ASILO
Asset Management

Asilo Argo Special Investment Fund · Portfolio Managers' Quarterly Letter Q2 / 2026

Q2 2026: From Subsidy to Scarcity

INTRODUCTION

In April, the chief technology officer of one of the world's largest consumer-technology companies admitted — almost in passing — that the firm had exhausted its entire annual artificial-intelligence budget in four months. Not because the projects had failed. Because they had worked. The agents were let loose, they produced, and they consumed. A line meant to last twelve months lasted one third of that.

We keep returning to that admission, because it is the most honest description we have of where the AI economy actually arrived this quarter.

The constraint moved.

For two years, the binding constraint on artificial intelligence lived in the world of ideas — better architectures, better training recipes, better ways to wire models into work. The cost of running these systems was, to a first approximation, free, because the laboratories were paying it for us. Inference was sold below cost to win users, in the time-honoured way of every platform that wanted to own a market before it charged for it. We have spent a decade learning to recognise that pattern; it is the pattern that built the consumer internet.

That era ended in the second quarter. Quietly, without a single defining headline, the constraint moved out of the world of ideas and into the world of physics — into compute, into power, and finally into silicon itself. The cost of intelligence stopped being something a balance sheet could absorb and started being something an economy has to ration. We think this shift — from the subsidy era to the scarcity era — is the most important development of the first half of 2026, and the one most likely to be misread by the people whose job is to price it. So we will spend most of this letter on it.

THE JEVONS THESIS

The intuition that a falling price reduces revenue is so deeply embedded in economic thinking that it passes as common sense. For most resources, it is. When the price of milk falls by half, the dairy farmer earns roughly half as much — partially cushioned by the marginal extra consumption that lower prices encourage, but not remotely reversed by it.

There is, however, a narrower category of resources to which this rule does not apply. William Stanley Jevons identified it in the nineteenth century studying coal: when the price of certain resources falls, total consumption rises so explosively that producer revenues increase rather than decline. The mechanism is not mysterious. A sufficiently large price reduction makes the resource economically viable in applications that were previously impractical, opening entirely new markets that dwarf the original ones. Electricity is the clearest modern example. The real price of electricity in the United States fell by approximately 80% between 1913 and 2024; over that same period, the inflation-adjusted consumer electricity market grew by a factor of more than thirty. The explanation is not that households grew extravagant with their lights: falling prices made electricity useful in ways that simply did not exist before — first as an upper-class novelty, then as a household utility, then as the substrate of an entire technological civilisation. The average Western home now contains sixty to seventy electric devices. None of that demand could have been imagined from the vantage point of a 1930s utility analyst.

The reason this matters now is that artificial intelligence has the structural characteristics of a resource to which the Jevons paradox applies — and, beyond that, of a General Purpose Technology: the specific category of innovation that reshapes entire economies rather than merely one sector. The steam engine, electricity, the internal combustion engine, and the internet each met the three defining criteria. Each was pervasive across industries. Each improved continuously. And each spawned complementary innovations — new business models, new products, new ways of organising work — that multiplied its original impact by orders of magnitude. AI meets all three. It is already embedded across healthcare, finance, manufacturing, transport, entertainment, agriculture, and education — in any domain, essentially, where data is generated or decisions must be made. Its capabilities are improving rapidly enough that this year's cutting edge is next year's baseline. And it acts as a platform that raises the productivity of innovation itself, accelerating scientific discovery, automating routine cognition, and generating entirely new product categories that in turn consume more AI.

The full productivity impact of General Purpose Technologies tends to arrive with a lag, as organisations develop the complementary skills, business models, and infrastructure needed to put them to work. Electricity took a generation to transform factory organisation; the productivity gains from electrification arrived most powerfully long after the wires were strung. We expect AI to diffuse faster — the software layer shortens many of the complementary lags — but the lesson from previous GPTs is not to mistake the early phase for the whole story.

The Jevons lens and the GPT lens converge on the same conclusion. As the cost of intelligence falls, the market for intelligence does not contract — it expands, in ways and at a pace that are structurally impossible to read from the installed base. That conclusion shapes everything that follows in this letter.

A CAPITAL BUBBLE IS NOT A COMPUTE BUBBLE

The bears have a tidy story, and it deserves a fair hearing before we explain why it is misreading the signs. The story is that the hyperscalers are building a 1999. Capital expenditure among the

largest platforms is running near $805 billion this year, up from an earlier $765 billion, with next year's figure lifted toward $1.1 trillion. To anyone who lived through the last great infrastructure overbuild — fibre, that time — these are the numbers that come right before the reckoning. Build it, the argument goes, and they will not come.

The error in the story, once you look closely, is that it quietly confuses two different things. A capital bubble is a financing event; a compute bubble is a physical one. Money shows up fast — it always has; capital is the easy part. A physical overbuild is a different matter: to overbuild compute, every single link in the physical supply chain would have to be loose at the exact same moment.

- **Silicon and chips.** Advanced foundry allocations, set years in advance.
- **Infrastructure.** Power grids, electrical substations, and the specialised cooling that data centres depend on.
- **Logistics and labour.** The physical buildings, and the specialised people required to build and run them.

Because each link has a distinct lead time — dictated by physics and local permitting rather than by financial term sheets — a simultaneous overbuild across all of them is structurally blocked. That is not what is happening.

Here is the giveaway. In a financing bubble, the order book empties as the speculative money burns off. Instead, the order book is filling. The largest platforms spent roughly half a trillion dollars over the past year, while their reported backlog — capacity that customers have already committed to buy — sits near $1.3 trillion. The gap between what is being built and what has already been ordered is not closing. It is widening. You do not get that shape from a bubble. You get it from a shortage.

WATTS AND WAFERS

Follow the chain to its narrowest point and you arrive, as you always do, at the foundry. At its annual shareholders' meeting in early June, TSMC's chief executive, C.C. Wei, was unusually plain: demand is running so far ahead of capacity that the company can only do so much, and it is working hard not to become the bottleneck for the whole industry. He was describing a shortage the company now expects to last for years. The American fabs that were meant to relieve it have slipped — permitting, a shortage of construction labour, the ordinary friction of building very hard things in places that have forgotten how to build quickly.

The Jevons logic applies here as forcefully as it does to AI demand itself. Every gain in model efficiency, every cheaper token, has not eased the pressure on the supply chain; it has multiplied the work thrown at it and pushed the bottleneck one layer further upstream — from the accelerators, to the high-bandwidth memory beside them, to DRAM, and down into the unglamorous passive components, the multilayer ceramic capacitors that every AI server swallows by the tens of thousands. This is where a good deal of our own research has gone this quarter, and where we do not believe the 2027 demand picture has yet reached consensus.

The cleanest way to hold the whole problem in one hand is to reduce the next phase of AI to two physical inputs: watts and wafers. Watts are the time-bound problem — the power shortage begins to ease through 2027 and 2028 as new generation comes online. Wafers are not. Foundry

capacity has no comparable relief valve on the timescale that matters. Watts can be added; wafers, mostly, cannot.

FEAR VS. EVIDENCE

While market analysts and commentators spent the quarter relitigating whether AI was overbuilt, we did what we did during last autumn's depreciation panic: we listened to what the people closest to the constraint were actually saying.

- **The model laboratories are behaving like sellers of a scarce good.** OpenAI told customers plainly the world would be "capacity constrained for some time," and began letting enterprises pay for guaranteed access to compute — not the conduct of a company chasing a soft market.
- **Even vast new supply barely moves the needle.** A frontier laboratory secured more than 200,000 GPUs through an unusual deal with a launch company — and the sharpest reaction was that, against its training and inference needs, even that is a drop in the ocean.
- **Token strategy has reached the C-suite.** Enterprise-software leaders now describe it as the single most heated subject among Fortune 500 CIOs and CFOs, with no one confident they have the answer.
- **The foundry is rationing aloud.** TSMC's own chief executive warned the shortage will run for years and that the company "can only support so much."
- **Value investors are arriving on the constraint.** A thirty-year veteran generalist called the semiconductor, capital-equipment and hyperscaler complex the most attractive sector available — and put the bulk of his capital there — on valuation, not narrative.

Last autumn the evidence cut against the bubble. This year it cuts against the glut. The drivers we identified have not merely held — they have tightened, quarter by quarter.

WHERE THE MISUNDERSTANDING LIVES

Regular readers know we do not chase mispricing. A mispricing is a gap between price and a value everyone can already see; it closes fast and it is competed away. We hunt misunderstanding — a gap between price and a reality the market has filed under the wrong heading, which can persist for years because closing it requires the consensus to change its mind about what it is looking at.

The misunderstanding here is the one we began with: the market keeps reading the abundance of capital as evidence of an abundance of compute, and so keeps pricing the build-out as a financing bubble waiting to pop. The same watts-and-wafers lens exposes an absurdity that captures it exactly. Semiconductor-capital-equipment companies trade at roughly forty times next quarter's annualised earnings; memory companies trade in the single digits. Both cannot be right. They describe two incompatible futures for one supply chain. That is not noise — it is the market openly disagreeing with itself about whether the constraint is durable.

Our exposure sits on the answer to that question, and it sits upstream. We own the bottleneck, not the application. Our positions in the memory and accelerator complex, in the foundry's most binding input layers, and in the passive components that scale one-for-one with server units are

not a wager on which model wins. They are a wager that intelligence — whoever makes it, however it is made — must pass through a handful of physical chokepoints that capital alone cannot conjure. We do not need the bubble debate settled. We need the constraint to be real and durable. The quarter says it is both.

THE DEMAND IS STRUCTURAL, NOT SPECULATIVE

One fair objection remains: perhaps the backlog is theatre — a few well-funded players booking capacity from one another in a circle. We take it seriously; the most thoughtful bear of the quarter reached past the dot-com era to the holding-company structures of 1920s electrification, with their layered financing and cash recycled between opaque entities. It is the right historical worry, and we keep it in view.

But we think the demand is real, and the reason is the shift from assisted to agentic AI. When a person prompts a model and waits, consumption is bounded by human patience. When that same person instead sets agents to work, consumption is bounded only by the task — and tasks are effectively unbounded. That is why a year's budget vanished in four months. It is why the organisations furthest ahead are no longer buying tools but rebuilding their operating models around fleets of agents, wrestling with questions as basic as how many agents one human can supervise. Demand of that shape is not an artificial result of financing. It is a new unit of economic activity, and it pulls tokens through the system in volumes the subsidy era never had to price.

A quieter consequence of the move from subsidy to scarcity is already visible in how intelligence is sold. In the subsidy era, a flat monthly subscription bought something close to frontier capability, because the laboratories were paying the inference bill themselves to win the market. As compute is rationed and finally priced, a fixed monthly price can only buy a rationed quality of intelligence — smaller models, smaller reasoning budgets, tighter usage limits — while the full capability migrates to usage-based enterprise contracts that pay for every token consumed. The flat-rate subscriber increasingly receives the cheaper, weaker intelligence; the enterprise willing to pay per token receives the frontier. You do not tier a good you hold in abundance. The tiering is itself evidence that the constraint is real — and token consumption, the meter on that contract, is becoming the cleanest measure we have of how much work is actually being done.

And the same shift does something more consequential than tier quality: it lifts the ceiling on revenue. A flat monthly subscription caps what a customer can pay no matter how hard they work the model; usage-based pricing removes that cap, so revenue scales with consumption rather than with seats. As agentic workloads turn a single user into a fleet of token-hungry agents, the monetisable surface expands by orders of magnitude — and because that demand lands on a supply chain that cannot flex, the incremental pricing power accrues to whoever owns the bottleneck. The subsidy era socialised the cost of intelligence; the scarcity era prices it, and the price flows to the constraint.

FROM SCARCITY TO SOVEREIGNTY: THE US–EUROPE DIVERGENCE

This quarter we give the regulatory-divergence thesis its own treatment, and the scarcity transition is precisely what makes it urgent. In a physically constrained world, the first-order question — what can be built — is quickly followed by a second-order one that begins to dominate returns: who is permitted to build it, and where the capital, the models and the operating companies actually sit. The answer, on every axis we can measure, is increasingly 'America'.

It is worth being precise about why this is not just another relative-value call. Most cross-Atlantic equity decisions are made in the language of cycles — currency, energy mix, the ECB versus the Fed, valuation gaps — and rest on a single premise: that the rules of the game are stable, and only the variables move. That premise is now wrong in one specific place, AI regulation, and because AI is the largest single source of equity value being created this decade, getting it wrong is the kind of mistake that makes a beautifully calibrated regional-rotation model look, a few years out, like a 1971 currency model fitted to a regime that no longer exists.

The new rulebook is being written in opposite directions on the two sides of the Atlantic. Washington has decided to treat AI as critical infrastructure, and two American legal frameworks make that posture coherent: Section 230 (which protects platforms from liability for the user content they carry), and a series of recent permissive fair-use copyright rulings covering the training of large language models. Europe is moving the other way:

- **Fines that reach 6% of global revenue.** The Digital Services Act's enforcement structure — already exercised in a €120 million action against X in December 2025 — is, on the administration's reading, a template more than a one-off.
- **Liability for what a model can reproduce.** A Munich court has already held a model provider liable for copyrighted lyrics its model had memorised, and the new Product Liability Directive — in national law by December 2026 — extends the EU's strict-liability regime to software and AI.
- **Copyright that runs against fair use.** The European direction of travel on training data runs counter to the American rulings, raising the cost of building a frontier model lawfully on the continent.

You do not need a political view on any of this to take an investment view. The investment view is simple: if these regimes harden, the cost of operating an AI business under European jurisdiction rises and the cost of operating one under US jurisdiction falls on a relative basis — not because of growth or rates, but because the rulebook itself is being rewritten beneath them.

This is where the scarcity letter and the divergence letter become the same letter. If token consumption is the meter on the cost of intelligence, it is also the cleanest available measure of how much cognitive work an economy is actually doing — and, increasingly, of how much it expects to do. When agents are set loose to 'try things', every attempt spends tokens; rising consumption is therefore a signal of both the volume of work and the effectivity an organisation expects to get back from it. On that measure the divergence is already stark. The United States, with roughly 340 million people, generates more than twice the LLM tokens of the entire European continent and its 740 million — with Germany, the largest single European source, generating only mid-single-digit shares of total routed usage. The rulebook is not merely shaping margins; it is shaping who does the work.

A frontier model is not a business. It becomes one through a go-to-market machine: capital to build capacity, enterprise channels to deploy it, and power to run it. In this cycle, that machine is the moat — because for the first time in software's history, distribution is capital-intensive. You do not reach the customer with a download link; you reach them with a data centre. And every link of that chain is, today, overwhelmingly American — with one Dutch exception we return to below. The five largest US hyperscalers guided to a combined $725 billion of 2026 capex, the majority pointed at AI infrastructure.

Behind them stand private-capital partners — Advent, Bain, Brookfield, TPG, Blackstone, Hellman & Friedman and Permira — that together manage roughly $2.7 trillion in assets, now

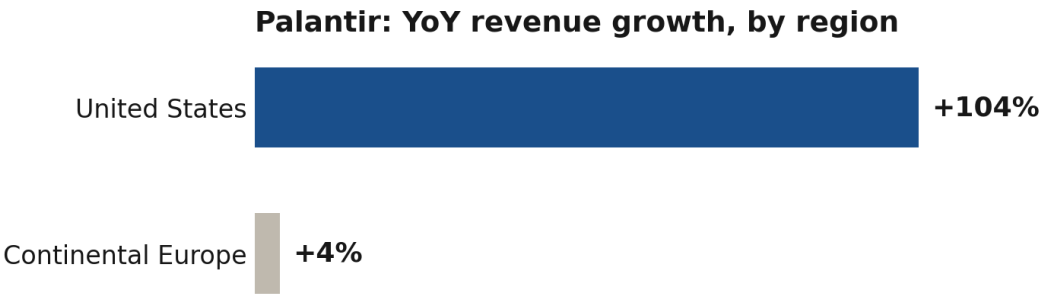
being pointed toward AI deployment through joint ventures with the frontier labs. A Bezos-led industrial-AI vehicle is reportedly raising around $100 billion. None of this is research spending. It is route-to-market spending — the machinery that turns model weights into invoices. Buy the factory, deploy the model — that is the business model. American capital, deployed into American models, sold through American channels, running on American power.

Europe is not blind to this; it is responding in the only register available to it. When Emmanuel Macron champions Mistral AI, or Ursula von der Leyen boasts of Europe's world-class supercomputing infrastructure, they are engaged in a kind of national prestige-building — proof that Europe is 'in the race'. Unable to match the sheer scale of American Big Tech, European leaders reach instead for the instruments they do control: public funding, a claim to open-source superiority, and regulatory frameworks — each repurposed as a vehicle for a unique, top-tier international status. It is a coherent strategy. But the gap it is meant to paper over is not a matter of narrative, and on the one number that matters most — revenue actually being earned — Europe's gem is not even close.

Source: Epoch AI tracker and company disclosures (Feb–Apr 2026). Annualised revenue run-rate.

And the gap is far more likely to widen than to close, because the industry now turns on a flywheel. Revenue becomes capital, capital buys compute, and compute trains the next, better model — which earns more revenue still. Scaling laws give that loop a winner-take-most character: the lab that can spend the most on compute tends to hold the best model, which pulls in the most revenue to spend again. From inside that dynamic it is hard to see how Mistral closes the distance. The know-how may well be there; the resources to compete with OpenAI and Anthropic round after round are not. Prestige, public funding and open-source positioning are real assets, but they do not fund a flywheel at that scale — and a portfolio cannot be paid in prestige.

None of this would matter much if the divergence were not already showing up in the income statement. It is. The clearest live demonstration is Palantir, where management has been unusually candid about which side of the divergence each of its dollars sits on: Continental Europe described, in successive calls, as 'sluggish' and growing in the low single digits, while the US business achieved triple-digit growth — 'Revenue in our US business grew 104% year-on-year,' management noted this quarter. The same product, sold by the same sales force into broadly the same enterprise budgets, growing at roughly twenty-five times the pace on one side of the Atlantic as the other, for two years running, is no longer a sales-cycle effect. It is the rulebook expressing itself in the income statement.



Most recent disclosure (Continental Europe Q4 2024; United States Q1 2026, +104%).

Source: Palantir Q4 2024 / Q1 2025 / Q1 2026 earnings calls (Quartr). Periods are not strictly comparable; the chart illustrates the magnitude of the regional dispersion management has repeatedly chosen to disclose.

None of this is a blanket 'short Europe' call. The West remains interdependent — there is no leading-edge US compute without ASML — and a 2025 statute can soften by the time it is enforced in 2027. Our point is narrower: the divergence is real, and it is best expressed at the level of individual securities rather than through a country bet.

OUTLOOK: THE SCRAMBLE, AND THE WEATHER AROUND IT

We expect the rest of 2026 to look much like its first half: periodic panics as old frameworks are forced onto a new reality, each one an opportunity for whoever has done the homework. The scramble for compute is now visible at the very top — frontier laboratories doing unusual deals for silicon, and behaving less like product companies than like utilities rationing a scarce supply. A quieter conversation has begun, too, about who ought to own all this: proposals for an AI token tax, for handing the public equity in AI companies, a renewed argument about AI and inequality. These are early and unserious in their specifics, but they are the predictable political weather that forms once a few physical chokepoints come to mediate a large share of output — against a backdrop of stubborn inflation and a rate-cutting cycle that has stalled, with markets now pricing hikes in Europe. We are watching it. We are not yet positioned for it.

OUR JOB

Our job is not to chase every small mispricing. It is to find the structural misunderstandings — the places where the market has the right facts filed under the wrong heading — because that is where the greatest alpha hides. This quarter the market is reading two such headings wrong at once. It reads the abundance of capital as an abundance of compute, and files a shortage under 'bubble'. And it reads a change in the rules of the game as a cyclical rotation between regions, and files a regime change under 'mean reversion'. We think the right headings are 'shortage' and 'regime change', and we expect the coming quarters to be unusually generous to investors who can tell the difference.

Thank you for your trust,

Henri Blomster & Ernst Grönblom