Nvidia's market capitalization assumes AI infrastructure spending will continue accelerating for years. The hyperscalers (Microsoft, Google, Amazon, Meta) are investing at rates that only make sense if AI fundamentally transforms their core businesses and the broader economy.

The bet is that training costs are front-loaded and inference revenue will follow. Build the models now, monetize them later. This is a reasonable bet. But it is a bet, and its reasonableness depends on the validity of the assumptions beneath it.

## 3.2 The Cracks

Several of those assumptions are under pressure.

First, the enterprise adoption gap. The promise of generative AI was that it would rapidly transform business workflows, driving productivity gains that would justify subscription and licensing revenue at scale. The early evidence is more ambiguous. Studies by McKinsey (2024) and BCG (2024) have found that the majority of companies experimenting with generative AI have not yet achieved meaningful, measurable returns. The pattern is familiar to anyone who has observed technology adoption cycles: impressive demonstrations, difficult deployment, unclear value at production scale (Gartner, 2024). This does not mean enterprise value will never materialize, but the timeline is longer and less certain than the investment thesis requires.

Second, the revenue concentration problem. Nvidia is generating historic profits selling the infrastructure of the AI boom (Nvidia, 2024). But the companies building on that infrastructure largely are not. OpenAI reportedly lost approximately $5 billion in 2024 (The Information, 2024). Most AI startups are burning capital without a clear path to profitability. When the primary beneficiary of a gold rush is the company selling shovels, it is worth asking who is finding gold.

Third, the demo-to-deployment gap. There is a persistent and well-documented gap between what AI systems can do in controlled demonstrations and what they can reliably do in production environments (Sambasivan et al., 2021). Benchmarks improve; operational reliability does not improve at the same rate. Hallucinations persist. Enterprise customers discover that "90% accurate" is not sufficient for mission-critical workflows where the cost of the remaining 10% is unacceptable. This gap is not merely a matter of fine-tuning. It reflects the architectural limitations discussed in Section 2: systems that interpolate from training data rather than reasoning about novel situations will produce confident errors that are difficult to predict or prevent.

## 3.3 Concentration Risk

Zoom out from individual companies, and the systemic picture becomes visible.

The Magnificent Seven technology companies increased their energy consumption by 19% in 2023 while the median S&P 500 company's consumption remained flat (Goldman Sachs, 2024). By some analyses, a relatively small set of AI-exposed companies accounted for roughly three-quarters of S&P 500 returns over the post-ChatGPT period (S&P Global, 2025). This is not a broad-based technology boom. It is a narrow bet by the market on a single thesis.

JPMorgan estimates that AI-related investment-grade bond issuances could reach $1.5 trillion by 2030 (JPMorgan Research, 2024). Much of this debt is predicated on productivity gains that may or may not materialize. If the gains come more slowly than projected, or in different forms than anticipated, the debt does not disappear. It becomes a drag on the companies and, indirectly, on the pension funds, insurance companies, and retirement accounts that hold those bonds.

The concentration extends beyond financial markets. Three companies (Nvidia, TSMC, and ASML) control critical chokepoints in the AI hardware supply chain (Semiconductor Industry Association, 2024). A small number of cloud providers control the infrastructure on which most AI applications run. This consolidation creates fragility: disruption at any single node, whether technical, geopolitical, or regulatory, propagates through the entire ecosystem.

## 3.4 Historical Parallels and Their Limits

A necessary clarification before proceeding. Identifying bubble-like structure is not the same as claiming AI is fraudulent, nor is it a prediction that a crash is imminent. The technology is real. The financial structure around it may nonetheless be mispriced. These are compatible observations, and conflating them is one of the most common errors in public discourse about technology investment.

We have seen versions of this pattern before, and the historical parallels are instructive precisely because they are imperfect.

The dot-com boom of the late 1990s left behind real infrastructure: fiber optic networks, data centers, a generation of internet-native companies, and foundational protocols. Most of the companies that raised money during the bubble failed. The Nasdaq fell nearly 80% from peak to trough. But the technology was real and eventually transformed the economy (Perez, 2002). The bubble was not wrong about the internet. It was wrong about the timeline and about which companies would capture the value.

The cryptocurrency boom left behind less. Blockchain has genuine use cases, but the speculative frenzy produced more fraud than lasting value. The infrastructure that survived is narrower, and the transformative promise remains largely unrealized.

AI will probably land somewhere between these precedents. The technology is real. Transformer architectures, large language models, and diffusion models are genuine innovations with genuine applications (Bommasani et al., 2021). The question is not whether AI is useful (it clearly is, in specific domains) but whether current valuations,

infrastructure spending, and debt levels are proportionate to the actual value being created on a timeline that justifies the capital at risk.

Useful and overhyped are not mutually exclusive. This is perhaps the most important sentence in this section. Much of the public discourse treats the AI investment question as binary: either AI is transformative, and all investment is justified, or AI is a bubble, and all investment is wasted. The more honest assessment is that AI is a genuinely important technology whose current financial structure has outrun its demonstrated economic returns, creating a gap that will eventually close in one direction or the other.

### 3.5 What Could Change the Picture

Intellectual honesty requires identifying the conditions under which this analysis would be wrong or premature.

If enterprise adoption accelerates significantly in 2026 and 2027, with measurable productivity gains translating to revenue growth for AI providers, the investment thesis strengthens. The current gap between investment and returns could simply be a timing issue, as it was for cloud computing, which took longer than expected to demonstrate ROI but ultimately justified the infrastructure spending.

If new architectures (whether LeCun's world models, hybrid approaches, or something not yet developed) unlock capabilities that current scaling cannot, the infrastructure already built could be repurposed in ways that validate the capital deployment even if the original thesis was wrong about the mechanism.

If geopolitical competition, particularly between the United States and China, sustains government-backed investment regardless of commercial returns, the market dynamics shift from commercial ROI to strategic spending, which follows a different logic and different timelines.

These are plausible scenarios, not speculative ones. Any honest assessment of structural fragility must acknowledge that fragility does not guarantee failure. Markets can remain stretched for longer than skeptics expect (a lesson Michael Burry, who is currently shorting Nvidia and Palantir, has learned more than once). The purpose of identifying fragility is not to predict timing but to understand what is at risk and who bears the consequences if the correction comes. That is the subject of Section 4.

## 4. Who Bears the Cost

Technology bubbles are often discussed in terms of valuations, stock prices, and investor returns. But the most consequential effects of market corrections are distributive. They fall unevenly, and they fall hardest on constituencies with the least influence over the investment decisions that created the exposure.

This section examines four categories of cost-bearing that tend to be underrepresented in mainstream coverage of AI market dynamics: retirement savings, regional economies, labor markets, and governance capacity.

## 4.1 The Retirement Savings Channel

When AI stocks decline, the most immediate and widely distributed impact flows through retirement accounts. The S&P 500 and related indices are heavily weighted toward the technology companies driving the AI boom. As of late 2025, the Magnificent Seven accounted for roughly 30% of the S&P 500's total market capitalization (S&P Global, 2025). This means that every index fund, target-date retirement fund, and pension allocation benchmarked to these indices carries significant concentrated exposure to the AI thesis, whether the account holder chose that exposure or not.

This is not a niche concern. Approximately 60% of American workers participate in employer-sponsored retirement plans, and the vast majority of those plans are invested in index funds or target-date funds with substantial large-cap technology exposure (Investment Company Institute, 2024). A 30% decline in the Magnificent Seven, which would be modest by the standards of historical technology corrections, would reduce the retirement savings of millions of Americans who never made a conscious decision to bet on AI.

The asymmetry is important. The executives and early investors in AI companies have access to diversification strategies, hedging instruments, and liquidity options that ordinary retirement savers do not. When the correction comes, the people who made the bet and the people who bear the consequences are largely different populations.

## 4.2 Regional Economic Exposure

The AI boom has concentrated infrastructure investment in specific geographies: Northern Virginia, central Oregon, west Texas, and parts of the Midwest, where cheap energy and favorable regulatory environments have attracted data center construction (CBRE, 2024). These communities have benefited from construction employment, tax revenue, and economic activity associated with the buildout.

But infrastructure booms create dependencies. Local governments issue bonds and make spending commitments based on projected tax revenue from facilities that may or may not operate at capacity over their expected lifetimes. Construction employment is temporary by nature; operational employment at data centers is minimal relative to the physical footprint. If infrastructure spending slows or facilities are underutilized, the communities that planned around the boom absorb the gap between projected and actual economic activity.

This pattern has precedent. Regions that experienced rapid buildout during the fiber-optic boom of the late 1990s saw significant economic disruption when the bubble burst, leaving behind dark fiber, unfinished developments, and municipal budgets planned around

growth that did not materialize (Blumenstein, 2002). The AI infrastructure buildout is larger in scale and more geographically concentrated, amplifying both the potential benefits and the potential disruptions.

## 4.3 Labor Market Displacement

The labor market effects of the AI boom operate through a mechanism that is both more complex and less visible than the direct job replacement that dominates public discussion.

The more immediate dynamic is displacement by attrition and restructuring. Companies adopting AI tools do not typically announce mass layoffs attributed to AI. Instead, they restructure roles, decline to backfill departing employees, and gradually shift task assignments to reduce headcount over time. Survey measures of AI adoption vary widely depending on the definitions used, but government and private data both show meaningful, rising adoption, especially in information and professional services. The U.S. Census Bureau's Business Trends and Outlook Survey found AI use approaching 10% of firms overall by mid-2025, with substantially higher rates in technology-intensive sectors (U.S. Census Bureau, 2025). Private surveys using broader definitions of AI adoption report considerably higher figures, underscoring that the measurement itself remains contested even as the direction of travel is clear.

It is important to be transparent about the limitations of the available data. The relationship between AI adoption and net employment is genuinely contested. Some studies suggest that AI adoption complements employment, creating new roles and increasing productivity, thereby generating additional labor demand (Acemoglu & Restrepo, 2019). Others find net displacement effects, particularly for mid-skill, routine cognitive tasks (Autor, 2024). The honest assessment is that we do not yet have sufficient longitudinal data to determine which effect dominates, and the answer likely varies by sector, geography, and time horizon.

What is less contested is the transitional cost. Even if AI ultimately creates more jobs than it displaces (a plausible but unproven claim), the transition is not seamless. Workers displaced from one sector do not automatically reappear in another. Retraining takes time and resources. Geographic mobility has costs. And the workers most vulnerable to displacement, those in routine cognitive roles, are often least equipped for the roles that AI adoption creates, which tend to require higher levels of technical skill or creative judgment (Autor, Levy, & Murnane, 2003; Frey & Osborne, 2017).

My own experience illustrates this dynamic at the individual level. In September 2023, after twenty-five years in technology consulting, I was part of a reduction-in-force. For the first twelve months, I did what most displaced professionals do: I operated on autopilot, driven by fear, trying to replace the role I had lost with something as similar as possible. The market for mid-career technology consultants had contracted, partly due to broader economic conditions and partly due to the early effects of AI adoption on consulting workflows. I sent applications, took calls, and competed for roles that were becoming scarcer.

Then I stopped. Not because I had given up, but because I recognized that the disruption was not just a career setback. It was an opportunity to take a genuine stock of what I wanted to do with the rest of my professional life. That pivot led me to the research and writing that produced this paper. I was fortunate: I had savings, a supportive family, and the intellectual background to redirect. Most displaced workers do not have those buffers. The transition cost is real, unevenly distributed, and largely invisible in the aggregate economic data that policymakers rely on.

### 4.4 The Governance Deficit

Perhaps the most consequential and least discussed cost of the AI boom is its effect on democratic governance capacity.

Four specific dynamics are worth identifying.

First, regulatory capture through complexity. AI systems are sufficiently complex that effective oversight requires technical expertise that most regulatory bodies do not possess (Cath et al., 2018). This creates an information asymmetry between the companies deploying AI and the institutions responsible for governing its use. The result is not overt corruption but structural dependency: regulators rely on the entities they regulate for the technical understanding necessary to regulate them.

Second, the speed mismatch. AI capabilities evolve on timescales measured in months. Legislative and regulatory processes operate on timescales measured in years. This gap means that by the time governance frameworks are developed, the technology they address has often moved on, creating a perpetual state of regulatory lag that effectively functions as deregulation by default (Marchetti, 2022).

Third, federal preemption of state and local governance. In the United States, there has been a concerted effort by industry to establish federal AI frameworks that would preempt state-level regulation (Engler, 2024). While proponents argue this prevents a patchwork of inconsistent rules, the practical effect is to concentrate governance authority at the level of government most susceptible to industry lobbying, while removing the ability of states and municipalities to address local impacts. This represents a form of moral hazard: the entities creating the externalities are actively shaping the governance structures that would hold them accountable.

Fourth, the democratic attention deficit. The sheer pace of AI development, combined with its technical complexity, overwhelms democratic publics' capacity to engage meaningfully with the policy questions it raises (Nemitz, 2018). Citizens cannot exercise informed judgment about AI governance if the technology, its risks, and its distributional consequences are not legible to them. This is not a failure of public intelligence. It is a failure of the information architecture that mediates between technological change and democratic deliberation.

These four dynamics compound each other. Regulatory complexity discourages engagement. Speed mismatches reduce accountability. Preemption concentrates

authority. And the resulting governance vacuum is filled by the same entities whose activities require governance, a feedback loop that is structurally difficult to interrupt.

### 4.5 The Distributional Pattern

Taken together, these four channels reveal a consistent distributional pattern: the benefits of the AI boom accrue primarily to a narrow set of companies, investors, and highly skilled workers, while the costs and risks are distributed broadly across retirement savers, regional economies, displaced workers, and democratic institutions.

This is not unique to AI. It is a recurring feature of technology-driven asset bubbles (Perez, 2002). But the scale of the current AI investment cycle, the concentration of market exposure, and the speed of potential labor displacement make the distributional stakes unusually high. Understanding this pattern is not pessimism. It is a prerequisite for designing responses that are proportionate to the actual risks.

## 5. The Alternative: Architecture Over Scale

The preceding sections have argued that the scaling thesis is under credible technical pressure, that the financial structure built on it is fragile, and that the costs of correction fall disproportionately on those least positioned to absorb them. A natural question follows: if not scaling, then what?

This paper does not claim to have a complete answer. But it does propose a directional alternative, one that has been developing across my earlier research and that draws on the same body of work the scaling critics cite, even if they have not yet assembled it in quite this way.

The alternative is architectural. And its organizing principle is augmentation rather than replacement.

### 5.1 The Biological Precedent

In an earlier paper (Maconochie, 2025a), I introduced the concept of the Evolutionary Processing Unit (EPU): the cumulative computational effort of approximately four billion years of biological evolution that produced the human brain. The central calculation is striking. The estimated total number of “brain-equivalent FLOPS” executed across all Homo sapiens who have ever lived is approximately 5.5 x 10^38. The most powerful supercomputer ever built would require roughly 10 trillion years to replicate that accumulated processing power. Evolution did not produce intelligence through brute-force computation. It produced intelligence through architectural innovation.

The product of that process, what I term the Biological Processing Unit (BPU), is not a monolithic processor. It is a confederation of specialized modules (sensory, memory, emotional, motor) coordinated by a dynamic executive function in the prefrontal cortex, all operating on approximately 20 watts of power (Maconochie, 2025b; Bennett, 2023). The

BPU's architecture exhibits four properties that current AI systems largely lack: modular specialization with integration, continuous plasticity and learning, embodied and causal reasoning, and attention as a resource allocation mechanism.

These are not incidental features. They are the reason the BPU achieves robust, generalizable intelligence at a fraction of the energy cost of current AI systems. The architectural lesson is that intelligence is not a function of scale. It is a function of how components are organized, coordinated, and dynamically allocated.

## 5.2 Augmented Human Intelligence

This architectural perspective leads to a different goal than the one currently driving most AI investment.

The prevailing paradigm aims at Artificial General Intelligence (AGI): autonomous systems that match or exceed human cognitive capability across all domains. The investment thesis described in Section 3 depends on this goal because only a transformative, general-purpose capability justifies the scale of capital being deployed.

The alternative I have been developing, Augmented Human Intelligence (AHI), starts from a different premise. Rather than building systems that replicate human intelligence, AHI designs architectures that enhance it. The distinction is not merely rhetorical. It implies fundamentally different design priorities.

AGI prioritizes autonomy. AHI prioritizes partnership. AGI seeks to minimize human involvement. AHI seeks to optimize it. AGI requires solving the full problem of general intelligence. AHI requires solving a more tractable problem: identifying which cognitive tasks are best handled by machines (retrieval, pattern recognition, tireless execution) and which are best handled by humans (judgment, meaning-making, contextual reasoning, genuine novelty), then designing the interface between them (Maconochie, 2025a; Maconochie, 2025b; Licklider, 1960; Engelbart, 1962).

This reframing changes the cost calculus described in Sections 3 and 4. If the goal is augmentation rather than replacement, the infrastructure requirements are different (more distributed, less concentrated), the labor market dynamics are different (complementary rather than substitutive), and the governance challenges, while still real, become more tractable because humans remain in the loop.

The economic implications deserve explicit attention because they connect the philosophical argument to the structural fragilities identified in Section 3. An augmentation-oriented paradigm shifts capital intensity away from hyperscale GPU clusters optimized for ever-larger training runs and toward human-computer interface design, domain-specific tooling, and distributed inference infrastructure. This reduces the capex-to-revenue mismatch at the heart of the current boom, because modular, specialized systems can demonstrate value in narrower domains on shorter timelines, rather than requiring transformative general-purpose capability to justify their cost. It also reduces concentration risk: an ecosystem of specialized AI tools serving specific

professional domains is structurally less fragile than one in which a handful of foundation model providers control the entire value chain. None of this eliminates investment risk, but it distributes it more broadly and ties it more directly to demonstrable, domain-specific returns.

## 5.3 Convergence with the Critics

What is notable is how closely the scaling critics' own positions converge toward augmentation, even when they do not frame it that way.

LeCun's world model architectures are designed to give machines a grounded understanding of reality, but his most recent public statements frame the near-term significance of AI explicitly as "an amplifier for human intelligence" rather than a replacement for it (LeCun, 2026). Brooks has consistently argued for constrained, domain-specific systems that work alongside humans rather than autonomously (Brooks, 2018). Marcus's hybrid architectures implicitly preserve a role for human oversight in tasks requiring genuine reasoning (Marcus & Davis, 2019). Even Chollet's ARC benchmark, by defining intelligence as fluid adaptation to novelty, points to a capability that humans possess and machines do not, suggesting that the most productive path is to leverage rather than replicate it (Chollet, 2019).

The pieces are present in the critics' work. They share a diagnosis (scaling is insufficient), and their prescriptions, while diverse, all imply architectures that are more modular, more grounded, and more oriented toward human-machine collaboration than the monolithic scaling paradigm permits. What has been missing is an explicit framework that names this convergence and articulates it as a design philosophy rather than merely a collection of critiques.

AHI is an attempt to provide that framework. It is not a complete technical specification. It is a directional claim: that the path to genuinely useful AI runs through architecture and augmentation, not through scale and autonomy, and that this path is more achievable, more economically sustainable, and more compatible with democratic governance than the alternative currently being pursued.

## 5.4 Attention as the Integrating Mechanism

One further connection deserves mention, because it links the technical argument to the distributional concerns raised in Section 4.

Across biological, artificial, and social systems, the mechanism that integrates specialized modules into coherent action is attention: the selective allocation of finite processing resources to the most relevant inputs. The prefrontal cortex performs this function in the brain (Miller & Cohen, 2001). The Transformer architecture (Vaswani et al., 2017) performs a version of it in LLMs, though without the contextual judgment that biological attention provides. And in human life, the deliberate management of attention is what enables

individuals to navigate complexity without being overwhelmed by it (Maconochie, 2025c; James, 1890; Kahneman, 2011).

The AI boom, as described in Sections 3 and 4, represents a failure of collective attention allocation. Enormous resources have been directed toward a single thesis (scale leads to AGI) while alternative approaches have been comparatively neglected. The governance deficit described in Section 4.4 is, at root, an attention deficit: democratic institutions cannot allocate sufficient attention to AI governance because the pace and complexity of the technology exceed their processing capacity.

AHI addresses this directly. By keeping humans in the loop as the executive function (the conductor, not the instrument), it preserves the capacity for contextual judgment, value-weighted prioritization, and adaptive response that attention provides. By designing modular, specialized AI components rather than monolithic general-purpose systems, it reduces the complexity that overwhelms governance. And by framing the goal as augmentation rather than replacement, it aligns the incentives of AI development with the interests of the broader populations who bear the costs when things go wrong.

## 6. Limitations and Open Questions

Any analysis that identifies structural fragility in a multi-trillion-dollar investment cycle and proposes an alternative framework has an obligation to be transparent about what it does not know, where its evidence is weakest, and what developments could render its conclusions premature or wrong.

### 6.1 Limitations of This Analysis

Several limitations should be stated directly.

First, the financial data cited in Section 3 is drawn primarily from industry reports, investment bank research, and business journalism rather than peer-reviewed academic sources. This is partly unavoidable: the AI investment cycle is moving faster than the academic publication cycle, and the most current data exists in industry research. But it means that some figures (energy consumption trends, bond issuance projections, enterprise adoption rates) carry the caveats inherent to their sources. Where possible, I have cited the originating institution so readers can assess each source's credibility and potential bias independently.

Second, the labor market analysis in Section 4.3 relies on data that is genuinely preliminary. The relationship between AI adoption and net employment is being studied in real time, and the longitudinal data necessary to distinguish between transitional displacement and structural job loss does not yet exist in sufficient depth. I have tried to represent the contested nature of this evidence honestly, but readers should understand that this is an area where confident claims in either direction outrun the available data.

Third, the scaling critics surveyed in Section 2, while credible and influential, do not represent the consensus view within the AI research community. Many serious researchers continue to believe that scaling, combined with architectural refinements, will yield further significant capability gains (Sutskever, 2023; Amodei, 2024). The survey approach used here (selecting four prominent critics and identifying their convergence) risks overstating the degree of professional consensus against the scaling thesis. The convergence I identify is real, but it coexists with a substantial body of opinion that disagrees.

Fourth, the AHI framework proposed in Section 5 is directional rather than operational. It articulates a design philosophy and connects it to biological precedent, but it does not provide the detailed technical architecture needed to implement it. This is a genuine limitation, and I want to be direct about its source. While I am technically literate, with a background in civil engineering and twenty-five years in technology consulting, I do not have the qualifications or expertise to take the AHI framework to a technical implementation specification. For that, I would need collaborators with complementary skills: researchers in cognitive architecture, modular AI systems, and human-computer interaction who share the conviction that this is a worthwhile alternative avenue of exploration. The framework is an invitation to that collaboration, not a substitute for it.

Fifth, my own position as a displaced technology professional, while it provides a relevant experiential perspective, also introduces potential bias. The experience of being part of a reduction-in-force may incline me to interpret labor-market signals as more threatening than they are and to favor frameworks that preserve human roles in AI systems. I believe my analysis is sound on its merits, but the reader should weigh this potential bias alongside the evidence presented.

## 6.2 Open Questions

Beyond the limitations of this specific analysis, several genuinely open questions will shape whether the concerns raised here prove prescient or premature.

Can scaling produce emergent capabilities that current benchmarks do not measure? It is possible that larger models will exhibit qualitatively new behaviors that current evaluation frameworks cannot capture. The history of AI development includes genuine surprises: capabilities that emerged at scale without explicit training (Wei et al., 2022). If such emergent capabilities materialize in commercially significant ways, the investment thesis strengthens even if the specific path to AGI remains unclear.

Will new architectures validate or strand existing infrastructure? The hundreds of billions invested in GPU clusters, data centers, and training infrastructure are currently optimized for the scaling paradigm. If alternative architectures (world models, hybrid systems, modular approaches) prove more productive, some of that infrastructure may be repurposable. But some may not. The degree to which current infrastructure transfers to future paradigms is an open question in engineering and economics.

How will geopolitical competition affect the timeline? The AI investment cycle is not purely commercial. It is embedded in strategic competition between the United States, China, and the European Union, among others (Ding & Dafoe, 2024). Government-backed investment may sustain infrastructure spending beyond what commercial returns alone would justify, altering the dynamics described in Section 3 in ways difficult to predict from economic analysis alone.

What governance models will emerge? Section 4.4 described a governance deficit, but deficits can be addressed. If democratic institutions develop effective AI governance frameworks, whether through technical standards, regulatory innovation, or international cooperation (Floridi et al., 2018), some of the distributional risks identified in Section 4 could be substantially mitigated. The governance challenge is real, but it is not necessarily permanent.

Is augmentation a stable equilibrium? The AHI framework assumes that human-AI partnership is a durable design goal rather than a transitional phase. But it is possible that, once built, augmentation systems create pressures toward greater automation over time, as economic incentives favor reducing the human component (Acemoglu & Restrepo, 2019). If so, AHI may be a waystation rather than a destination, and the labor and governance concerns raised in Section 4 would reassert themselves in a different form.

These questions are not rhetorical. They represent genuine uncertainties that will be resolved by evidence, not argument. The purpose of raising them is not to hedge but to demonstrate that the analysis presented here is offered in the spirit of intellectual honesty rather than certainty. The structural fragilities are real. The distributional risks are real. The alternative framework has genuine merit. But the future is not determined, and any analysis that claims otherwise is selling something that nobody is in a position to sell.

## 7. Conclusion

This paper began with an orchestra. It is worth returning to one.

The AI industry has spent the better part of a decade assembling the largest orchestra ever constructed. More instruments, more musicians, more volume. The hypothesis has been that if you build it big enough, something extraordinary will emerge. And in fairness, something has. Large language models are genuinely impressive instruments. They retrieve, synthesize, and generate language with a fluency that was unimaginable a decade ago.

But fluency is not intelligence. Volume is not music. And the question that the scaling thesis has never adequately answered is not whether bigger models can do more things, but whether doing more things is the same as doing the right things, in the right way, for the right reasons.

The scaling critics surveyed in Section 2 have identified, from different vantage points and with different proposed solutions, that the current architecture has fundamental limits. The financial analysis in Section 3 has shown that the investment structure built on the assumption that those limits do not exist, or will be overcome by further scaling, exhibits concentration risks and structural fragilities that are historically consistent with asset bubbles. And Section 4 has traced the distributional consequences: when corrections come, the costs do not fall on the people who made the bets. They fall on retirement accounts, regional economies, displaced workers, and democratic institutions that were never consulted about the level of exposure they were absorbing.

None of this means the music stops tomorrow. Markets can remain stretched for extended periods. New capabilities may emerge that change the calculus. Geopolitical dynamics may sustain investment beyond what commercial logic would dictate. The limitations section of this paper is not a formality. It reflects genuine uncertainty about timing, magnitude, and resolution.

But the structural pattern is legible, and it points in a direction that deserves more attention than it is currently receiving.

The alternative proposed in Section 5, Augmented Human Intelligence, is not a prediction. It is a design philosophy. It says: rather than building ever-larger monolithic systems in pursuit of autonomous general intelligence, design modular, specialized architectures that enhance human judgment rather than attempt to replace it. This is what evolution did over 4 billion years on 20 watts. It is what the scaling critics are converging toward, even when they do not name it. And it is what a Turing Award laureate described, just days ago, as AI's most significant near-term role: an amplifier for human intelligence.

I do not claim that AHI is a complete solution. It is a directional framework that requires collaborators with expertise I do not possess to develop into operational specifications. But I believe the direction is sound because it aligns the technical evidence, economic incentives, and human interests in a way the scaling paradigm currently does not.

The orchestra metaphor holds one final lesson. The greatest performances are never about the size of the ensemble. They are about the score, the conductor, the acoustics of the hall, and the discipline of the musicians. They are, in a word, about architecture.

The AI industry has built an enormous orchestra. The question now is whether we will write a score worth playing, in a hall designed to carry the sound to everyone, conducted with the judgment and restraint that the moment demands.

If we do not choose the architecture, the architecture will choose us. And its choices will not optimize for the things we care about.

# References

## Academic and Peer-Reviewed Sources


Acemoglu, D., & Restrepo, P. (2019). Automation and new tasks: How technology displaces and reinstates labor. *Journal of Economic Perspectives*, 33(2), 3-30.

Amodei, D. (2024). Machines of loving grace: How AI could transform the world for the better. *Dario Amodei's Blog*.

Autor, D. (2024). Applying AI to rebuild middle class jobs. *NBER Working Paper No. 32140*.

Autor, D., Levy, F., & Murnane, R. J. (2003). The skill content of recent technological change: An empirical exploration. *Quarterly Journal of Economics*, 118(4), 1279-1333.

Bender, E. M., & Koller, A. (2020). Climbing towards NLU: On meaning, form, and understanding in the age of data. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 5185-5198.

Bennett, M. (2023). *A Brief History of Intelligence: Evolution, AI, and the Five Breakthroughs That Made Our Brains*. New York: Mariner Books.

Blumenstein, R. (2002). Fiber-optic glut and telecom bust leave towns dark. *The Wall Street Journal*, June 18, 2002.

Bommasani, R., Hudson, D. A., Adeli, E., et al. (2021). On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*.

Cath, C., Wachter, S., Mittelstadt, B., Taddeo, M., & Floridi, L. (2018). Artificial intelligence and the "good society": The US, EU, and UK approach. *Science and Engineering Ethics*, 24(2), 505-528.

Chollet, F. (2019). On the measure of intelligence. *arXiv preprint arXiv:1911.01547*.

Ding, J., & Dafoe, A. (2024). Engines of power: Electricity, AI, and general-purpose military transformations. *Journal of Strategic Studies*, 47(1), 1-32.

Engelbart, D. C. (1962). Augmenting human intellect: A conceptual framework. *SRI Summary Report AFOSR-3223*.

Floridi, L., Cowls, J., Beltrametti, M., et al. (2018). AI4People: An ethical framework for a good AI society. *Minds and Machines*, 28(4), 689-707.

Frey, C. B., & Osborne, M. A. (2017). The future of employment: How susceptible are jobs to computerisation? *Technological Forecasting and Social Change*, 114, 254-280.

James, W. (1890). *The Principles of Psychology*. New York: Henry Holt and Company.

Kahneman, D. (2011). *Thinking, Fast and Slow*. New York: Farrar, Straus and Giroux.

Kaplan, J., McCandlish, S., Henighan, T., et al. (2020). Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*.

Kiela, D., Bartolo, M., Nie, Y., et al. (2021). Dynabench: Rethinking benchmarking in NLP. *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics*, 4110-4124.

LeCun, Y. (2022). A path towards autonomous machine intelligence. *OpenReview preprint*.

Licklider, J. C. R. (1960). Man-computer symbiosis. *IRE Transactions on Human Factors in Electronics*, HFE-1, 4-11.

Marchetti, R. (2022). The ethics of AI governance: Regulatory gaps and the speed of innovation. *Philosophy & Technology*, 35(3), 72.

Marcus, G. (2020). *Rebooting AI: Building Artificial Intelligence We Can Trust* (with E. Davis). New York: Vintage Books.

Marcus, G., & Davis, E. (2019). *Rebooting AI: Building Artificial Intelligence We Can Trust*. New York: Pantheon.

Miller, E. K., & Cohen, J. D. (2001). An integrative theory of prefrontal cortex function. *Annual Review of Neuroscience*, 24, 167-202.

Mitchell, M. (2023). How do we know how smart AI systems are? *Science*, 381(6654), adj5957.

Nemitz, P. (2018). Constitutional democracy and technology in the age of artificial intelligence. *Philosophical Transactions of the Royal Society A*, 376(2133), 20180089.

Perez, C. (2002). *Technological Revolutions and Financial Capital: The Dynamics of Bubbles and Golden Ages*. Cheltenham: Edward Elgar Publishing.

Sambasivan, N., Kapania, S., Highfill, H., Akrong, D., Paritosh, P., & Aroyo, L. (2021). "Everyone wants to do the model work, not the data work": Data cascades in high-stakes AI. *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, 1-15.

Sutskever, I. (2023). An observation on generalization. Keynote address, NeurIPS 2023.

Vaswani, A., Shazeer, N., Parmar, N., et al. (2017). Attention is all you need. *Advances in Neural Information Processing Systems*, 30.

Wei, J., Tay, Y., Bommasani, R., et al. (2022). Emergent abilities of large language models. *arXiv preprint arXiv:2206.07682*.

## Industry Reports and Market Research

ARC Prize Foundation (2024). ARC-AGI benchmark results. https://arcprize.org

BCG (2024). From potential to profit with GenAI: A CEO's guide to scaling value. *Boston Consulting Group*.

U.S. Census Bureau (2025). Business Trends and Outlook Survey: Artificial intelligence use by businesses. *U.S. Department of Commerce*.

CBRE (2024). North American data center trends report, H2 2024. *CBRE Research*.

Gartner (2024). Gartner hype cycle for artificial intelligence, 2024. *Gartner, Inc*.

Goldman Sachs (2024). AI, data centers and the coming US power demand surge. *Goldman Sachs Equity Research*.

Investment Company Institute (2024). Investment Company Fact Book: A review of trends and activities in the investment company industry, 64th edition.

JPMorgan Research (2024). AI infrastructure: The next trillion-dollar opportunity. *JPMorgan Chase & Co*.

McKinsey & Company (2024). The state of AI in early 2024: Gen AI adoption spikes and starts to generate value. *McKinsey Global Institute*.

Nvidia (2024). Nvidia Corporation Annual Report, fiscal year 2024. *SEC Filing*.

S&P Global (2025). S&P 500 concentration and performance attribution, Q4 2025. *S&P Global Market Intelligence*.

Semiconductor Industry Association (2024). 2024 State of the U.S. semiconductor industry. *SIA Annual Report*.

The Information (2024). OpenAI's losses nearly doubled to $5 billion in 2024. *The Information*, October 2024.

## Journalism and Public Statements

Brooks, R. (2018, updated annually). Predictions scorecard. https://rodneybrooks.com/predictions-scorecard

Brooks, R. (2024). Interview on AI capability assessment. Various public appearances.

Chollet, F. (2024). Public statements on intelligence measurement and ARC benchmark results. Various platforms.

Engler, A. (2024). The push for federal AI preemption. *Brookings Institution*.

LeCun, Y. (2024). Public statements on LLM limitations and world models. Various platforms.

LeCun, Y. (2026). Remarks at the AI Impact Summit, New Delhi, February 2026.

MIT Technology Review (2026). Yann LeCun's new venture is a contrarian bet against large language models. January 22, 2026. https://www.technologyreview.com/2026/01/22/1131661/yann-lecuns-new-venture-ami-labs/

### Author's Prior Work

Maconochie, J. (2025a). Beyond Scale: A Modular Architecture for Adaptive Artificial Intelligence. Unpublished whitepaper.

Maconochie, J. (2025b). The Evolutionary Processing Unit and the Biological Processing Unit: Lessons from Four Billion Years of Architectural Innovation. Unpublished whitepaper.

Maconochie, J. (2025c). The Architecture of Language: Vanishing Constraints and the Crisis of Shared Meaning. Unpublished whitepaper.

Maconochie, J. (2026a). More instruments, same tune: The growing case against scaling. *Architecture & Attention*, Substack.

Maconochie, J. (2026b). When the music stops: The structural fragility of the AI boom. *Architecture & Attention*, Substack.

Maconochie, J. (2026c). After the music stops: Who bears the cost of the AI correction. *Architecture & Attention*, Substack.

## Appendix: Evidence Map for Key Claims

Several claims in this paper rely on data that is either contested in its measurement, varies by definition, or originates from industry sources rather than peer-reviewed research. In the interest of transparency, the following notes flag the most significant areas where readers should apply their own judgment.

**AI adoption rates (Section 4.3).** Survey measures of enterprise AI adoption range from approximately 10% (U.S. Census Bureau BTOS, using a narrow definition of AI integration) to upward of 50% (private surveys using broader definitions that include any use of generative AI tools). The direction of travel is consistent across measures; the magnitude is not. This paper cites the Census figure as a conservative anchor but acknowledges the definitional uncertainty.

**Market concentration metrics (Section 3.3).** The share of S&P 500 returns attributable to AI-related companies varies by timeframe, index methodology, and how "AI-related" is defined. Figures in financial journalism have ranged from roughly 60% to over 80% depending on the period measured and the companies included. This paper uses "roughly three-quarters" as a directional estimate consistent with multiple credible analyses, not as a precise measurement.

**AI company financial losses (Section 3.2).** Revenue and loss figures for private AI companies, including OpenAI, are based on business journalism reporting rather than audited financial statements. These figures should be treated as informed estimates, not verified actuals.

**Bond issuance projections (Section 3.3).** The JPMorgan estimate of $1.5 trillion in AI-related bond issuances by 2030 is a forward projection, not an observed data point. Projections of this kind carry inherent uncertainty and should be understood as indicative of the scale of anticipated debt, not as a prediction of the actual figure.

**Labor displacement vs. complementarity (Section 4.3).** The academic literature on whether AI adoption produces net job displacement or net job creation is genuinely unsettled. This paper cites studies on both sides and does not claim to resolve the question. The transitional costs of displacement, however, are documented independently of the net employment outcome.

**Energy consumption trends (Section 3.3).** The Goldman Sachs figure on Magnificent Seven energy consumption growth is drawn from equity research, not from company disclosures or government data. It should be treated as an analyst estimate based on available information.