XAMUN POSITION PAPER NO. 7

THE WORKFORCE.

The GPU is in Dubai. Where is the decision?

Sovereign compute is necessary and not sufficient. What an autonomous workforce still has to answer once the data centre is certified — where each number came from, which rule applied, who signed, and how much autonomy that decision has earned.

Arup Maity, Founder and CEO, Xamun Technologies

Position paper · Version 1.0 · September 2026

SUMMARY OF THE ARGUMENT

The UAE's infrastructure operators have made the case for sovereign compute, and made it well: an agentic economy cannot run on foreign data centres, uncertified clouds, or fragmented local hardware. That argument is now settled at the national level. This paper begins where it stops.

A certified data centre answers one question — *where did the work run?* It does not answer the four an auditor, a regulator, or a board will ask next: where did this number come from, which rule applied and on what date, who decided and why, and how much autonomy did this decision earn. Those are questions about the **decision**, not the infrastructure, and no certificate on the building settles them.

The paper sets out what it takes to answer them: a roster of agents each serving one named role on a schedule; a fixed order of two minds in which code computes and decides and a language model only drafts; rules cited to their instrument as of the date of the event; a single place where a person is interrupted; autonomy that is earned per decision and capped for irreversible work; and a ledger in which hours returned are measured against a declared baseline rather than claimed. It closes with the questions a buyer should put to any vendor of an "autonomous workforce", including this one.

Model choice — open-weight on sovereign hardware or frontier behind a residency boundary — turns out to be a consequence of this architecture rather than a feature of it. When the model never holds a number and never makes a decision, it becomes the one component that can be swapped without touching the record.

IN THIS PAPER

1 The push is real, and the infrastructure case has been made
2 Three sovereignties, and which one a data centre supplies
3 A roster, not a chatbot
4 Two minds, in a fixed order
5 The rule is cited, as of the date
6 The fork: the only place a person is interrupted
7 Autonomy is earned per decision and paced by reversibility
8 Hours returned are measured, not claimed
9 The model is a component
10 What a workforce reports upward
11 Six questions for any vendor of an autonomous workforce

1 The push is real, and the infrastructure case has been made

In September 2026 du Tech published a white paper on what it calls the sovereign agentic infrastructure: a 1GW, liquid-cooled AI park; Oracle Cloud run in-country under a sovereign-operator model; high-density compute nodes; a national hypercloud certified by the UAE Cyber Security Council; and a hybrid framework that routes lighter agent workloads to public edge nodes and anchors regulated processing on private infrastructure. It followed a June memorandum with an orchestration partner, signed in the presence of the Council's head, to run sovereign GPU orchestration on that hypercloud. Earlier in the year the Government of Dubai set a direction for private-sector adoption of agentic AI, and the federal target of half of government services delivered through AI agents by 2028 stands behind both.

This paper takes that infrastructure argument as correct. A workforce of agents that reasons over confidential records, transacts on behalf of an institution, and runs continuously cannot depend on compute that sits outside the jurisdiction, on a cloud the national security authority has not examined, or on hardware that a supply shock can take away. Papers earlier in this series argued that the value in enterprise AI concentrates at the point where the workflow itself is rebuilt around governed intelligence; that rebuild needs a floor to stand on, and the floor is now being poured. Sovereign compute is a national good, and the operators building it are doing the country a service.

It is also, on its own, an answer to a question nobody in a bank, a hospital, or a ministry will be asked. When a regulator opens the file, the question is not *where did this run?* It is *where did this number come from, under which rule, and who decided?* Those questions belong to a different layer. The infrastructure paper defines its subject clearly — the compute spine — and leaves that layer open. This paper is about that layer.

> *A certificate on the building says where the work ran. It says nothing about what was decided, under which rule, or by whom.*

2 Three sovereignties, and which one a data centre supplies

The word *sovereign* is doing three different jobs in the current conversation, and the three are routinely collapsed into one. It helps to separate them.

Sovereignty	The question it answers	What supplies it
Infrastructure	Where does the work run, and who can reach it?	In-country data centres, certified clouds, national networks
Model	Whose weights read the organisation's data, and where do they execute?	Open-weight models on sovereign hardware; frontier models behind a residency boundary
Decision	Where did this number come from, which rule applied on that date, who signed, and	The architecture of the agent itself: what computes, what drafts, what records, and who is interrupted

Sovereignty	The question it answers	What supplies it
	how much autonomy has this decision earned?	

The first is the subject of the infrastructure operators, and they are right that it comes first. The second is a live question for every regulated buyer and is addressed in Section 9. The third is the one this series has been circling since its first paper, and it is the one an autonomous workforce cannot delegate to its host.

Consider the definition of an agent the infrastructure paper offers: instead of executing pre-programmed code, an autonomous agent uses a deep reasoning model to break down abstract goals, optimise its own tasks, and collaborate with other agents. That is a fair description of what most agent frameworks now do. It is also a precise description of the model being the decision-maker. If the model decomposes the goal, chooses the steps, and executes them, then the model produced the number in the report, the model chose which rule to apply, and the model decided the case was ready. Certify the data centre as heavily as you like; the provenance of that decision is still a language model's output, and *trust me* is still not a control.

The use cases the infrastructure paper describes make the point sharper, because they are the right use cases. Municipal payments processed by agents. Visa allocations and corporate licences completed within minutes. Compliance certifications issued without a person in the loop. These are exactly the routine, high-volume, rule-heavy tasks where a governed workforce returns the most hours. They are also customer-facing, regulator-facing, and in several cases irreversible. Whether an agent may do them at all is not the question. The question is how fast it earns the right to do them unattended, and what evidence that right rests on. That is decision sovereignty, and it is the subject of the rest of this paper.

3 A roster, not a chatbot

The first architectural choice is the one most vendors get wrong by default, because it is the one the tooling makes easiest: the chat box. A chat assistant presents a blank field, waits for a request, and answers it. It is a general-purpose surface, which is exactly why nobody can say in advance what it will be asked to do or what it will produce. A general-purpose surface cannot be governed, only monitored after the fact.

A governed workforce has no box to type into, by design. It is a roster. Each agent serves one named role — sales analyst, business analyst, compliance analyst, vendor manager — and does one task for that role: the daily sales pack, the weekly management pack, the KYC readiness check, the vendor SLA reconciliation. Each runs on a schedule or when a record arrives, through the same gate, and each run produces a row. The organisation decides what the agent does when it is configured, not when it is prompted.

This sounds like a limitation and is the opposite. It is what makes the other properties in this paper possible. Because the task is fixed, the calculation can be written in code and tested. Because the role is named, the person who signs the output is known before the run starts.

Because the trigger is a schedule or an event, nobody can ask the agent to do something outside its remit, and there is no free-text instruction to audit. And because every run is a row, the question *what did this agent do last Tuesday?* has a single, complete answer.

What a row contains

- The data the run loaded, and from which table or extract.
- Every figure the run computed, and the calculation that produced it.
- Every rule the run applied, with its instrument, article, and the version of the rule pack in force.
- Whether a language model was called, what it was shown, and what it returned.
- Which validators ran on the draft, and which passed or failed.
- What a person decided at any fork, in what role, with what reason.
- What the run returned in hours, against which declared baseline.

A chat assistant produces none of this. An RPA script produces the first two and hard-codes the third. The row is the unit of decision sovereignty; everything else in this paper is a rule about what may write to it.

4 Two minds, in a fixed order

Papers One and Two in this series introduced the architecture this series calls two minds: a deterministic mind that computes and decides, and a language model that drafts, over one shared memory. This paper describes how that order is enforced inside a workforce, because *enforced* is the operative word. The order is not a prompt instruction. It is a property of the code and of the tests that gate the build.

The deterministic mind runs first, and may run alone

Code loads the data. Code computes every figure the output will contain — the gross, the mix, the variance against the trailing mean, the overrun against the contract clause. Code resolves which rules are in force on the date being judged. And code decides whether the run may proceed at all: if a blocking rule fires, the run stops at the gate and the model is never called. A KYC case with a screening hit does not get a beautifully drafted readiness memo. It gets a row that says which rule blocked, and a fork for a compliance reviewer.

The model runs second, and sees no values

When the run does proceed to a draft, the language model is shown the *names* of the figures, never their values. It writes prose around placeholders. Every number in the final output is substituted from the calculation record; a figure the model invents, or a placeholder it fails to use, fails the run. The model cannot round, cannot estimate, cannot adjust, because it never had the number to begin with. This is the mechanical answer to the hallucination problem that Paper One reframed: hallucination is a feature of a language model, and it is never allowed in the ledger because the ledger is never in the model's hands.

No free text from a person ever reaches the model

The third rule is the one buyers find hardest to believe and vendors find hardest to keep. No free-text input from a person — not a question, not a note, not the reason a reviewer gave at a fork — is ever passed to a language model. The reason is recorded for the audit and stops there. This closes the injection surface that every chat-shaped product leaves open, and it is enforced by a guard test that fails the build if any code path passes a person's text to a model. The property is not promised; it is tested.

Validators sit between the draft and any reader

Before anyone sees a draft, a set of deterministic validators checks it: that every placeholder was substituted; that mandatory disclosures are present; that no personal data appears where it should not; that a bilingual output is at parity across languages; that the wording obeys the rules in force for that document type. A validator failure opens a fork. It does not reach an inbox.

> *The model is shown the names of the figures, never their values. A number it invents fails the run. That is not a prompt. It is a test.*

5 The rule is cited, as of the date

A rule that lives in a script is a rule nobody can inspect. When a compliance reviewer asks *why did this block?*, the answer must name the instrument, the article, and the version of the rule pack that was in force, and it must say whether the rule was resolved from a live regulatory memory or from an offline copy. A governed agent hard-codes nothing. It asks a regulatory knowledge base which obligations apply to a record of this type, in this jurisdiction, on this date, and it writes the answer — with citations — into the row.

As-of discipline

The phrase *on this date* matters more than it looks. Regulation changes. A case submitted in June is judged by June's instrument even if the run happens in September, and the row says so. A workforce that applies today's rule to last quarter's record produces a decision that is wrong on its face and indefensible in front of a regulator. As-of resolution is a property of the knowledge base, not a feature the agent can be told to remember.

Obligations block; placeholders warn

Not every rule in a regulatory memory is settled. Some entries are curated obligations, tied to a published instrument and reviewed by a person; some are placeholders awaiting curation, marked as such. The distinction is visible on every screen and enforced in the gate: a curated obligation may block a run; a placeholder may only warn. A workforce that lets an unreviewed rule stop a payment, or lets a reviewed one merely warn, has confused its own confidence with the law's.

No personal data leaves for the lookup

The regulatory memory receives amounts, flags, and risk tiers. It never receives names, identifiers, or nationality. The rule is resolved against the shape of the record, not its subject. This is one of the few places where infrastructure sovereignty and decision sovereignty touch directly: the knowledge base can sit anywhere the organisation's residency policy allows, because nothing that identifies a customer ever crosses to it.

6 The fork: the only place a person is interrupted

The infrastructure paper describes agents that complete transactions without human intervention, and treats that as the goal. This series has argued since Paper Two that the goal is subtler: a person should be interrupted at judgement forks and nowhere else. Too many interruptions and the workforce has returned no hours. Too few and the organisation has delegated its judgement to a model. The fork is the design that holds the line.

A fork opens in four circumstances only: a rule blocks; a required fact is missing; a draft fails a validator; or a decision is held for sign-off because its autonomy stage requires it. A fork never asks a question. It names the thing — the rule's own sentence, the record it concerns, the citation beneath — and it offers a choice. Two forks from one agent are never the same card twice, because each is bound to its own record and rule.

Routing by role, refused by the server

Who may close a fork is not a matter of who is available. A fork raised by a statutory, regulatory, or Sharia obligation belongs to a compliance reviewer, whoever the pack was addressed to. Everything else is signed by the role its decision names. The server refuses any other signatory. This is the point at which the organisation chart becomes a control: the person who signs is the person whose role the rule names, and the row records that it was so.

A decision is a record

At a fork, a person chooses, gives a reason, and the choice is recorded. Nothing is preselected, so silence is not consent. Nothing is undone, so a decision is not a draft. Approving a pack sends it. The reason is kept for the audit and, as Section 4 said, never reaches a model. The result is a decision trail in which every human judgement has a name, a role, a timestamp, a reason, and the rule it concerned — the record a regulator asks for, existing before the regulator asks.

> *Too many interruptions and the workforce has returned no hours. Too few and the organisation has delegated its judgement to a model. The fork holds the line.*

7 Autonomy is earned per decision and paced by reversibility

The phrase *autonomous workforce* implies a binary: a task is either done by a person or done by an agent. That binary is where most agentic deployments fail, because it forces the organisation

to choose between trusting a vendor's promise and getting nothing. Autonomy should be a dial, and the dial should be per decision, not per agent, not per system.

Stage	What happens	What the person does
Suggest	The system computes, applies the rules, drafts — and proposes.	Acts on the proposal, or not. The human process continues as before.
Approve	The system does the work and holds the result.	Signs it. Approving releases it.
Autonomous	The system does the work and releases it.	Audits a sample. Marks what was right and wrong.

Three rules govern the dial.

1. **The stage governs release, never the gate.** A blocking rule stops a run at every stage. Autonomy is about whether a person must sign what the run produced; it is never about whether the rules apply.
2. **Irreversible and outward-facing decisions are capped.** Anything sent to a customer or a regulator is approved by a person every time and cannot be promoted to autonomous. The server refuses the promotion. This is the rule that separates a governed workforce from the use cases in the infrastructure paper: the visa allocation and the municipal payment are precisely the decisions that stay at *approve*.
3. **Promotion needs the record, and one piece of evidence against is enough to pull back.** A decision is recommended for promotion after a run of clean executions with a high approval rate, no failures, and no reader marking the output wrong. A single thumbs-down from whoever read the output holds a promotion until it is answered. The system recommends; a supervisor moves the dial, with a reason, and the move is itself a row.

The consequence is that autonomy in a governed workforce is always a statement of fact about the past, not a promise about the future. The board can ask *which decisions run unattended, and on what evidence?* and receive a list, with the runs and the approvals and the reader feedback behind each entry. That is a different thing from a vendor's assurance of absolute trust, and it is the only version of trust an auditor will accept.

8 Hours returned are measured, not claimed

The loudest claim any workforce product makes is the hours it saves, so the ledger has to be built to survive the question that follows the claim. Most vendors report a figure computed from an assumption — an average task time, a benchmark, an estimate supplied by the vendor. The figure is unfalsifiable and the board knows it.

A measured ledger works differently. It begins with a baseline: a declaration, by a named person, of how long the human process took, with a stated basis — a time study, a manager's estimate, the organisation's own timestamps. Every line of the ledger shows who declared the baseline, when, and on what. Then:

- **No baseline, no number.** A decision nobody has priced reads *not measured*, never zero. The absence of a figure is itself information.
- **A run stopped at the gate returns nothing.** Blocked work is not saved work.
- **A run a person had to sign returns the baseline minus the review time**, and the result can be negative. An agent that costs more in review than it saves in labour is exactly what a pilot exists to find, and a ledger that cannot show a negative number cannot be trusted to show a positive one.
- **Baselines are kept as they stood.** A later revision rewrites nothing; each entry carries the baseline in force when the run happened. A first declaration reaches back over earlier runs and says that it did.
- **No trend line until there is a month to draw one from.** A chart drawn from a week of data is a claim in disguise.

This is the same discipline Paper Three applied to delivery timelines and Paper Six applied to outcome-based pricing: the number has to be one the buyer could have computed themselves, from evidence they hold. A ledger of this kind does not make the workforce look better. It makes the workforce look true, which is the only property a board can act on.

> *An agent that costs more to review than it saves is exactly what a pilot exists to find. A ledger that cannot go negative cannot be trusted when it is positive.*

9 The model is a component

The infrastructure paper's mandate to enterprise leaders includes an instruction that deserves to be taken seriously: do not run critical business workflows on black-box, off-shore models. The question is what follows from it, and the answer depends entirely on where the model sits in the architecture.

If the model is the decision-maker — decomposing goals, choosing steps, producing the numbers — then model choice is the most consequential decision in the deployment, and it is also a trap. An open-weight model on sovereign hardware satisfies the residency requirement and may lag a frontier model on reasoning; a frontier model behind a residency boundary offers the capability and imports a dependency. Either way the organisation has bet its decision quality on a vendor's weights, and cannot change the bet without changing every decision the workforce has made.

If the model only drafts, the question dissolves. Because the model never holds a value and never makes a decision, it is the one component in the architecture that can be replaced without touching the calculation, the rule resolution, the gate, the fork, or the ledger. An agent whose outputs are regulator-facing can run on an open-weight model inside the certified hypercloud, where nothing sensitive would be at risk even if the draft were poor, because the draft is validated before anyone reads it. An agent whose prose demands more can run on a frontier model through a residency-compliant boundary, and the same validators apply. The two can coexist in one roster. Either can be swapped per agent, and the row records which was used.

This is what it means to say that model sovereignty is a consequence of decision sovereignty. An organisation that has put the decision in code and the draft in the model has made the choice of model reversible — which is the only kind of choice a regulated institution should make about a technology that changes every quarter. The infrastructure operators are right that the model should be local where it can be. The architecture is what makes *where it can be* a question the organisation can answer differently for every agent, and change its mind about without cost.

10 What a workforce reports upward

Paper Four argued that an AI strategy for a mid-sized company is a ranked list of operations, not a list of tools, and that the strategy should be revisited quarterly against evidence. A governed workforce is where that evidence comes from. Because every run is a row, the workforce reports upward in a form a strategy layer can read — not summaries, not claims, but the entries themselves.

What the strategy layer asks	What the workforce reports
Where are hours actually being returned, by role?	The ledger's entries, each with its baseline and the provenance of that baseline.
Which decisions run unattended, and on what evidence?	The autonomy stage of every decision, and the runs, approvals, and reader feedback behind each promotion.
What is our compliance posture on the routine work?	Every rule evaluation with its instrument and citation — what blocked, what warned, what is still a placeholder awaiting curation.
Where is human judgement still required, and why?	The fork record: how often a person was needed, in which role, for which rule or missing fact.

One property makes this reporting trustworthy rather than merely available: the strategy layer and the workforce reason from the same regulatory memory. The law the strategy reasons about is the law each agent applied, on the date it applied it. A strategy built on a different reading of the rules than the operations that execute it is the ordinary condition of most organisations, and a shared substrate is the only thing that removes it.

This is also the answer to the question the infrastructure paper's readers raised most often in response: how does sovereign compute reach the small and mid-sized businesses that make up most of the economy? Not by giving them GPUs. By giving them a roster of agents that does the routine work of named roles, reports in rows, and returns hours they can count — running on whatever sovereign compute the operators provide.

11 Six questions for any vendor of an autonomous workforce

The series has ended each paper with something a buyer can use. These six questions apply to any product that calls itself an agentic or autonomous workforce, including the one this paper describes. A vendor who cannot answer them in a live system, from the record, is selling infrastructure or selling trust; neither is a workforce.

1. **Show me a number in last week's output. Where did it come from?** The acceptable answer is a calculation on a run record, from which the number was substituted into the draft. The unacceptable answer is *the model*.
2. **Which rule applied to this decision, and which instrument is it from?** The acceptable answer names the article, the pack version, and the date as of which it was resolved. *It's in the prompt* and *it's in the script* are the same answer, and both fail.
3. **Who decided, and can I read their reason?** The acceptable answer is a named role, a reason, and a row that cannot be edited. *Nobody* is the answer most products give when pressed.
4. **How much autonomy does this decision have, and what did it earn it?** The acceptable answer is a stage, per decision, with the runs and approvals behind it — and a list of decisions that can never be promoted.
5. **How many hours has it returned, and against whose baseline?** The acceptable answer shows the declaration, the declarer, and the basis, and includes at least one decision that reads *not measured* or returned a negative number.
6. **Can a person type a request into it?** The acceptable answer is *no, by design, and here is the test that fails the build if they could*.

None of these questions concerns the data centre. All of them will be asked by the same regulator who was reassured by the certificate on the building. The infrastructure operators have answered the question of where the work runs, and the country is better for it. The question of what was decided, under which rule, by whom, and on what evidence is the one the organisation cannot outsource — and the one a governed workforce is built to answer, one row at a time.

About this series

This is the seventh paper in a series on enterprise AI as an operating system rather than a tool. Paper One established where the value is: not in better models, but at level five of a maturity ladder, where the workflow is re-engineered around governed intelligence. Paper Two set out the method for re-deriving a process from its outcome. Paper Three answered the question of delivery. Paper Four turned the method into a strategy for a company of 50 to 500 people; Paper Five applied it to the accounting and advisory practice; Paper Six separated the commercial model from the delivery model that makes an outcome contractable. This paper describes the workforce that does the routine work once the strategy is written.

Every paper is readable in full at xamun.ai/whitepapers. The placement mechanics — how the deterministic mind intercepts, binds, and audits the model's output — are the subject of implementation work and outside the scope of this series. The regulatory substrate referred to throughout is compliance support, not a legal opinion.

Sources and references

- Al Awadi, J., *The Sovereign Agentic Infrastructure: Orchestrating the Autonomous AI Workforce for the UAE's Digital Core*, du Tech, September 2026.
- du / Open Innovation AI memorandum of understanding on sovereign GPU orchestration, signed at the Digital Readiness Retreat, Dubai, June 2026, in the presence of the UAE Cyber Security Council.
- Government of Dubai, May 2026 direction on private-sector agentic AI adoption; UAE federal target of 50% of government services delivered via AI agents by 2028.
- MIT NANDA, *The GenAI Divide: State of AI in Business* (2025) — the 95% of enterprise pilots reporting no measurable P&L impact.
- McKinsey & Company, *The State of AI* (2025) — workflow redesign as the strongest correlate of EBIT impact.
- Gartner (2025) — forecast that over 40% of agentic AI projects will be cancelled by end-2027.
- Xamun Position Papers No. 1 to No. 6 (August–September 2026), xamun.ai/whitepapers.

No. 7 in the Xamun position-paper series. Version 1.0, September 2026.