tss thinktank
software solutions

RESEARCH PAPER

The Token Reckoning

Enterprise AI's costliest mistake, and what comes next

THE ARGUMENT IN BRIEF

For eighteen months, enterprises measured AI progress by how many tokens their people could burn, and called it tokenmaxxing. The bills are now arriving, and they are large: exhausted budgets, scrapped leaderboards, and a widening gap between consumption and value. Then, on a Friday in June 2026, a government directive switched a frontier model off worldwide with minutes of notice, and the conversation changed from return on investment to something harder, namely who controls the intelligence an enterprise now runs on. This paper traces the arc from the drug to the reckoning, and sets out a framework for AI that is measurable, governable and resilient.

Greg Strydom
Chief Executive Officer, Think Tank Software Solutions

JUNE 2026

SECTION 01

This Problem Is Not New

Before we name tokenmaxxing, we should be honest about what it is. It is not the first time enterprise technology has cost a fortune and struggled to prove it was worth it. It is the latest instance of a pattern that is now almost forty years old, and the organisations that will navigate it best are the ones that recognise the pattern rather than the novelty.

The frame was set in 1987. The economist Robert Solow, who would win the Nobel Prize that year, looked at two decades of corporate computer investment and the national productivity data that was supposed to reflect it, and found a gap he could not close. "You see the computer age everywhere but in the productivity statistics," he wrote in the New York Review of Books that July. The line endured because it captured something uncomfortable. American multi-factor productivity growth had run at 1.9 percent a year from 1948 to 1973, then collapsed to 0.2 percent afterward, precisely as businesses were pouring money into mainframes and personal computers. The technology was visible on every desk. The return was invisible in the numbers.

The most useful explanation came not from Solow but from the economic historian Paul David, who pointed back to the electric motor. Factories began buying electric motors in the 1890s, yet American manufacturing productivity barely moved for roughly forty years. The motors worked. The problem was that factories had been designed around a single central steam engine, and simply bolting electric motors onto that layout changed nothing. The productivity explosion arrived only once factories were physically redesigned around distributed power. The lesson generalises cleanly: personal efficiency is not systemic productivity. Giving every worker a faster tool does not reorganise the work. Until the organisation rewires itself around the technology, the gains stay trapped at the level of the individual and never reach the balance sheet.

Enterprise software spent the next three decades proving the point. ERP in the 1990s became a byword for budget overruns and abandoned projects, with implementations regularly failing to deliver the return they promised. CRM in the 2000s was worse. Across more than a decade of analyst research, the reported CRM failure rate has sat stubbornly between 30 and 70 percent, with Gartner putting it at 50 percent and Forrester at 47. Hershey is the textbook case: an overly ambitious implementation that triggered severe operational disruption and a drop in its stock price. And yet the same research shows that when these systems were implemented with discipline, the payback was large. Nucleus Research found that well-run CRM returned $8.71 for every dollar spent. The technology was rarely the variable that decided the outcome. The failures clustered around the same root cause every time: organisations treated a change pro-

gramme as a software purchase. When CRM initiatives were treated only as software deployments, they frequently failed.

Digital transformation in the 2010s ran the cycle again under a new name, with the same split between the firms that rewired their processes and the firms that bought the tools and waited for magic that never came.

Which brings us to 2026, and the part that should make every enterprise AI leader sit up. The Solow paradox has returned almost word for word. AI is everywhere in headlines, board papers and capital expenditure plans, and almost nowhere in the productivity data. Recent analysis drawing on Goldman Sachs data found that roughly 70 percent of S&P 500 firms mention AI, but only about 1 percent have quantified its impact on earnings. The computer age was visible in equity markets and balance sheets long before it appeared in the national accounts. AI is following the same script. You can see the AI age everywhere but in the productivity statistics.

This is the frame for everything that follows, and it is worth being clear about what it implies. The token ROI failure is not an AI story. It is a governance story, and governance stories have known solutions. The organisations that eventually extracted real value from ERP, CRM and digital transformation did it by measuring outcomes rather than activity, by redesigning process rather than layering tools onto old habits, and by refusing to confuse motion with progress. That blueprint exists. The reason it cannot simply be lifted and reused is that the AI wave changes two variables the earlier waves never had to reckon with. The cost scales faster, because tokens meter continuously and agentic systems consume them at rates no licence model ever anticipated. And the dependency runs deeper, because the entire stack now rents its intelligence from a handful of providers who can change the price, or switch the model off, with no notice at all. The disease is familiar. The clock is faster, and the exposure is structural.

SECTION 02

The Drug *Why Tokenmaxxing Made Sense*

It is easy to mock tokenmaxxing in hindsight. It is more useful to understand why intelligent organisations adopted it on purpose, because the logic was sound right up to the point where it failed.

Start with the measurement problem. When a new technology arrives faster than anyone can build metrics for it, leaders reach for the proxy they can see. Outcomes from AI are hard to measure and slow to materialise. Usage is trivial to measure and instant. Every dashboard could show tokens consumed, prompts sent, seats active. None could show judgment improved or decisions made better. Faced with a board asking whether the AI

investment was working, executives pointed at the number that moved. If you cannot measure what matters, you measure what is countable, and then you manage to it.

The cultural pressure ran in the same direction. Through 2025, the dominant corporate fear was not overspending on AI. It was being left behind. That fear pushed adoption targets into job descriptions. Meta's chief people officer told employees that AI-driven impact would be a core expectation for 2026. When usage becomes a performance expectation, employees do the rational thing and use more. The incentive was explicit, top-down, and impossible to misread.

The novelty was real, and it mattered. Early frontier models genuinely rewarded exploration. The tenth prompt taught you something the first did not. Satya Nadella, Microsoft's chief executive, described the pull precisely when he admitted to it himself on the Hard Fork podcast in June 2026: it is addictive, the impulse to keep going just to see what the model can do next. That is not a character flaw. It is what a powerful, responsive, general-purpose tool does to a curious professional. For a while, more usage really did mean more learning.

And the perceived cost to the individual was zero. The employee throwing every task at the most expensive model paid nothing. The bill landed on a central budget line, abstracted away from the person generating it. Jensen Huang, chief executive of the chipmaker Nvidia, captured the prevailing attitude at the top of the industry when he suggested that a 500,000-dollar engineer who is not consuming a quarter of a million dollars in tokens should set off an alarm. In that worldview, restraint was the failure mode and consumption was the signal of a serious operator.

Put these four forces together, a measurement vacuum, an adoption mandate, a genuinely rewarding novelty, and a cost that was invisible at the point of use, and token-maxxing was not a mistake anyone made. It was the equilibrium the system was built to produce. The problem was never that people behaved irrationally. The problem was that the incentives were rational and the measurement layer was missing.

SECTION 03

The Addiction *How It Spread and Self-Reinforced*

A behaviour becomes a culture when it acquires its own reinforcement loop. Token-maxxing got two: status inside the firm, and a parallel habit that employees had already formed outside it.

The status loop ran on leaderboards. Once a company ranked people by token consumption, the ranking stopped being a measurement and became a target. This is Goodhart's Law in its purest form: any measure that becomes a target ceases to be a good measure.

Employees did not consume more because they had more to do. They consumed more because the board rewarded consumption, and the board could not tell the difference between work and noise. The metric was gamed precisely because it was prominent, and its prominence came from the absence of any better number to compete on.

The deeper loop was already running before the enterprise formalised it. MIT's NANDA group published a study in 2025 with a finding that should have stopped the industry cold. Despite 30 to 40 billion dollars of enterprise spending on generative AI, 95 percent of organisations were seeing no measurable return. The report called it the GenAI Divide, separating the small minority of deployments that delivered real value from the overwhelming majority that produced pilots and press releases and nothing on the profit and loss statement.

The South Africa-based practitioner Cobus Greyling, writing on the same data, drew out the part that reframes everything. The official enterprise initiatives were stalling, but a shadow AI economy was thriving underneath them. Around 90 percent of employees were already using personal AI tools for work, often paying for their own ChatGPT, Claude or Gemini subscriptions because the consumer models were better than whatever their employer had sanctioned. Ethan Mollick at Wharton named these workers the secret cyborgs. The driver was model-lag: by the time an enterprise procured and approved a model, a far better one was available to any individual for twenty dollars a month.

This is where a distinction matters that the leaderboards erased. The shadow AI economy and tokenmaxxing are not the same phenomenon, and conflating them hides what actually went wrong. The shadow economy is a story about breadth: a marketing manager cleaning up copy, an analyst summarising a report, a lawyer pressure-testing an argument, each quietly using a personal AI subscription to do the job they already had. That usage is cheap, a chat turn costs a fraction of a cent, and it is mostly productive. Tokenmaxxing is a story about intensity, and it lives at the other end of the distribution: the power user, very often a developer, pointing agentic tools at the codebase to generate whole features, run test loops and refactor at scale. That usage is not a fraction of a cent. It is the activity that consumes tokens at tens of times the rate of a single prompt, and it is where the budgets actually detonated.

The connection is that the enterprise inherited both and measured neither. The shadow economy proved the demand was real and the workforce was already fluent. When the company formalised AI usage, added the gamification and the central bill, it did not just sanction the cheap everyday use. It poured fuel on the expensive intensive use, and then celebrated the resulting token counts as evidence of progress. It skipped the one thing the shadow economy never needed and the enterprise desperately did: a measurement layer that tied consumption to outcomes. The result was a workforce that had learned to use AI constantly, inside an organisation that had learned to celebrate that without ever

checking what it produced, or what it cost. At Amazon, the FT reported employees spinning up AI agents to perform meaningless tasks purely to keep their token counts high. The habit had become its own justification.

SECTION 04

The Hangover *What the Data Actually Showed*

The reckoning did not arrive as an argument. It arrived as a set of invoices, and they were specific enough to end the debate.

The emblem of the whole era was Meta's Claudeonomics leaderboard. An employee built it on the company's internal platform, and it ranked more than 85,000 staff by token consumption, handing out titles like Token Legend and Cache Wizard to the heaviest users. The top-ranked individual averaged 281 billion tokens in a single month. Neither Mark Zuckerberg nor CTO Andrew Bosworth made the top 250. Over one 30-day window, tracked consumption climbed from roughly 60 trillion tokens to 73.7 trillion as people competed. Then the dashboard leaked externally and Meta took it down within about two days. Bosworth's verdict, in a follow-up memo, was the epitaph for the strategy: all motion is not progress, and token usage alone is not a measure of impact of any kind. Meta is now replacing the leaderboard with a central monitoring system it calls AI Gateway and imposing formal token budgets. Amazon had already pulled an equivalent leaderboard after watching employees game it the same way.

The budget overruns were worse, because they hit finance rather than culture. Uber's chief operating officer, Andrew Macdonald, disclosed that the company had burned through its entire 2026 AI budget in the first four months of the year. The mechanism was agentic coding tools whose adoption had been deliberately accelerated by an internal leaderboard, and whose cost nobody had metered against output. Uber's own CTO reportedly spent more than 1,200 dollars of tokens in a single two-hour demo. Uber has since capped employees at 1,500 dollars per month per coding tool. ServiceNow reportedly exhausted its annual provider budget within months as well. Salesforce was reported to be facing a 300-million-dollar bill from a single AI provider for the year.

The productivity case, when anyone finally measured it, was thin. Jellyfish, an engineering management platform, found that the heaviest token users were about twice as productive as lighter users, but spent roughly ten times the tokens to get there. Per-developer consumption on its platform rose about 18.6 times in nine months, driven largely by agentic features. Faros AI, studying 20,000 developers, found output rising but bugs and rewrites rising alongside it. The pattern across all of this data is the same: consumption and value are correlated, but weakly and non-linearly, and the heaviest spending produces the worst ratio of cost to outcome.

The macro view confirmed it. Ramp, processing corporate card data across its customer base, found average monthly AI token spend up roughly 13-fold in a year, with costs capable of spiking tenfold overnight when an agent was left running. By May 2026, Fortune had run the headline declaring tokenmaxxing dead, and the Wall Street Journal was reporting that Uber, Microsoft, Meta and Salesforce had begun rationing access to the expensive tools. The era did not end with a strategic insight. It ended with a CFO reading an invoice.

SECTION 05

The Reckoning *What Token Governance Actually Looks Like*

Killing the leaderboards was the easy part. The harder work is replacing a consumption culture with an outcome culture, and the organisations doing it well are moving on three fronts at once.

The first lever is model routing, and it is the fastest win available. Most token spend is wasted not on hard problems but on easy ones sent to expensive models out of habit. Nadella's instruction to Microsoft was a single sentence that doubles as a governance principle: do not use frontier models for non-frontier problems. The implementation he pointed to is Copilot's Auto mode, which selects an appropriate model for each task rather than defaulting every request to the most capable and most expensive one. A summarisation does not need a frontier reasoning model. A routine classification does not need the flagship. Routing the work to the cheapest model that can do it well is the difference between a defensible AI bill and an indefensible one, and it requires no change in what employees are trying to achieve.

The second lever is architecture. Greyling's analysis of emerging agent designs found that single-agent systems with tools reduce token usage by 54 percent compared with multi-agent setups on sequential tasks, because they avoid the overhead of agents passing context back and forth between each other. The caveat matters: single agents hit a sharp capacity ceiling as the task library grows, and the practical path in 2026 is a hybrid of streamlined workflows and modular skills. The general principle holds. A great deal of token spend is structural waste created by over-engineered architectures, and a great deal of it can be designed out before a single prompt is sent.

The third lever, and the one that actually defines the new discipline, is measurement. The question shifts from how many tokens were consumed to what those tokens shipped, resolved or converted. The right metrics already exist inside the business: shipping velocity and merged pull requests for engineering, defect and rework rates for quality, resolution time and first-contact resolution for support, revenue per interaction for sales. None of these is a token count. All of them are outcomes the organisation cared

about before AI arrived. The emerging FinOps discipline is racing to formalise this, with a proposed tokenomics standards body and new metrics like cost-per-intelligence and tokens-per-watt. The underlying move is unglamorous and decisive. You stop paying for motion and start paying for results, and you build the plumbing to tell the two apart. A recent academic study of agentic coding made the case unanswerable: higher token usage did not translate into a higher task-resolution rate. Spending more was not the same as achieving more, and now there is data to prove it.

SECTION 06

The Pricing Problem *A Cost You Cannot Negotiate*

Every cost discussion so far has assumed the price per token is a given. It is worth pausing on who sets that price, because the answer is uncomfortable. The model providers set it, unilaterally, and they can change it whenever model economics or competitive strategy dictates.

This is a different kind of exposure from anything in the traditional software stack. A SaaS licence is a fixed annual contract negotiated between two parties. A token rate is a posted price the buyer accepts or declines, with no negotiating leverage until volumes reach a scale almost no enterprise commands. The closest analogy is not software at all. It is a commodity input like energy: a continuously metered cost, set by the supplier, that the buyer can hedge and optimise around but cannot control.

The headline numbers make this look benign, and that is the trap. Per-token prices have fallen dramatically. Anthropic cut its flagship Opus rate from 15 dollars per million input tokens and 75 dollars output at the Opus 4 and 4.1 generations down to 5 and 25 at Opus 4.6 and later, a threefold reduction while capability improved. A live price war is now underway, with OpenAI reportedly weighing aggressive cuts to stay ahead of Anthropic. One analysis captured the paradox in a single line: token prices fell by most of their value over the period, and enterprise AI bills tripled anyway. Cheaper tokens did not produce cheaper bills, because consumption rose faster than price fell. Deflation per unit is not deflation in aggregate when the units multiply without limit.

That raises the question every procurement leader should be asking. Is the current cheapness permanent, or is it a land-grab? There is a strong case that it is the latter. The leading providers are widely understood to be pricing below the true cost of frontier inference to win market share, and the structural pressure points one way. Anthropic filed confidentially for an IPO this month at a reported 47-billion-dollar revenue run rate and a 965-billion-dollar valuation. Public-market discipline does not reward subsidised pricing indefinitely. The most likely trajectory is continued per-token deflation at the com-

modity tiers, where competition is fierce, alongside margin recovery at the frontier, where switching costs are highest and a small number of providers hold pricing power.

The AWS parallel is the one to keep in mind. Cloud was cheap and frictionless to adopt, and the costs became uncomfortable only once everything ran on it and switching had become expensive. Token providers are reading from the same playbook. The hedges available are real but partial: route aggressively to cheaper models, keep at least two providers live, exploit caching and batch discounts, and design workloads so they can migrate. None of these gives the buyer control over the price. They give the buyer the ability to walk, which in a market with no negotiating leverage is the only power worth building.

SECTION 07

The Agentic Cost Explosion *A Budget Problem Nobody Has Priced*

There is a specific reason the bills outran the budgets, and it is not carelessness. It is that the deployment model changed underneath the budget assumptions. When AI stops answering questions and starts running multi-step work, the relationship between a task and its token cost breaks.

A single prompt is a bounded transaction. An agent is not. It plans, calls tools, reads the results, reflects, retries when it fails, and loops until it decides the task is done. Each of those steps consumes tokens, and the accumulated context of the whole trajectory is resent on every iteration, so input tokens compound. A rule of thumb that an agentic task costs 10 to 50 times a single prompt for equivalent work is a reasonable planning baseline, but the worst cases are far worse. A systematic study of agentic coding published in 2026 found that agentic tasks consumed on the order of a thousand times more tokens than a simple chat exchange, that input tokens rather than output drove the cost, and that two runs of the same task could differ by up to 30 times in total consumption. The cost is not just high. It is high and stochastic, which makes it nearly impossible to forecast at the level of an individual task.

The real-world spread bears this out. On one platform, the median developer consumed around 51 million tokens a month, costing roughly 52 dollars. A power user at the 90th percentile ran nearly eight times that. At the extreme, a single engineer at Vercel reportedly spent around 10,000 dollars of tokens in one day pointing agents at a hard problem, while Uber's CTO spent 1,200 dollars in a two-hour demonstration. Salesforce's Agentforce layer alone processed 3.2 trillion tokens in a single quarter. Most enterprise AI budgets were drawn up when AI meant a chat window, before any of this was the dominant pattern.

The point of modelling this is not precision. It is order-of-magnitude awareness. Consider three representative cases. A compliance-checking agent in banking that reads a transaction, retrieves policy, reasons across multiple regulations, and documents its conclusion is not one prompt; it is a long trajectory run thousands of times a day. A network fault-resolution agent in telecoms that diagnoses, tests hypotheses, and retries is inherently iterative and therefore inherently token-heavy. An inventory-optimisation agent in retail that runs continuously against changing data has no natural stopping point at all. In each case the unit economics that looked trivial in a pilot become material at production scale, and they scale with the volume of business events rather than the number of employees. An enterprise that has not stress-tested its agentic workloads against its budget has not made a forecasting error. It has not yet made the forecast.

SECTION 08

The Sovereignty Problem *When Your AI Gets Switched Off*

Everything to this point is a cost-and-value story. On the evening of 12 June 2026, it became something else.

At 5:21pm Eastern time that Friday, the US government issued an export control directive to Anthropic, citing national security authorities, prohibiting access to its two newest and most capable models, Claude Fable 5 and Claude Mythos 5, by any foreign national anywhere in the world. The prohibition extended to foreign nationals inside the United States, including Anthropic's own non-citizen employees. Because the company could not filter foreign nationals out of its user base in real time, it had no compliant option other than to disable both models for everyone. It did so within hours. Access to Anthropic's other models, including Opus 4.8, was unaffected. The models had launched only 72 hours earlier, on 9 June. The reported trigger was a claimed method of bypassing Fable 5's safety classifier, surfaced shortly after launch through a coordinated multi-agent attack. Commerce Secretary Howard Lutnick's letter to Anthropic's CEO reportedly offered no detailed explanation of the concern.

Two things about this event need to be held at once. The first is that it is genuinely unresolved. Anthropic has publicly stated that it believes the directive rests on a misunderstanding and that it is working to restore access as soon as possible. This may yet be reversed. The second is that the reversibility does not soften the lesson, because the lesson is not about whether Fable 5 comes back. It is that a publicly deployed frontier model, relied on by paying customers, was taken offline worldwide by a government with five minutes of notice and no workaround at the API level. That had never happened before. It can happen again, to any provider, on any Friday.

The implications run far past one company. Map the dependency graph of a typical enterprise AI stack. The frontier model comes from OpenAI, Anthropic or Google. It runs on Microsoft Azure, Google Cloud or Amazon Web Services. The retrieval layer, the embedding models and the vector database that complete the system come from a similarly small cluster of US technology firms. For an enterprise headquartered outside the United States, every one of those layers sits under US law, US export-control authority and US geopolitical priority. The exposure is not to a single vendor that might raise a price or suffer an outage. It is systemic, and it is correlated, because the whole stack answers to the same jurisdiction. June 12 was the first time that abstract dependency produced a concrete, dated, worldwide shutdown.

Then there is the contract, and this is where most buyers will discover they are unprotected. Standard enterprise AI agreements were not written with a government-mandated takedown in mind. Force majeure clauses typically excuse a provider from performance when an event outside its control prevents it, which means such a clause is far more likely to shield the provider from liability for the shutdown than to give the customer any remedy for the disruption it caused. The realistic answer for an enterprise that lost access on June 12 is that its contract offered little recourse and its continuity plan, if it had one, did not contemplate this failure mode. That gap is now a live procurement and legal risk, and it has been sitting in plain sight in contracts that nobody read for this scenario.

SECTION 09

The Vendor Concentration Map *All Roads Lead to Virginia*

The sovereignty risk is not a tail event made vivid by one Friday. It is the predictable consequence of how concentrated the AI supply chain has become, and the concentration is worth stating plainly.

At the frontier, three firms dominate: OpenAI, Anthropic and Google. The cloud infrastructure those models run on is controlled by three more, two of which are the same companies: Microsoft, Google and Amazon. The connective tissue of a production AI system, the embedding models, vector databases, API gateways and orchestration layers, comes from a small cluster of mostly US firms. And the physical footprint is more concentrated than most buyers realise. Northern Virginia is the largest data-centre market on earth, with close to 4,900 megawatts of capacity in 2025, more than twice its nearest global rival and roughly an eighth of all live capacity worldwide. It is also home to AWS's oldest and busiest region, the one whose periodic outages have a habit of taking large parts of the internet down with them. The popular claim that 70 percent of global internet traffic passes through the area is disputed and probably overstated, but the underlying point survives the correction: a remarkable share of the world's cloud and AI work-

loads runs through a single cluster of counties outside Washington. For a bank in Lagos, a telco in Nairobi or an insurer in Johannesburg, the intelligence layer of the business can depend, at several points in the stack, on infrastructure governed by one country's law. There is no close precedent for this degree of concentration in the history of enterprise technology.

The obvious hedge is open-weight models, and here the news is better than it was a year ago. The capability gap has largely closed for standard work. By mid-2026, GLM-5.1 was matching Claude Opus on coding benchmarks, Qwen 3.7 was competitive with GPT-5.5 on reasoning, and DeepSeek V4 Pro was tying the closed frontier on SWE-Bench Verified at around 80 percent. These are not curiosities. They are production-grade models an enterprise can self-host, keeping data and inference inside its own jurisdiction and removing the remote-shutdown risk entirely. Licensing has loosened too, with Mistral, Qwen and several others now under permissive Apache 2.0 or MIT terms.

But sovereignty through open weights comes with two honest caveats. The first is operational: the strongest open models are very large, often needing multiple high-end GPUs to run, so self-hosting trades a vendor-dependency problem for an infrastructure-and-talent problem that not every African enterprise is positioned to take on. The second is geopolitical, and it is the one rarely said out loud. Most of the leading open-weight models, Qwen, DeepSeek, GLM and Kimi, are Chinese in origin. Diversifying away from dependence on US providers can mean adopting models developed under a different government's influence, which is a different risk rather than the absence of one. The genuinely non-aligned options at the open frontier are narrower: Mistral from France, and Meta's Llama from the US, the latter carrying its own usage-cap and regional licence restrictions. The realistic posture for most enterprises is not pure independence but deliberate optionality: a hybrid stack that keeps frontier closed models for the work that needs them, runs open models for sensitive or high-volume workloads, and is architected so that no single provider, and no single government, can take the whole thing down at once.

SECTION 10

The Cognitive Offloading Risk *What AI Is Doing to Human Judgment*

The risks so far are measurable in money and uptime. The last one is slower, quieter, and potentially more expensive, because it erodes the asset that does not appear on any balance sheet: the judgment of the workforce.

The mechanism is called cognitive offloading, and the evidence for it is no longer speculative. A 2025 study by Michael Gerlich, surveying 666 participants across age and edu-

cation groups, found a negative correlation between heavy AI tool use and critical-thinking ability, mediated by the habit of routing problems through AI rather than reasoning through them. He described the result as a kind of cognitive laziness, a declining inclination to engage in deep, reflective thought. The effect was strongest among the youngest users, those aged 17 to 25, and higher education appeared to offer some protection. More striking still, researchers at the MIT Media Lab ran an EEG study of essay writing and found that participants using a language model showed lower cognitive engagement while working, and when later asked to perform the same task unaided, performed measurably worse than those who had never used AI assistance. They named the phenomenon cognitive debt. A 2026 conceptual review described the underlying dynamic as a delegation feedback loop: the more fluently AI solves a problem, the more readily the user hands the next one over, and the user's own capacity atrophies through disuse.

The organisational version of this risk is precise. If junior analysts, lawyers, compliance officers and engineers stop developing independent reasoning because AI handles the first draft of every judgment, the capability base of the organisation degrades in a way that is slow, invisible and expensive to reverse. The danger is sharpest exactly where it matters most: in the high-stakes, novel situations where there is no precedent to retrieve and human judgment is the legal and operational backstop. A workforce that has only ever supervised AI outputs has not been trained to produce them.

And here the two halves of this paper meet. The cognitive-offloading risk and the sovereignty risk are the same risk viewed over different timescales. Ask the uncomfortable question that 12 June posed in miniature: what happens to an organisation whose analysts have quietly stopped thinking for themselves when the model goes dark on a Friday evening with no warning? The capability you offloaded is the capability you no longer have when the tool is withdrawn. The mitigation is not to use less AI. It is to use it deliberately, preserving the human reasoning loop on purpose. The research points the way: guided, scaffolded use that keeps the human doing the hard cognitive work, with AI as a check rather than a substitute, preserves critical thinking where unstructured reliance erodes it. Judgment, like the muscle it resembles, is kept by being used.

SECTION 11

The Path Forward *Architecting for Value, Resilience and Sovereignty*

The argument of this paper is not against enterprise AI. It is for enterprise AI done with discipline, and the discipline now has a recognisable shape. Four pillars hold it up.

The first is value governance. Token volume is retired as a primary metric and replaced with outcome-linked measurement drawn from the business the organisation was already running: shipping velocity, defect and resolution rates, revenue per interaction,

time to close. The governing question changes from how much AI did we use to what did the AI help us ship, resolve or earn. This is the same move that eventually separated the ERP and CRM winners from the casualties, and it works for the same reason: it makes motion and progress distinguishable, so the budget can follow the second and starve the first.

The second is cost architecture. Model routing becomes default rather than exception, so frontier capability is reserved for frontier problems. Agentic workloads are explicitly cost-modelled before deployment, not discovered on the invoice. Budgets are stress-tested against the stochastic, compounding cost behaviour of agents rather than the tidy economics of a chat window. The organisation treats tokens the way a manufacturer treats a volatile input cost: metered, hedged, and continuously optimised.

The third is sovereignty resilience. No critical workflow depends on a single provider that a single jurisdiction can switch off. The stack is built for optionality: at least two frontier providers kept live, open-weight models in production for sensitive and high-volume work, contracts renegotiated to address service continuity and government-action scenarios rather than assuming force majeure cuts only one way, and deployment spread across more than one cloud where the workload justifies it. The goal is not to predict the next 12 June. It is to make the next one survivable.

The fourth is human capability. AI productivity is pursued without quietly hollowing out the workforce's ability to reason. Offloading is made deliberate rather than ambient, with the human reasoning loop preserved in the roles and decisions where judgment is the point. The organisation builds AI use that keeps people sharp rather than dependent, because the capability it offloads carelessly is the capability it will not have in the moment it most needs it.

None of these four is exotic. Each is the AI-era expression of a governance lesson the enterprise has already learned the hard way at least once. The organisations that internalise them do not use less AI than their competitors. They use it knowing what it costs, what it produces, and what happens when it is taken away.

SECTION 12

The Implementation Partner in the Value-First Era

A shift of this kind does not only change what enterprises buy. It changes what they need from the people who help them buy it, and the implementation partner's role is being redefined in the process.

Until now, the work that has paid the bills for AI implementation partners has been access and deployment. Get the tools in. Train the users. Integrate the APIs. Stand up the

pilot. That work was valuable while AI was scarce and unfamiliar, and for a lot of partners it still is. But it is commoditising fast, because the tools are getting easier to deploy and the deployment has stopped being the hard part. A partner whose whole value proposition is getting AI switched on is selling a service whose price is falling toward zero.

The work that is coming is everything this paper has described, and it is harder to commoditise precisely because it is judgment rather than installation. It is governance design: building the measurement layer that ties token spend to business outcomes. It is model selection strategy: deciding which workloads route to which models, when an open-weight model in the customer's own environment beats a frontier API, and how to keep optionality alive. It is cost architecture: modelling agentic workloads before they reach production and stress-testing budgets against them. It is sovereignty risk management: mapping the dependency graph, renegotiating the continuity terms, and designing a stack that survives a provider being switched off. None of that is deployment. All of it sits closer to the decisions a board actually loses sleep over.

So the real distinction is not between an old kind of partner and a new one. It is between the partner who answers how do we use this and the partner who answers should we, where, at what cost, and what is our exposure if it disappears. The first is a vendor, and the vendor's margin is under pressure already. The second is an adviser, and that work is only becoming more valuable as the cost and sovereignty risks this paper has traced come into focus. The market has not finished making that shift, which is exactly why the position is still open. The partners who get there first will be the ones whose customers skip the waste phase and architect for value, resilience and sovereignty from the start.

REFERENCES

Sources

Robert Solow, "We'd Better Watch Out," New York Review of Books, 12 July 1987 (productivity paradox).

Paul David, research on the electric motor and productivity lag.

Artorius Wealth Management, "The Solow Productivity Paradox" (March 2026), citing Goldman Sachs data on S&P 500 AI disclosure.

Johnny Grow, CRM Failure Report (2025); Nucleus Research CRM ROI study; Rand Group on ERP failure rates.

Satya Nadella, live taping of NYT "Hard Fork" podcast (June 2026), reported by Business Insider, Benzinga, Windows Central and others ("I'm a tokenmaxxer too, it's addictive"; "Don't use frontier models for non-frontier problems").

MIT NANDA, "The GenAI Divide: State of AI in Business 2025" (July 2025); Campus Technology summary.

Cobus Greyling, "The Shadow AI Economy" (September 2025) and "The Hidden Shadow Economies of AI" (January 2026); "AI Agent Architectures" (January 2026, single-agent 54% finding).

The Information, reporting on Meta's Claudeonomics leaderboard; Fortune (9 April 2026); MLQ News and Edgen on Meta's AI Gateway and Bosworth memo.

Uber: COO Andrew Macdonald on 2026 AI budget; Shopifreaks, The D[AI]LY Brief and others on Uber, ServiceNow and the Salesforce $300M provider bill.

Jellyfish and Faros AI developer productivity studies, via TechCrunch, “The token bill comes due” (5 June 2026).

Ramp Builders Blog, “Building a Unified Pipeline for AI Token Spend” (April 2026); The Next Web on token-price deflation versus rising bills; FinOps Foundation tokenomics standards effort.

arXiv 2604.22750, “How Do AI Agents Spend Your Money?” (agentic token consumption study).

Anthropic API pricing (Opus, Sonnet, Haiku); OpenTools and Finout on the OpenAI/Anthropic price war and Opus price cuts; Anthropic confidential IPO reporting (Quartz, WSJ).

Anthropic statement on the Fable 5 and Mythos 5 suspension (12 June 2026); Fortune, Al Jazeera, Quartz, Snyk, MarkTechPost and DevToolLab on the export-control directive.

Open-weight model benchmarks and licensing (2026): Codersera, AceCloud, ComputingForGeeks, BuildFastWithAI (GLM-5.1, Qwen, DeepSeek V4 Pro, Mistral, Llama).

Gerlich (2025), Societies 15(1), on cognitive offloading and critical thinking; Kosmyna et al. (2025), MIT Media Lab EEG study (“cognitive debt”); Kim et al. (2026), delegation feedback loop.