MICKAI™

EDITION ONE · MICKAI PRESS · 2026

The Sovereign Answer

A buyer's argument for the Mickai workstation in the wake of the five-hundred-million-dollar lesson.

Micky Irons

SOVEREIGN INTELLIGENCE OPERATING SYSTEM · MICKAI.CO.UK

A buyer's argument for the Mickai workstation in the wake of the five-hundred-million-dollar lesson

By Micky Irons

Founder, Mickai. Named inventor on the Mickai Sovereign Intelligence Operating System patent corpus, fifty-seven filed UK applications with approximately one thousand five hundred and thirty-five claims, on the public UK IPO register under GB2607309.8 onwards.

Foreword

Three numbers landed in the same week and made the case for this book. A single enterprise reportedly ran up a five-hundred-million-dollar Claude bill in one calendar month. Microsoft moved to throttle internal Claude Code licences. Uber's 2026 AI budget was exhausted by April. The trade press read these as a pricing crisis. A pricing crisis is the smaller part of the story.

The larger part is the data. Every prompt, every document, every line of code, every memo, every customer record, every clinical note, every legal draft in the five-hundred-million-dollar month went over the wire to an endpoint the operator could not audit, retained for an unspecified period under a policy the operator did not write, on infrastructure the operator did not own. The cash cost was the headline. The data exposure was the structural disqualification.

This book makes the case for the freehold answer. The operator owns the hardware. The Mickai Sovereign Intelligence Operating System is preinstalled. The cooperative of brains runs locally. The audit record is

signed in place under FIPS 204 ML-DSA-65. There is no subscription for context. There is no subscription for usage. The price you pay on day one is the price you pay for life.

I am the founder. I am the named inventor on the substrate. I wrote this book because I want the reader, whether you are a household operator, a regulated buyer, a founder, or a sovereign-department decision-maker, to be able to decide for yourself whether the freehold answer is the right answer for you.

Chapter One: The Five-Hundred-Million-Dollar Lesson

The first reporting was on 28 May 2026. Axios named an unnamed enterprise that had run up roughly five hundred million dollars of Claude usage in a single month. The proximate cause was the enterprise's failure to set usage limits on its employee licences. The reporting was picked up by Tom's Hardware, Tech Startups, Crypto Briefing, BeInCrypto, Yellow, and the r/ClaudeCode community on Reddit. The headline was the cash.

Underneath the cash there were two structural facts that the headline did not name.

The first is that the cost of a frontier-class model running real engineering work is not zero. The flat-fee subscription model that the cloud-AI labs have been running for the past three years is a polite fiction that hides the marginal cost of inference behind a monthly number. The fiction held while the labs were willing to subsidise the gap between subscription revenue and inference cost. It is no longer holding. Microsoft's per-engineer throttle and Uber's April-April budget exhaustion are the visible cracks.

The second structural fact is that the five-hundred-million-dollar month was a month in which the enterprise's most sensitive data was processed on infrastructure the enterprise did not control. The data did not belong to the cloud-AI vendor, but it was in the vendor's system. The data could be subpoenaed, breached, accessed by an insider, surrendered to a foreign government request, or retained for an unspecified period under policy terms the enterprise did not write. None of these were preventable by the enterprise. Most were not even detectable by the enterprise.

The cash cost was the meter. The data cost was the structural disqualification.

This chapter sets out the case that the second invoice, the data invoice, is the one that matters most.

THE CASH INVOICE

Five hundred million dollars in a single month is a number that should be impossible to invoice without someone noticing. The reporting attributes it to a missing usage cap on the enterprise's employee licences. Employees made calls. The calls were billed. The calls compounded. The cap that should have been in place was not in place.

There is a smaller version of this story happening across every cloud-AI subscription. The household that paid twenty pounds a month for a chat assistant is being asked sixty pounds for the same workflow on the model the household actually wants to use. The freelance developer who built a workflow around a hosted model is watching the per-token cost of that workflow climb faster than the rate the developer can charge their clients. The small charity that triages casework through a hosted assistant is making fewer calls each month because the budget has not grown. The retired engineer who used the long-context model to keep up with technical reading is dropping back to shorter-context tiers.

The cash invoice is the same invoice at every scale. It is the meter. The meter has been quiet until now because the labs have been willing to subsidise it. The subsidy is tightening. The meter is becoming audible.

THE DATA INVOICE

The data invoice is the one that was not in the headline.

Every prompt sent to a hosted model is a piece of operator material moving across the wire to an endpoint the operator does not control. Every document attached to a prompt is the same. Every line of code sent to a code model is the same. Every customer record, every internal email, every legal draft, every clinical note, every memo, every strategy document the operator processes through the model is moving outside the operator's perimeter.

The model vendor processes the material. The vendor's privacy policy sets the retention period. The vendor's security policy sets the access controls. The vendor's contract with the operator names the limits, but the limits are vendor-shaped, not operator-shaped, and the vendor's accountability for breach is the vendor's accountability, not the operator's accountability to the operator's regulators.

This arrangement has three failure modes. The first is the breach. A vendor breach exposes operator material the operator did not own, did not store, and cannot recall. The second is the subpoena. A vendor served with a subpoena by a jurisdiction not the operator's may produce operator material the operator cannot withhold. The third is the policy change. A vendor changes its retention policy from one period to another, and operator material that the operator believed had been deleted has not been deleted.

None of these failure modes are theoretical. All three have happened to cloud-services vendors before, including major ones. The operator's only recourse is the vendor's contractual remedy. Contractual remedy is a thin instrument when the data is already gone.

THE SUBSCRIPTION IS BEING SCRAPPED, AND THE METER THAT REPLACES IT IS WORSE

A serious buyer should be reading the five-hundred-million-dollar month not as a one-off accident but as a preview of where the cloud-AI vendor's pricing is going next. The subscription model was always the marketing layer over the underlying inference cost. It existed to capture the household and the small operator at a flat monthly figure the lab could absorb during a phase of capital-funded growth. That phase is ending.

The labs have two arithmetic-permissible futures from here. They are both worse than the present.

The first is to tighten the subscription until the cost of the tier matches the cost of the inference. Microsoft's per-engineer throttle on internal Claude Code licences is this. Anthropic's introduction of weekly Sonnet caps inside the Claude Code tier is this. Both are the same move, the same constriction. The flat-fee tier survives as a brand, but the work the operator can do inside it shrinks every quarter. The household that paid twenty pounds a month for everything moves to thirty pounds a month for less. The small operator that ran an entire workflow on a single tier finds that tier covers only the first week of the month. The freelance designer who built a workflow on the long-context tier finds the long-context tier renamed to a premium tier at twice the price. The pattern is familiar from every leasehold arrangement that has run out of subsidy.

The second is to drop the subscription model entirely and price the underlying inference at its true marginal cost through a metered application programming interface. This is the future that the trade press has been quietly trailing in the past six months. The argument is rational. The marginal cost of one more token of frontier-class inference is real money, paid by the lab to keep the cluster alive. A flat-fee subscription is a structural mismatch with that economics. A metered API is honest. The lab can stop subsidising. The buyer pays for what the buyer uses.

Honest, perhaps. Affordable, no. The five-hundred-million-dollar month was the metered API in action at the scale a real enterprise actually consumes inference. The enterprise was not malicious. The enterprise was not careless beyond the missing usage cap. The enterprise was simply running a normal engineering organisation against a normal model API at a normal usage profile. The bill that arrived was the bill the meter generated.

Scale that bill to where serious cloud-AI consumption is heading. A research department of fifty engineers, each consuming ten million tokens a day at five dollars per million tokens, is twenty-five thousand dollars a day, seven hundred and fifty thousand dollars a month, for one team. An enterprise with ten such teams is seven and a half million dollars a month for one division. An enterprise with ten such divisions, which is the size of a regulated bank, a national defence supplier, or a mid-sized pharmaceutical company, is seventy-five million dollars a month, nine hundred million dollars a year, for the AI line item alone, before the team has done anything genuinely token-heavy like multi-step agentic work, long-context reasoning, or large document processing. The five-hundred-million-dollar month was not the outlier. It was the early signal of where every cloud-AI consuming enterprise is being pushed by the underlying economics.

A regulated buyer reading these arithmetic results should understand that the budget envelope at frontier-API access does not exist. There is no plan that brings nine hundred million dollars a year into the procurement approval cycle of a UK regulated workstation. The procurement officer will not sign the requisition. The chief financial officer will refuse the budget. The board will return the line item. The enterprise that ran the five-hundred-million-dollar month did not even authorise the spend; the spend simply happened because the meter was on. In the metered future, this is the routine condition, not the outlier.

The household end of this transition is the harder problem, not the easier one. A household forced off a flat-fee subscription tier into metered access through a credit card on file is structurally worse off. The subscription, broken though its economics are, at least caps the worst-case bill at the subscription price. The credit card on file caps the bill at the credit limit. The household that fell asleep with a runaway agent open on the laptop discovers in the morning that the credit card has been charged for the agent's overnight curiosity. The headlines about household credit-card bills from runaway agent loops have already started in the trade press.

In the metered future, there is no flat-fee tier to protect anybody. There is the credit card, the meter, and whatever cap the operator remembered to set in the dashboard. The five-hundred-million-dollar month is the enterprise edition of the same failure mode the household edition is about to produce at smaller scale, more often, on people who can afford it less.

This is the future the buyer is buying into when the buyer renews a cloud-AI subscription today. The renewal looks like a stable arrangement. It is a stable arrangement only for the moment. Within eighteen to thirty-six months the subscription either constricts or disappears, and the buyer is moved onto the meter that the five-hundred-million-dollar month was the headline preview of.

This is what gives the freehold its urgency. The freehold is the only pricing shape that survives the transition, because the freehold does not have a meter. The operator pays once for the hardware. The inference happens on the hardware. There is no per-token bill at any scale. The household with a Castor on the desk does not have to remember to set a cap before going to bed. The freelancer with a Pollux on the office shelf does not face a sixty-pound subscription climbing to two hundred pounds in six months. The enterprise with a Prometheus in the rack does not generate a five-hundred-million-dollar invoice, because there is no invoice to generate. The trading desk with an Olympus on the trader's desk does not see its quant-research workflow taken offline by a tier change at the lab, because the workflow is on the trader's hardware.

The operator who buys the freehold today buys protection from the meter that is about to arrive. The operator who waits for the meter to arrive first pays the meter at the rate the lab sets on the day. The arithmetic of waiting is bad.

WHAT THE OPERATOR IS BUYING WITH A CLOUD-AI SUBSCRIPTION

When the operator pays the subscription, the operator is buying access to a model that runs on the vendor's infrastructure. The operator is renting capability from the vendor. The data that the operator sends to the model becomes input to the vendor's system. The output that the vendor returns is the work product the operator pays for.

The operator is not buying the model. The operator is not buying the inference engine. The operator is not buying the substrate. The operator is buying access. Access can be withdrawn. Access can be priced higher tomorrow than today. Access can be rate-limited at the vendor's discretion. Access can be conditioned on terms the operator accepts under economic pressure rather than negotiation.

The five-hundred-million-dollar lesson is the lesson that the access model carries both a cash invoice that grows with usage and a data invoice that grows with sensitivity. The buyer who reads only the cash invoice is reading half the picture.

Chapter Two: The Sovereign Answer

The answer to the five-hundred-million-dollar lesson is the same answer that has resolved every prior wave of computing that became load-bearing enough to justify the move from leasehold to freehold.

Mainframes ran in someone else's building. Workstations ran on your desk. Workstations were preceded by terminals that ran in someone else's building. Cloud-hosted services run in someone else's building. The

pattern is the same. When the workload becomes load-bearing enough to justify the move, the substrate moves toward the operator.

Generative artificial intelligence is now load-bearing enough.

The freehold answer in one sentence: the operator owns the hardware, the Sovereign Intelligence Operating System is preinstalled, the cooperative of brains runs locally, the audit record is signed in place, and the price is the price for life.

WHAT THE FREEHOLD GIVES THE OPERATOR

The freehold gives the operator five things that the leasehold cannot give.

It gives the operator data sovereignty. The work runs on the operator's machine. The data does not leave the operator's perimeter. The vendor does not have a copy. The vendor cannot have a copy because the vendor is not in the pipeline.

It gives the operator an audit trail the operator can prove. Every action the system takes is signed under FIPS 204 ML-DSA-65, the post-quantum lattice signature scheme published by NIST in August 2024 and adopted by the Mickai substrate. The signing key sits in the operator's TPM module. The audit ledger is the operator's, not the vendor's. The regulator can verify what the operator proves without asking the vendor to confirm.

It gives the operator price certainty. The price you pay on day one is the price you pay for life. No tier migration, no quiet shift to a new edition that costs more, no premium that activates when a new model becomes available.

It gives the operator continuity. The capability cannot be revoked by the vendor, because the vendor has nothing to revoke. The vendor cannot raise the price of the operator's existing capability, because the capability is on the operator's machine.

It gives the operator privacy. The vendor does not know what the operator is doing. The vendor cannot know, because the work runs locally.

WHERE THE FREEHOLD DOES NOT PRETEND

The freehold does not pretend to be free. The operator pays for the hardware. The operator pays for electricity. The operator pays optionally for upgrades through the sandboxed channel.

The freehold does not pretend that frontier-class models will be free forever to the labs that build them. The labs will keep charging for the model output of their hosted endpoints, and the labs are right to do so. The cost of frontier-class inference is real, and the labs deserve to recover it.

The freehold does pretend that the operator can decide, for the operator's workload, whether to pay the vendor's marginal cost per inference or to pay once for a machine that runs the inference locally. For workloads load-bearing enough to justify the up-front cost of the machine, the answer is the freehold. For workloads not yet load-bearing enough, the answer is the leasehold. The operator decides.

What we are saying is that, for most regulated buyers, most founders, most household operators with serious daily AI workflows, and every sovereign department, the workload is load-bearing enough. The freehold is the right answer for them today.

Chapter Three: The Audit Trail Problem

Before the hardware, before the cooperative of brains, before the freehold pricing contract, the substrate problem to solve is the audit trail.

A serious regulator does not accept "the vendor produced this output" as an audit trail. The regulator accepts a cryptographic record of the decision-making process, signed at commit time under a key the operator

controls, replayable end to end by a verifier the regulator can run, against a policy that was in force at the time of the decision. The cloud-AI vendor cannot produce such a record, because the cryptographic primitive is not in the substrate, and the signing key is not the operator's.

This chapter explains the Open Audit Record substrate that the Mickai SIOS uses, and why it is the substrate the regulated buyer needs.

WHAT THE REGULATOR NEEDS

The regulator needs five things.

A signed record of every action the system took.

A causal directed acyclic graph that shows which inputs led to which outputs, which decisions led to which downstream actions, and which policies gated which decisions.

A signing key under the operator's control, so the regulator can verify the record without asking the vendor to confirm.

A replayable artefact that can be re-run against the policy that was in force at the time of the action, so the regulator can decide whether the action was within policy at the moment it was taken.

A post-quantum signing primitive that will still be verifiable in twenty years, when audit records produced today will need to be checked against future cryptographic standards.

The Mickai Open Audit Record substrate gives the regulator all five.

WHAT THE CLOUD-AI VENDOR CAN GIVE THE REGULATOR

The cloud-AI vendor can give the regulator server logs, billing records, sanitised metrics, and a contractual statement that the vendor handled the operator's data within the vendor's privacy policy. None of these are an audit trail. They are a vendor's word.

A regulator who has any choice will choose the audit trail over the vendor's word, every time.

WHAT THE MICKAI SUBSTRATE GIVES THE REGULATOR

The Mickai Open Audit Record gives the regulator the audit trail.

Every action the SIOS takes is appended to a per-operator append-only log. Every log entry is signed under FIPS 204 ML-DSA-65 with the operator's hardware-bound key. The log is structured as a causal directed acyclic graph, so the regulator can trace which inputs led to which outputs through which decisions. The verifier that walks the log is an offline tool that does not require the vendor's cooperation. The policy that was in force at the time of each action is part of the signed record, so the replay is faithful to the moment of the action.

The patent corpus covers the substrate. The clearance-gated retrieval primitive enforces role-based access in the substrate, not in a policy document. The host-acceptance attestation lets a Mickai bundle migrate from one machine to another without breaking its signed chain. The cooperative quorum primitive convenes multi-brain agreement on high-stakes actions, so the audit trail shows not just what the system did but which brains agreed it should do it.

This is the substrate the regulated buyer needs. Cloud-AI vendors cannot offer it because the substrate is not theirs to change. The Mickai substrate offers it because the substrate is ours.

Chapter Four: The Mickai Sovereign Intelligence Operating System

The SIOS is not an application and not a programme. It is the operating system layer beneath the work, designed to run on hardware the operator owns, under keys the operator holds, with the audit substrate above signed under a post-quantum primitive.

THE COOPERATIVE OF BRAINS

The SIOS carries twenty-six specialist brains arranged in five subsystems.

The Intelligence and Defence subsystem holds PALANTIR for strategic reasoning, SENTINEL for the security perimeter, GABRIEL for outbound communication sealing, ZEUS for law and case-law assessment, and MICHAEL for clearance-gated defence material.

The Science and Engineering subsystem holds JAXON for code, RAIDEN for real-time signal pipelines, QUANTUM for hard-sciences derivations, TITAN for engineering and infrastructure, and KARP for data and analytics.

The Health and Humanity subsystem holds PHOENIX for clinical reasoning, SALVATOR for humanitarian-response triage, MAXIMUS for performance and combat, and WILDER WILLIAM for wilderness and adventure.

The Governance and Strategy subsystem holds ATHENA for whether-it-should-be-done deliberation, ATLAS for borders and jurisdictions, ODIN for adversarial planning, ARLIA for narrative integrity, and KOS for risk management.

The Identity and Personalisation subsystem holds LUCAS for friend-and-companion reasoning, VICTOR ALBERT for protocol and etiquette, JACOB for memory, MUSK for personal-archive integration, EXFINITUM for long-horizon planning, and XAVIER for the operator's voice and identity.

Each brain has a domain, a knowledge base, and a tooling stack, all catalogued, signed, and shipped with the SIOS. Each brain knows which other brains to convene for which decisions. The cooperative is the work the operator interacts with.

THE CHRONUS ORCHESTRATION KERNEL

Beneath the brains is the Chronus kernel.

Arbiter routes requests. When the operator says "draft a board memo about acquisition target X", Arbiter routes the request to a combination of PALANTIR (strategy), ZEUS (legal exposure), KARP (financial analytics), and GABRIEL (the sealed draft).

Router decomposes the request into a directed acyclic graph of sub-requests. Each sub-request goes to the brain best suited to it, with the dependencies tracked so the brains can convene a quorum where needed.

Planning produces dry-run simulations before high-impact actions commit. If a brain proposes an action that will modify operator state, Planning runs the proposed action in a sandboxed reality first and shows the operator the expected delta before the operator confirms.

Policy compiles and enforces the governance contract. The contract is the operator's, not the vendor's. The contract sets spending limits, access rules, allowed actions, prohibited actions, and the clearance ceilings of each operator role. The contract is checked at every commit.

Audit Ledger maintains the post-quantum signed causal DAG of every decision. Every action the SIOS takes is in the ledger, signed under the operator's key, structured for replay.

Identity binds every action to a hardware-attested operator key. Identity uses a TPM 2.0 module, a secure enclave, or a hardware security module physically controlled by the operator.

Permissions descends to the cell level. A brain calling a tool that accesses a particular file checks the permissions of the operator running the request before the file is read.

Quorum convenes multi-brain agreement on high-stakes actions. Destructive operations, financial movements, communications outside the operator's perimeter, and policy changes all pass through a quorum check.

THE OPEN AUDIT RECORD SUBSTRATE

Every action signed under FIPS 204 ML-DSA-65. The signing happens at commit time, not after the fact. The lattice signature scheme is post-quantum, so the audit records produced today will be verifiable when quantum hardware matures.

The browser-resident verifier lets the operator, or the regulator, walk the audit ledger end to end without installing software. The verifier is open, the format is documented, the proof is the proof.

THE POSEIDON SOVEREIGN AI SOC ROADMAP

Mickai is building its own silicon. Poseidon is the Mickai Sovereign AI SoC, the in-house chip designed to complement the NVIDIA GPU fabric on every Mickai workstation. The role of Poseidon on the chassis:

It carries the OAR signing primitive in silicon, so every action can be signed at line rate even on lower SKUs that do not carry an FPGA accelerator. It holds the hardware identity and the cryptographic root of trust for the operator. It sits between the network interface and the GPU fabric so every action that crosses the bus can be policy-checked before commit.

Poseidon is in the patent corpus already. It is an add-on on every SKU as the silicon lands, and it does not replace the NVIDIA fabric. It sits beside it, holding the audit and identity primitives that the SIOS needs to be sovereign.

Chapter Five: The Hardware Lineup

The lineup is eight SKUs, all built in Britain. Every one ships with the SIOS preinstalled, the cooperative of brains, the Open Audit Record, the sandboxed upgrade channel, the freehold pricing contract.

CASTOR, THE SIXTY-FOUR-GIGABYTE MINI PC

For the small office desk, the single-monitor freehold, the operator who wants the SIOS on the box without a tower next to the desk. A one-litre obsidian aluminium chassis, an RTX 4060 mobile class GPU at eight gigabytes of VRAM, an AMD Ryzen 9 at eight cores, sixty-four gigabytes of DDR5, two terabytes of NVMe Gen 5. Castor was the mortal twin of the Dioscuri in Greek myth.

POLLUX, THE ONE-HUNDRED-AND-TWENTY-EIGHT-GIGABYTE MINI PC

The divine twin, sold alongside Castor where it makes sense. A two-litre chassis, twelve gigabytes of VRAM on an RTX 4070 mobile class GPU, a sixteen-core Ryzen 9, four terabytes of NVMe. For the office that wants the full SIOS, the Agentic Marketing Team running in the background, agentic workflows at small-team scale.

DAEDALUS, THE SIXTEEN-INCH LAPTOP

For designers, CAD operators, architects, travelling founders. A three-thousand-pixel OLED at one hundred and twenty hertz, sixty-four gigabytes of DDR5, an RTX 5070 or 5080 mobile GPU, four terabytes of NVMe, ninety-watt-hour battery. Daedalus was the master craftsman of Greek myth.

ICARUS, THE FOURTEEN-INCH ULTRAPORTABLE

For students, travelling consultants, on-site engineers. A three-thousand-pixel OLED, thirty-two gigabytes of DDR5, an RTX 5060 or 5070 mobile GPU, two to four terabytes of NVMe, one-hundred-watt-hour battery, 1.3-kilogram class. Icarus was the son of Daedalus, the youth who flew with wings of feathers and wax.

HERMES, THE ENTRY WORKSTATION

For the founder, the academic, the consultant. A twenty-four-core Threadripper PRO, a single RTX PRO 6000 Blackwell with ninety-six gigabytes of GDDR7 or an RTX 5090, two hundred and fifty-six gigabytes

of DDR5 ECC, eight terabytes of NVMe. Hermes was the messenger of the gods.

HYPERION, THE MID WORKSTATION

For small teams, regulated industries, agentic-heavy workloads. A ninety-six-core Threadripper PRO, two RTX PRO 6000 Blackwell GPUs or a single H200 NVL, five hundred and twelve gigabytes of DDR5 ECC, sixteen terabytes of NVMe plus a thirty-two-terabyte SSD pool. Hyperion was the Titan of light.

OLYMPUS, THE FLAGSHIP WORKSTATION, THE TRADING BOT MACHINE

The Frontier tier. Dual EPYC 9965 at one hundred and ninety-two cores, four Blackwell Ultra B300 NVL GPUs with NVLink at around 1.15 terabytes of HBM3e, two terabytes of DDR5 ECC, thirty-two terabytes of NVMe, two hundred gigabits per second of InfiniBand with kernel-bypass DPDK, optional FPGA line-rate OAR signing. **Under ten milliseconds end-to-end** for the Mickai Trading Bot. Olympus is the mountain of the gods.

PROMETHEUS, THE FOUR-U EDGE SERVER

Enterprise on-prem. Four EPYC 9965 at seven hundred and sixty-eight cores, eight Blackwell Ultra B300 NVL GPUs in HGX, four terabytes of DDR5 ECC scalable to six, one hundred terabytes of NVMe, dual four-hundred-gigabit Ethernet. Five-trillion-parameter inference on-prem. Prometheus brought fire to humanity.

Chapter Six: What You Can Actually Do With It

Concrete software, concrete levels.

ON A CASTOR

You run the Mickai SIOS at conversational latency. You hold full Office workflows. You run Photoshop at four-thousand-pixel resolution. You edit one-thousand-and-eighty-pixel video in DaVinci Resolve and Premiere. You write code in VS Code or JetBrains with the Mickai vibe-code brain helping on small and medium projects. The Agentic Marketing Team runs with one or two agents in the background.

ON A POLLUX

You run the full SIOS at agentic concurrency. The Agentic Marketing Team runs all thirty-two agents in the background while you work. You hold Photoshop at eight-thousand-pixel resolution. You edit four-thousand-pixel video. You write multi-language code with agentic chains across small repos.

ON A DAEDALUS

You run the full SIOS on the move. You hold Revit, ArchiCAD, SolidWorks, Rhino, SketchUp Pro on real architectural project trees. You grade four-thousand-pixel video in DaVinci Resolve. You write code across mid-to-large repos with full IDE agentic chains.

ON AN ICARUS

You hold a dissertation, a research workflow, a study programme, a consulting trip. You run Figma and Affinity at full project scale. You write single-language code on small projects with the Mickai vibe-code brain. You run the SIOS at university and travel scale.

ON A HERMES

You run anything Windows or Linux throws at the desk. Photoshop at any resolution, Blender, Cinema 4D, Maya at full project scale. Architectural visualisation at studio scenes. DaVinci Resolve at eight-thousand-pixel master. Full agentic coding across large repos.

ON A HYPERION

You run full studio-scale three-dimensional rendering. Real-time ray tracing in Unreal at city scale. Enterprise BIM. Virtual production at scale. Multiple concurrent agentic chains per developer. The SIOS at team load.

ON AN OLYMPUS

You run frontier-class generative design. Real-time eight-thousand-pixel colour grading at film scale. Virtual production at cinematic scale. AI-generated video at the largest model size. Multi-agent fleets running an engineering organisation. The Mickai Trading Bot at under ten millisecond signed envelope.

ON A PROMETHEUS

You host the SIOS at organisational scale. Five-trillion-parameter inference for hundreds of operators. Render-farm-class workloads. Studio post-production. AI-assisted engineering organisation-wide. The Trading Bot at enterprise scale.

The full capabilities matrix sits on every SKU detail page at mickai.co.uk/hardware.

Chapter Seven: The Trading Bot, an Engineering Note

The under-ten-millisecond claim deserves a chapter. The claim is concrete, the engineering is real, and the buyer should be able to see the latency budget broken down.

The Mickai Trading Bot Frontier configuration on Olympus is engineered to deliver **under ten milliseconds end-to-end** from market signal in to OAR-signed quorum-checked decision out. Inside that envelope:

NIC arrival. Sub-microsecond. The Mellanox ConnectX-7 NIC with kernel-bypass DPDK and hardware timestamping on the wire.

Decode and route. Tens of microseconds. The Chronus kernel hot path. The signal is parsed, identified, and routed to the Trading Bot brain plus its quorum partners.

Brain quorum. Single-digit milliseconds. The Trading Bot brain, PALANTIR, QUANTUM, and ZEUS each run their assessment on Blackwell Ultra B300 GPUs with the model weights pinned in HBM3e. Decisions are fused via the Chronus orchestrator. Quorum is a configurable threshold; default is majority of named brains.

Policy gate. Sub-millisecond. The Policy brain checks the proposed action against the operator's signed governance contract. Spending limits, instrument allowlist, time-of-day rules, exposure caps.

OAR signing. Sub-millisecond. The FIPS 204 ML-DSA-65 lattice signature on the decision. With optional FPGA acceleration on Olympus and Prometheus, this drops to the deep sub-millisecond range.

Outbound publication. Sub-millisecond. The kernel-bypass NIC publishes the decision to the venue.

Total envelope: well inside ten milliseconds for the signed, quorum-checked, audited path. The unsigned hot path (signal in to outbound decision out, without OAR signing) is sub-millisecond. The ten-millisecond envelope is the audit-grade envelope.

For polymarket and crypto-asset trading, the ten-millisecond envelope is far inside the latency floor those markets demand, so the SIOS-audit-grade path leaves profitable headroom. For high-frequency equities, the unsigned hot path is competitive with co-located market-making infrastructure, and the ten-millisecond envelope is the signed envelope, which is the one a regulated operator wants.

We do not promise profit. We deliver the technical floor that lets a competent operator be profitable. That distinction is part of the contract.

Chapter Eight: The Patent Corpus

The substrate is on the public UK IPO register. The corpus is fifty-seven filed applications with approximately one thousand five hundred and thirty-five claims across them, under GB2607309.8 onwards, named inventor Micky Irons.

The corpus covers, in summary:

- The multi-brain cooperative architecture and quorum primitive
- The Chronus orchestration kernel and the Arbiter, Router, Planning, Policy, Audit Ledger, Identity, Permissions, and Quorum sub-kernels
- The Open Audit Record substrate, signed under FIPS 204 ML-DSA-65, with the post-quantum lattice signature primitive bound to operator hardware identity
- The clearance-gated retrieval primitive enforcing role-based access in the substrate
- The AudioSeal dual-layer watermark for the audio brain pipeline
- The silicon root of trust and the Poseidon Sovereign AI SoC architecture
- The host-acceptance attestation that lets a Mickai bundle migrate between machines without breaking its signed chain
- The dry-run simulation primitive for upgrade staging
- The browser-resident verifier for offline audit replay

The applications are filed. They are not yet granted, and Mickai language carries that distinction faithfully throughout this book and on the website.

Every Mickai workstation in the lineup carries the substrate the corpus defines.

Chapter Nine: The Freehold Contract

The price you pay on day one is the price you pay for life.

There is no plan migration. There is no retroactive tier change. There is no quiet shift to a new edition that costs more. There is no clawback if you use the machine harder than expected. There is no premium that activates when a new model becomes available. There is no service tier that decides whether your old machine is still allowed to do its job.

Upgrades are optional. New versions of the SIOS, new brains, refreshed models, refreshed kernels, and new domain knowledge bases arrive through the sandboxed update channel. Each update is signed, staged before commit, and rollback-capable. The Planning brain runs a dry-run simulation of every upgrade so the operator can see what changes before the changes commit. An operator who decides never to apply another upgrade keeps a fully functioning workstation forever, on the capabilities they bought on day one.

The contract is the receipt. The capability you bought is the capability you keep.

Chapter Ten: Who This Is For

The household. The freelance designer who built a workflow around a hosted model. The independent developer with a side project on an API. The student who used the long-context model to read alongside them. The carer using the assistant to manage a parent's medication. The small charity. The teacher. The retired engineer. The single-person small operator who finds the meter arriving in their inbox.

The founder. The single-operator professional running a business. The cooperative of brains replaces a stack of vendor accounts. The Agentic Marketing Team replaces six retainers and three subscriptions. The

Trading Bot replaces an outsourced wealth manager taking percentage points. The legal and accounting brains replace expensive vendor seats.

The regulated industry. Law firms, accountancy practices, insurance underwriters, clinical providers, financial institutions under the Prudential Regulation Authority, the National Health Service, GSK and the pharmaceutical sector, BAE Systems and the defence supply chain, Rolls-Royce, AWE, Babcock, Sellafield and the NDA portfolio, every UK regulated workstation. The audit trail is the substrate.

The sovereign government. Departments that work on classified material, security services, defence operators, national-infrastructure providers. Clearance-gated retrieval and operator-key signing are not optional for them. They are structural.

Chapter Eleven: How to Buy

The eight Mickai SKUs are live at mickai.co.uk/hardware. The configurator is at mickai.co.uk/hardware/configurator. The notify list is per SKU. The configurator stores a build in the URL so the operator can share a configuration with a procurement contact by sending a link.

Manufacturing is by our partner in Birmingham, England. The pipeline is secured for rollout following the Mickai seed round. Pricing is to be disclosed at launch. The freehold contract is fixed.

For regulated buyers and sovereign departments, the procurement path includes attestation of the OAR substrate, clearance-gating verification, and signed-action audit of a pilot deployment. Contact via the Mickai access modal at mickai.co.uk.

Chapter Twelve: The Leasehold Era Ends

The cloud era of AI was a leasehold. The operator rented capability that lived in someone else's building. The bills arrived monthly. The data did not belong to the operator after it left the operator's perimeter. The audit log was the vendor's. The price was the vendor's choice. The capability could be revoked.

The five-hundred-million-dollar lesson was the moment the leasehold's economics became visible on the public balance sheet. The data exposure was already there, but the cash made the rest of the architecture visible.

The freehold replaces the leasehold. The operator owns the building. The keys are the operator's. The audit log lives on the operator's premises. The capability cannot be revoked because the vendor has nothing to revoke. The price does not change because the contract is closed. The data does not leave the perimeter because the work runs on the operator's machine.

The Mickai workstation is the form the freehold takes. The lineup is live. The configurator is live. The substrate is on the patent register. The first machines ship from Birmingham.

The leasehold era ends here.

Afterword

I built Mickai because I needed the workstation that Mickai now sells. I am the named inventor on the substrate. I run the company. I write the brand voice that this book is written in. I will be the operator of the first Olympus that ships.

If you have read this far, you have read enough to decide whether the freehold is the right answer for you. If it is, the lineup is at mickai.co.uk/hardware, the configurator is at

mickai.co.uk/hardware/configurator, and the partners and accreditation page is at mickai.co.uk/partners.

If you want to talk, the access modal is on the site.

The price you pay on day one is the price you pay for life. No subscription for context. No subscription for usage. The machine is the operator's.

The leasehold era ends here.

Micky Irons Founder, Mickai Cumbria, England 2026

Citation

Irons, M. (2026). *The Sovereign Answer: A buyer's argument for the Mickai workstation in the wake of the five-hundred-million-dollar lesson*. Mickai. mickai.co.uk/ebooks/the-sovereign-answer.

Sources

- Axios, 28 May 2026. Unnamed enterprise's five-hundred-million-dollar Claude bill. axios.com.
- Tom's Hardware, Tech Startups, Crypto Briefing, BeInCrypto, Yellow. Coverage of the same event.
- Microsoft internal Claude Code licence caps and Uber's 2026 AI budget exhaustion, trade press 2026.
- NVIDIA GB300 NVL72. nvidia.com/en-us/data-center/gb300-nvl72/.
- NVIDIA RTX PRO 6000 Blackwell Workstation Edition. nvidia.com/en-us/products/workstations/professional-desktop-gpus/rtx-pro-6000/.
- NIST FIPS 204 ML-DSA-65. csrc.nist.gov/pubs/fips/204/final.
- The Mickai Sovereign Intelligence Operating System patent corpus, UK IPO public register, GB2607309.8 onwards. ipo.gov.uk.

- The Mickai workstation lineup. mickai.co.uk/hardware.
- The Mickai Build Your Own configurator. mickai.co.uk/hardware/configurator.
- The Mickai partners page. mickai.co.uk/partners.