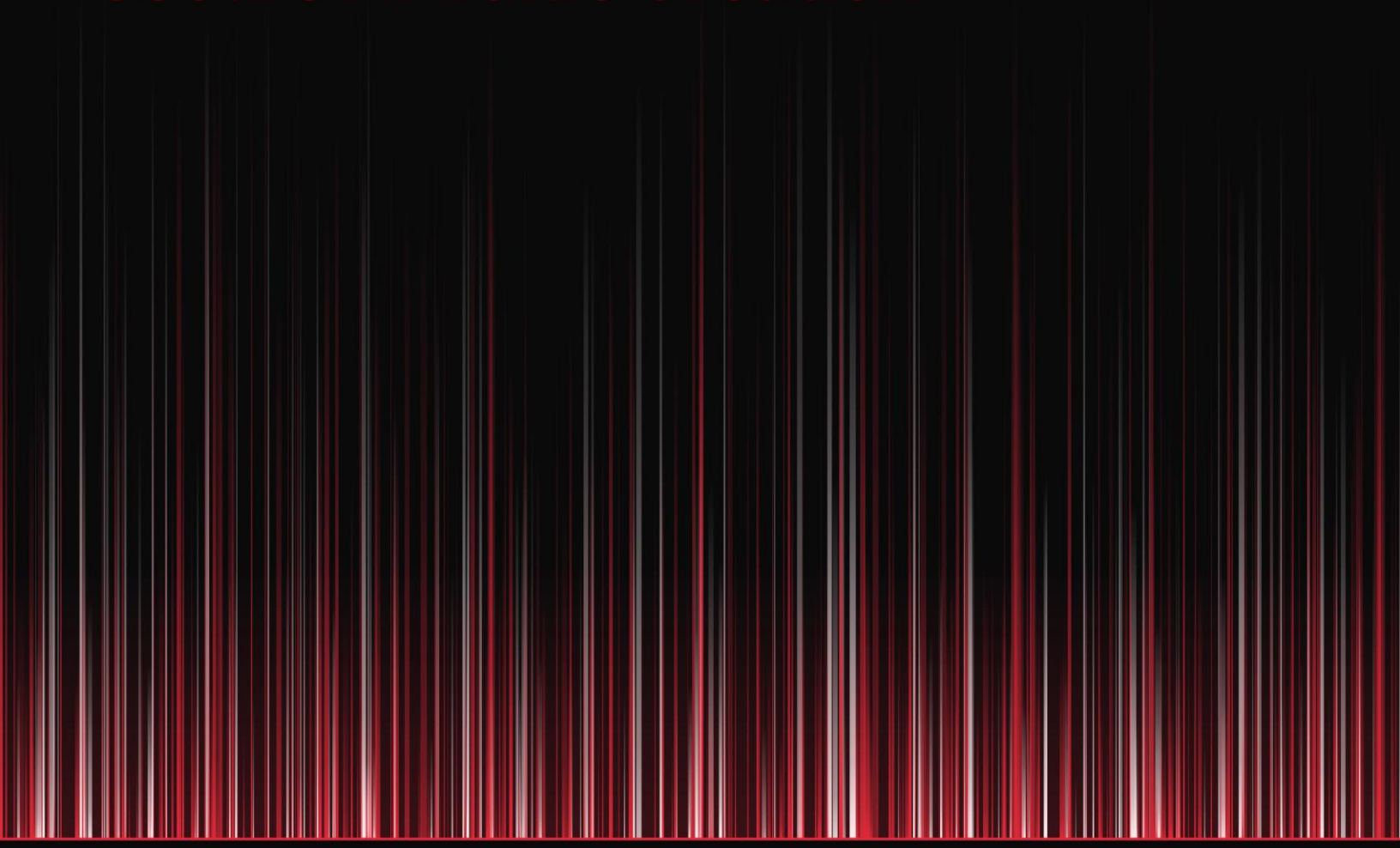
DataKrypto

WHITE PAPER

From Adoption to Advantage

The convergence behind **Secure AI Value Creation**



By **Angelo Fienga and Paolo Campoli**

CONTENTS

1 Structure, not enthusiasm, separates the winners 3

2 Success creates a new asset, your Capital Intelligence, and it is exposed 5

Where the asset lives decides whether you can defend it 6

3 Measure first: the hidden bill of tokens and agents 7

4 A structured model strategy: routing and on-premises 8

Can a fine-tuned model match a frontier model on enterprise work? 9

5 Protecting the new intelligence from the inside 10

Closing the loop: three elements, one asset 11

© 2026 DataKrypto. All rights reserved.

From Adoption to Advantage

The Convergence Behind Secure AI Value Creation

By Angelo Fienga and Paolo Campoli

IN BRIEF

The enterprises that win with AI aren't simply automating tasks, they're encoding how the business thinks into agentic systems. That encoded intelligence becomes an asset only when it's chosen deliberately, made economical to run, and protected, including while it's in use.

Two years of enterprise experimentation have made one thing clear: adopting AI and capturing value from it are different problems. The organizations that turn a portfolio of pilots into durable advantage get three things right at once, and it is how these three work together that separates the leaders from the stuck.

1. **Disciplined selection.** Leaders choose use cases that move a real business metric and secure the data those use cases depend on.
2. **Capital Intelligence.** As companies put AI to work, they create new frameworks, redesigned processes and operating logic encoded in their systems. Together these form an emerging proprietary asset, and one that can be defended only once it is recognized.
3. **Economic literacy.** Leaders understand, measure and optimize what AI costs to run. The token has become the central unit of consumption, but it is only part of the bill.

These three do not form a checklist to be worked in sequence, they converge, and out of that convergence comes the asset this paper is really about. Disciplined AI adoption and economic optimization together convert an enterprise's operational knowledge into Capital Intelligence, the encoded operating logic that becomes its newest strategic asset. That Capital Intelligence is infused in models, data and workflows that are expensive to run and increasingly exposed, and the economics decide whether the advantage is sustainable and where the asset physically lives. Read together, they point to a fourth conclusion most enterprises have not yet internalized: the more successful this system becomes, the more valuable and the more vulnerable the intelligence at its center, so protecting it while it is in use becomes central to keeping the advantage. What follows traces that logic in order: where AI succeeds, what that success creates, what it costs, and how to keep it.

© 2026 DataKrypto. All rights reserved.

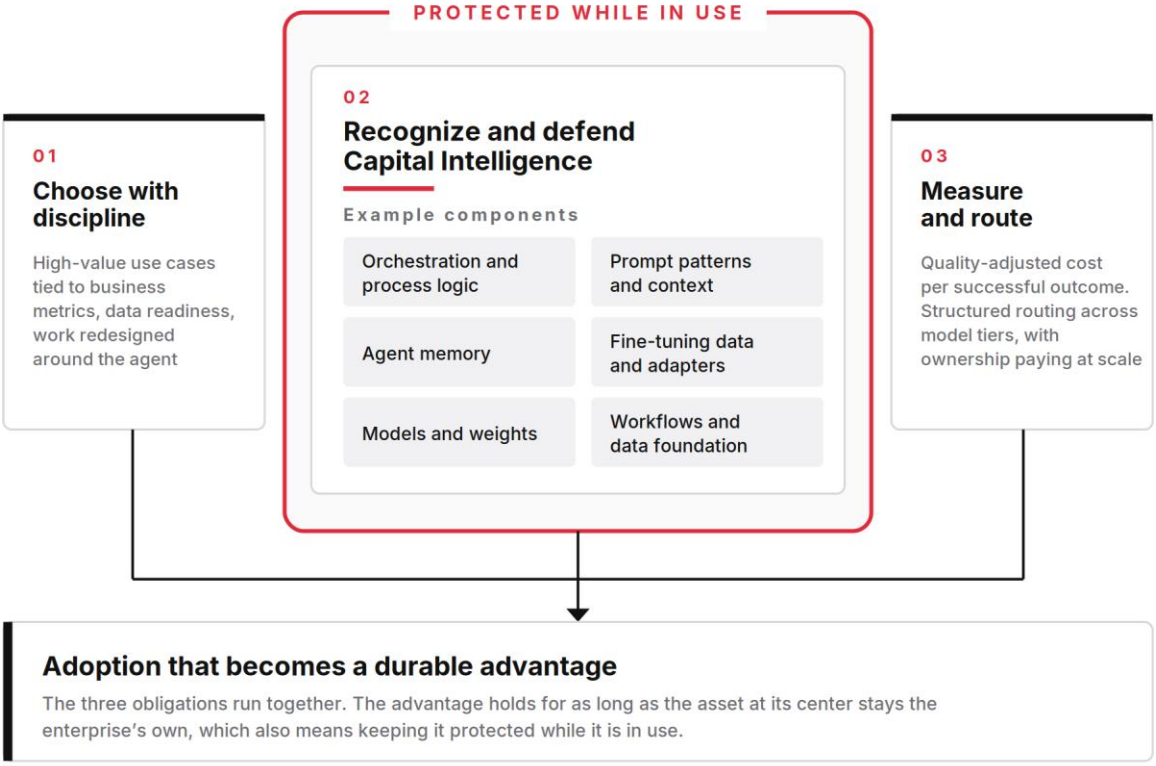


FIGURE 1 An organization's advantage rests on three things: choosing well, measuring honestly, and protecting what we call Capital Intelligence. The first two are practices. The third is an asset, and when it's lost, the advantage goes with it.

1 Structure, not enthusiasm, separates the winners

The evidence of the last two years is consistent: broad adoption of generative AI has not, by itself, produced broad financial return. Deloitte's State of AI in the Enterprise: The untapped edge frames the gap as a move from ambition to activation. The share of leaders reporting a genuinely transformative effect from AI has risen to 25 percent, more than double the 12 percent of a year earlier. Yet only about 20 percent of organizations say they are already growing revenue through AI while 74 percent still describe revenue growth as an aspiration. Access to AI has expanded far faster than the ability to turn that access into organization-wide impact.

McKinsey's work sharpens the same picture. In Seizing the Agentic AI Advantage, it finds that nearly eight in ten companies use generative AI in at least one function while roughly the same share report no material impact on earnings, the pattern it calls the "gen AI paradox." It estimates that about 90 percent of higher-impact, function-specific use cases never leave the pilot stage. Deloitte quantifies the wall those pilots hit: in Scaling Agentic AI to Realize Business Value, produced with Google Cloud, only about a quarter of

© 2026 DataKrypto. All rights reserved.

organizations have moved 40 percent or more of their experiments into production. Gartner adds the necessary counterweight, projecting that more than 40 percent of agentic AI projects will be canceled by the end of 2027 because of escalating costs, unclear business value or inadequate risk controls. The constraint is not enthusiasm, it is operational discipline.

> **40%+**
>
> of agentic AI projects will be canceled by the end of 2027 because of escalating costs, unclear business value or inadequate risk controls.
>
> GARTNER

Success, in other words, is not random, and maturity is what distinguishes it. Deloitte's multicounty survey of 3,235 IT and business leaders across 24 countries, reported in Agentic AI is scaling faster than guardrails, found that only 21 percent of organizations have a mature governance model for agentic AI even as 74 percent expect to be using agents at least moderately by 2027. The organizations succeeding with agentic AI, it notes, take a measured approach, beginning with lower-risk use cases, building governance capabilities, and scaling deliberately. The State of AI report reinforces the point from the value side, observing that enterprises where senior leadership actively shapes AI governance realize significantly greater business value than those that delegate the work to technical teams alone. Structure, not experimentation, is the differentiator.

What the structured minority does in practice is recognizable. They select a small number of high-value, function-specific use cases tied to core business metrics, where the data is structured and a credible path to return exists even when the payoff is not immediate, and they redesign the workflow around the agent rather than bolting the agent onto an unchanged process. Deloitte describes this as an AI-native approach that redesigns work holistically instead of layering AI onto legacy processes, a shift that also depends on data readiness, since legacy data and infrastructure cannot power real-time, autonomous AI. McKinsey's Building the Foundations for Agentic AI at Scale reaches the same conclusion from the constraint side: eight in ten enterprises cite data limitations as the primary roadblock to scaling agents.

Where does this land first? The most-cited use cases cluster around procurement, research and development, the supply chain, marketing and sales, and customer service. The proof points are concrete rather than aspirational: McKinsey documents a retail bank that rebuilt its credit-memo process to pull data from more than ten systems and generate first drafts, reporting turnaround roughly 30 percent faster and analyst productivity gains

© 2026 DataKrypto. All rights reserved.

of up to 60 percent. When an enterprise does this well, it does not merely cut cost. It creates something new.

2 Success creates a new asset, your Capital Intelligence, and it is exposed

A redesigned, agent-driven process can become a defensible asset in its own right. The asset is not the model alone and not the data alone, it is the encoded operational logic of how the business is improved by AI and agents, the sequence of decisions, exceptions, tool calls and judgments that used to live in the heads of a company's best people. Andreessen Horowitz makes the point sharply in Is Software Losing Its Head?: agents may erode old moats such as user habit, but they make operational logic and context more important, because an agent needs explicit rules, permissions and process definitions to act, and that institutional logic is the hardest thing for a competitor to reconstruct. McKinsey frames the same asset as an opportunity: firms can encapsulate internal expertise, legal reasoning, tax interpretation and procurement best practice directly into agents. Call this encoded know-how the enterprise's **Capital Intelligence**, the strategic asset that emerges when AI converts operational knowledge into encoded, executable form. It is a construct rather than a settled legal or accounting category, but a competitive asset that may contain proprietary workflows, fine-tuning data, model weights, prompt patterns and operational know-how, and parts of it may qualify as trade secrets where they derive value from being secret and are kept so, though that protection is never automatic. Like any asset, it can be protected or surrendered.

CAPITAL INTELLIGENCE

EXAMPLE COMPONENT	PRINCIPAL EXPOSURE
Orchestration and process logic Agent definitions, tool calls, exception handling, escalation rules	Built from templates every customer of the platform has
Prompt patterns and context System prompts, retrieval context, reasoning traces	In cleartext at the moment of inference
Agent memory Persistent state carried across sessions and between agents	Readable by anyone with privileged platform access
Fine-tuning data and adapters Curated training sets, LoRA and QLoRA adapters	Trade-secret value gone the moment it is copied
Models and weights Base and customized model artifacts, intellectual property in themselves	Model theft; architecture recoverable by side channel
Workflows and data foundation Redesigned processes and authorization-aware retrieval	Hard to extract, inspect or redeploy once embedded

FIGURE 2 Examples of what Capital Intelligence can contain, and the exposure each one carries. Capital Intelligence is a construct, so these components are illustrative and the list stays open.

© 2026 DataKrypto. All rights reserved.

Where the asset lives decides whether you can defend it

Building on a vendor's platform

The first risk arises when the innovation is built inside a vendor's proprietary agent platform such as SAP Joule, Salesforce Agentforce or ServiceNow Now Assist. Here precision matters: the popular claim that "your data trains their model" does not necessarily hold for the leading vendors. SAP, for example, states that it contractually prohibits its LLM providers from training any model on customer data. The real exposure is more structural, and it has three parts.

Shallow differentiation. The building blocks used to innovate (the agent templates, prebuilt skills and orchestration patterns) are, by design, available to every customer of that platform, so the differentiation created with them is shallow and shared rather than proprietary, and often easy to replicate.

Know-how you can't protect. The enterprise may not control the inference path, the memory architecture or the orchestration loop, which means its encoded know-how lives inside infrastructure it cannot instrument or protect with its own technical controls. As Waehner (2026) argues, competitive advantage in agentic AI increasingly depends on what the enterprise builds with AI rather than on what a vendor builds for it.

Harder to leave. Lock-in increases: the more process innovation is embedded inside a vendor platform, the harder it becomes to extract, inspect, protect or redeploy elsewhere. A platform may behave impeccably with customer data and still leave the enterprise's Capital Intelligence less portable and harder to defend.

When you own the pipeline

The second risk applies even when the enterprise owns the entire pipeline: the model and its weights are now intellectual property in their own right. The OWASP GenAI Security Project recognizes model theft (the unauthorized access, copying or exfiltration of proprietary models, weights and parameters) as a direct threat to intellectual property and competitive advantage, with attack paths spanning infrastructure compromise, insider leakage and API extraction. The exposure now reaches the hardware layer: research presented as "ModelSpy" at NDSS 2026 demonstrated that the electromagnetic emanations of a GPU during inference can be used to reconstruct a neural network's architecture from several meters away. The result is narrow. It recovers architecture under specific experimental conditions rather than weights or prompts, but the direction of travel is clear. Protecting access to the building is no longer the same as protecting the asset inside it.

© 2026 DataKrypto. All rights reserved.

3 Measure first: the hidden bill of tokens and agents

The next discipline is economic. Deloitte's The pivot to tokenomics argues that traditional total-cost-of-ownership frameworks miss the reality of AI, because AI capabilities run on tokens (the units of text, image or audio that models process in training, inference and reasoning) and every interaction is metered in real time. The token, in this framing, is an increasingly important unit of consumption and an economic signal, though not a substitute for a full view of total cost of ownership. It can be governed, but only if token consumption is treated as a managed budget rather than a byproduct of experimentation, and the metric that ultimately matters is not cost per token but quality-adjusted cost per successful business outcome. Without that instrumentation, every discussion about model choice, routing or on-premises inference is guesswork.

Once the meter exists, a counter-intuitive dynamic becomes visible. Deloitte reads AI economics through a Jevons-like rebound, in which efficiency gains tend to be outpaced by rising total consumption rather than producing net savings. Average inference cost is projected to fall from roughly $0.04 per million tokens in 2025 to about $0.01 by 2030. And yet enterprise spending keeps surging, because agentic systems and multi-agent workflows expand token demand faster than efficiency gains can offset it. The effective cost per task can stay stubbornly high: teams default to premium frontier models, agents consume more tokens through multi-step reasoning and error correction, and usage expands the moment people discover what agents can do. Consumption is also extremely skewed. Bain, in The Future of Opex in the Agent Economy, reports early data suggesting the top 5 percent of users can consume more tokens than the remaining 95 percent combined, concentrating cost in exactly the people an organization can least afford to throttle.

> **5%**
>
> of users can consume more tokens than the remaining 95 percent combined, concentrating cost in exactly the people an organization can least afford to throttle.
>
> BAIN & COMPANY

In 2026 this turned from theory into headlines.

- The Verge reported that Microsoft began canceling most internal Claude Code licenses across its Experiences and Devices division (including teams around Windows, Microsoft 365, Outlook, Teams and Surface), with a June 30, 2026 deadline, steering employees toward GitHub Copilot CLI.

© 2026 DataKrypto. All rights reserved.

- Business Insider reported that Amazon shut down an internal token-usage leaderboard called KiroRank after concerns that it encouraged excessive consumption rather than meaningful business outcomes.

- Forbes and Fortune reported Uber burned through its 2026 AI budget in four months while struggling to connect Claude Code usage to measurable consumer-facing innovation. What is evident is the budget model can fail when usage is not metered per task, per workflow and per outcome.

4 A structured model strategy: routing and on-premises

Once the meter exists, the largest lever is to stop running everything on frontier models. The most vivid public example comes from AT&T. Facing roughly eight billion tokens a day, its team rebuilt the orchestration layer so that large "super agents" route tasks to smaller, domain-specific worker models instead of pushing everything through frontier reasoning models. The company reported up to 90 percent cost savings while processing far more tokens than before (on the order of 45 billion a day, more than a fivefold increase) from a deliberate re-architecting of its orchestration layer rather than a single tweak. Structured routing between frontier cloud models, smaller specialist models and self-hosted models is therefore not a technical nicety but one of the largest levers in the AI operating-expense equation.

> **90%**
>
> cost savings reported by AT&T after routing tasks from frontier reasoning models to smaller, domain-specific worker models, while processing more than five times the tokens.
>
> VENTUREBEAT

The supply side has matured to support this. Technavio put the open-source LLM market, a label that in practice spans the open-weight models enterprises can self-host, at roughly $21 billion in 2025, growing at about a 34 percent compound annual rate through 2030. On-premises was the largest deployment segment, explicitly driven by demand for data sovereignty and cost control, and with on-premises inference estimated to cut operating expenditure by up to a third versus token-based APIs. The adjacent fine-tuning market is smaller but growing quickly. Fortune Business Insights estimates LLM fine-tuning services at $1.9 billion in 2025 and projects $9.0 billion by 2034, an 18.7 percent compound annual rate. These are commercial market estimates with limited methodological transparency, useful as directional signals of a maturing supply side rather than as proof that on-premises is cheaper for any particular enterprise. The driver is straightforward: enterprises want models that understand their domain, run at lower cost, and keep sensitive data

© 2026 DataKrypto. All rights reserved.

closer to the business. Parameter-efficient methods such as LoRA and QLoRA have lowered the customization barrier, though they do not remove the need for evaluation, governance and secure operations.

Can a fine-tuned model match a frontier model on enterprise work?

On bounded, well-defined tasks, increasingly yes. On open-ended, high-stakes reasoning, not consistently. A 2026 survey of dynamic model routing and cascading argues that smaller models can be sufficient for routine queries while complex tasks still require more capable ones, and that well-designed routing can balance performance, cost and specialization across model families, sometimes outperforming even the single most capable model (Moslem and Kelleher, arXiv). Reuters likewise highlighted emerging research suggesting that small local models can match or outperform larger ones on many average tasks while still falling short on the hardest. A concrete datapoint shows how large the gain can be on a narrow task: a base Qwen3-8B model scored 41 percent accuracy, while the same model with a lightweight LoRA fine-tuning adapter nearly doubled that to 78 percent, in an example reported as drawn from Stanford research infrastructure. It is a single reported experiment on a specific task, not evidence that an 8-billion-parameter model can replace a frontier one in general, but it shows how large the gain can be where the task is well bounded. The conclusion is routing, not wholesale replacement: use frontier models where they are economically justified, and route bounded, repetitive, domain-specific work to smaller or fine-tuned models when they meet the quality bar.

Crucially, the appeal of open-weight and on-premises models is not only price. It is sovereignty and control: sensitive data and prompts can stay inside the enterprise boundary rather than moving through external APIs and foreign-hosted infrastructure, a requirement that grows more important in regulated and critical-infrastructure settings. But on-premises is not free. As Deloitte's tokenomics analysis stresses, it shifts cost from API operating expense to capital expenditure, energy, GPU operations, model-lifecycle management, evaluation, security and talent, and its total-cost-of-ownership modeling shows that ownership overtakes consumption only at scale, where a dedicated "AI factory" can see more than a 90 percent drop in cost per billion tokens between its first and third years. The right question is therefore not "cloud or on-premises?" but which tasks justify frontier models, which justify smaller or self-hosted ones, and what the fully loaded cost per task is for each.

There is a further consequence that leads directly to the final question. Moving inference on-premises reduces provider exposure and can reduce token cost, but it does not protect intellectual property by itself, it relocates the exposure. The model weights, fine-tuning data, prompt patterns, reasoning traces, retrieval context and persistent agent memory

© 2026 DataKrypto. All rights reserved.

now sit inside infrastructure operated by people with privileged access: database administrators, virtual-machine and container-platform operators, backup and observability teams, and, in hybrid deployments, cloud infrastructure providers. Transport-layer and at-rest encryption protect data in transit and at rest, but not while it is being used, and if any part of the pipeline still runs on an external platform, the platform-control risk described earlier returns. In other words, the more successful the structured strategy becomes, the more concentrated, valuable and exposed the enterprise's Capital Intelligence is. On-premises can lower the API bill and improve sovereignty, but unless the asset is protected while in use, the company has simply moved the horse into a different barn.

5 Protecting the new intelligence from the inside

Protecting Capital Intelligence is not a single move. Most of the work is done by controls enterprises already know, identity and privileged-access management, network segmentation and isolation, key management, data minimization, authorization-aware retrieval, and the contractual and trade-secret protections that govern what a vendor may do with the asset. Confidential computing based on trusted execution environments adds hardware-backed isolation at runtime. Each of these is necessary, and none of them closes the same gap. The one exposure they leave open is the cleartext moment, when data, prompts and model must be decrypted to be processed, and it is that residual gap, rather than the whole of security, that in-use protection is meant to address.

This is the exposure that DataKrypto approaches through what it describes as Confidential Agentic AI: protecting the AI itself, together with the data and the prompts that move through it, from the inside rather than only around it. The principle is simple to state and hard to achieve. Extend encryption to data while it's in use, not only while it's in transit and at rest, so that sensitive information stays protected even at the moment it's being processed. Applied to an agentic architecture, this closes the cleartext gaps that perimeter controls, network encryption and runtime isolation leave open across the four principal exposure surfaces: model inference, the retrieval and knowledge layer, persistent agent memory, and agent-to-agent communication.

The same principle protects the fine-tuned model that encodes the enterprise's Capital Intelligence. An enterprise can therefore pursue the cost advantage of on-premises, open-weight inference and the differentiation of models trained on its own data without surrendering that asset to privileged insiders or external platforms, with protection becoming a property of the infrastructure rather than a rewrite of the application. The honest claim is not "zero risk" (no serious security posture should promise that) but something more defensible: closing the cleartext exposure that perimeter controls, transport encryption, disk encryption and runtime isolation leave open.

© 2026 DataKrypto. All rights reserved.

Closing the loop: three elements, one asset

The three elements introduced at the outset are, finally, one story. Choosing the right use cases and the data to support them is what produces value, the Capital Intelligence that AI converts operational knowledge into is the new asset a company must recognize and defend, and the economics of tokens determine whether that asset is sustainable and where it lives.

Each element intensifies the others, concentrating an enterprise's hardest-won advantage (its process logic, routing decisions, fine-tuning data, model weights, memory and prompts) into a single, valuable and increasingly exposed core.

Selecting use cases with discipline, recognizing and building Capital Intelligence, and measuring cost while routing models all remain necessary. But none of it compounds into lasting advantage on its own. The intelligence an enterprise encodes into its agents is an advantage only for as long as it stays its own, and it stays its own only if the asset itself is protected while it is being used.

© 2026 DataKrypto. All rights reserved.

ABOUT

About the authors

Angelo Fienga is a Senior Business Advisor at Deloitte Italy focused on AI, sustainability and cloud. He brings more than 30 years of experience in technology strategy, innovation and business transformation, including senior roles at Cisco, where he led Sustainable Solutions across EMEAR. His career has spanned AI, cloud, IoT, telecommunications and digital infrastructure, with previous experience at Ericsson and Qualcomm. Fienga holds a degree in Electronic Engineering and a Master's in Data Intelligence and Decision Science from Sapienza University of Rome.

Paolo Campoli is Senior Vice President and Chief Growth Officer at DataKrypto, where he leads global growth, strategic partnerships and enterprise adoption of Confidential AI. He previously spent 26 years at Cisco, most recently as Vice President and Global Head of the Service Provider Sector, leading strategy across one of the company's largest business segments. Earlier, as CTO of Cisco's Service Provider business in EMEAR, he helped drive industry adoption of 5G, automation and software-defined networking. Campoli has more than three decades of experience across telecommunications, cloud and digital infrastructure and holds a doctorate in Electronics and Telecommunications from Politecnico di Milano.

About DataKrypto

DataKrypto reimagines security for artificial intelligence through encrypted execution. FHEnom for AI™ combines Fully Homomorphic Encryption (FHE) with Trusted Execution Environments (TEEs) to protect AI models, data, prompts and intermediate information throughout AI execution, including during GPU processing.

This paper is provided for informational purposes only and does not constitute legal, financial or security advice. Third-party research, market estimates and news reports are cited as published; figures are directional and have not been independently verified by DataKrypto. All trademarks are the property of their respective owners.

© 2026 DataKrypto. All rights reserved.