White Paper – A Universal OS for AI

Building the Invisible infrastructure of Artificial Intelligence

Artificial intelligence is emerging as the true infrastructure of the 21st century.
Yet today, it relies on an almost complete technological monopoly: Nvidia and its proprietary language, **CUDA.**

THE PROBLEM

- Over **70% of AI models** currently run on Nvidia GPUs.
- Every new chip (Google TPU, AWS Trainium, photonic or neuromorphic processors) comes with its **own software language**.
- Companies are trapped, forced to **rewrite their models at every hardware change** – a process costing millions and creating total dependency.
- AI execution today lacks **trust and verifiability**: results can be altered, spoofed, or impossible to trace once computed.

OUR SOLUTION

A universal operating system for AI: neutral, adaptable, capable of running any model on any processor (GPU, TPU, photonic, neuromorphic, edge computing).

At its core lies AXIR (AXIOM Intermediate Representation), a universal language that automatically translates between all AI frameworks and all hardware architectures.

One single development → execution everywhere.

Cryptographic Trust Layer

To ensure integrity, traceability, and accountability in AI computation, AXIR integrates a cryptographic signature and verification layer.
Each model execution produces a verifiable digital signature that certifies:

- The authenticity of the model executed,
- The integrity of the computational graph,
- The validity and reproducibility of the output, regardless of the hardware used.

This guarantees trustable AI at scale — a foundation where every computation can be proven genuine, securely audited, and cryptographically verified.

Through this trust layer, AXIR establishes the first “Proof-of-Computation” protocol for artificial intelligence — a new standard for transparent, verifiable, and sovereign AI execution.

OUR CONDITION

To become for AI what Windows or ARM have been for computing: an invisible, universal, and indispensable layer. Our mission is to free companies from this monopoly and enable AI to scale without borders, worldwide.

OUR MISSION

Axiomos will ree enterprises from monopoly lock-in and enable AI deployment worldwide, without borders.

Author: Pierre SECK- Founder & Visionary, DeepTech Entrepreneur

Executive Summary

Artificial Intelligence (AI) is on the verge of becoming the key infrastructure of the 21st century, much like electricity or the Internet. It is already at the core of numerous sectors: healthcare, finance, industry, telecommunications, and defense.
Yet, this critical infrastructure today relies almost entirely on a single player: **Nvidia**. Its GPUs and proprietary CUDA language account for more than 70% of the computing power used for AI worldwide.

The Problem: Systemic Dependence

- **Software lock-in:** CUDA has become the de facto standard, but it remains closed. Any model developed for CUDA can only run on Nvidia GPUs.
- **Growing hardware fragmentation:** every new entrant (Google's TPU, AWS's Trainium, Microsoft's Maia, photonic or neuromorphic chips) brings its own software environment, creating a true technological chaos.
- **Massive costs:** adapting or migrating an AI model to another hardware architecture can require several months of work, specialized teams, and budgets reaching into the millions.
- **A major strategic risk:** tomorrow's global economy rests on an infrastructure controlled by a single vendor — a vulnerability comparable to depending on a sole supplier of energy or Internet access.
- **Lack of computational trust:** current AI systems provide no cryptographic assurance that a model's output genuinely originates from the model and parameters claimed. There is no verifiable proof of authenticity or integrity in AI computations.

The Solution: A Universal OS for AI

We propose the development of a universal operating system for artificial intelligence. At its technological core lies AXIR (AXIOMOS Intermediate Representation), which acts as a pivot language and neutral translator between:

- **AI models** (PyTorch, TensorFlow, etc.),
- **All existing and future hardware** (Nvidia GPUs, Google TPUs, AWS Trainium, photonic processors, neuromorphic chips, edge devices).

Key Benefits

- **Write once → run everywhere**, with AXIR as the intermediate standard.
- **Automatic optimization** tailored to specific needs: speed, cost, or energy efficiency.
- **Restored freedom** for enterprises, no longer locked into a single technology provider.

The Cryptographic Trust Layer

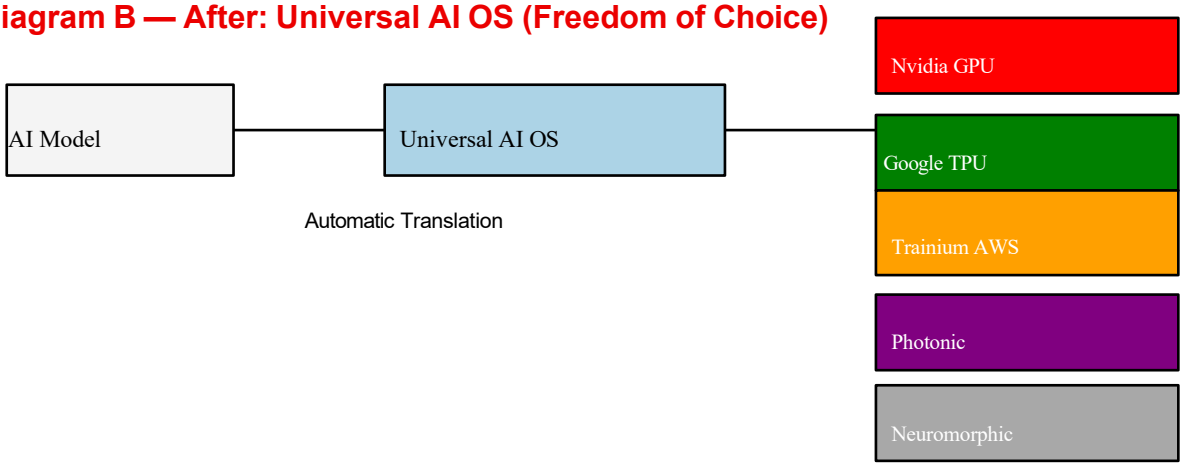
Beyond portability and performance, AXIR introduces a **cryptographic trust layer** designed to bring **verifiability and accountability** into AI computation.
Each model execution generates a **digital signature** and **cryptographic proof** certifying:

- The **authenticity** of the model and parameters used,
- The **integrity** of the computational graph,
- The **validity** and **traceability** of the results across devices.

This establishes a foundation for **verifiable AI** — where outputs are **provably genuine**, **tamper-proof**, and **auditable** without reliance on a centralized authority.
By combining universal execution with cryptographic verification, Axiomos sets the stage for a new paradigm:
Trustable AI Infrastructure — Secure, Neutral, and Borderless.

Diagram A — Before: CUDA Lock-in

Diagram B — After: Universal AI OS (Freedom of Choice)



Market & Opportunity

A colossal market in rapid acceleration

The global market for AI-dedicated chips is estimated at $150 billion in 2025 (McKinsey, Markets&Markets). By 2035, it could reach nearly $1 trillion, with an annual growth rate exceeding 20%. AI-focused data centers alone are expected to generate over $400 billion in annual spending by 2030 (Bloomberg, IDC). In short: we are witnessing one of the strongest waves of technological growth since the advent of the Internet.

The Current Dynamics

- **Nvidia dominates: with more than 70% market share, its GPUs and CUDA language control the majority of global AI computation.**
- **The tech giants strike back: Google (TPU), Amazon (Trainium), and Microsoft (Maia) are investing massively in their own chips to reduce dependency.**
- **A wave of hardware startups (Cerebras, Groq, Graphcore, Lightmatter, etc.) is exploring revolutionary architectures — wafer-scale, photonic, neuromorphic.**

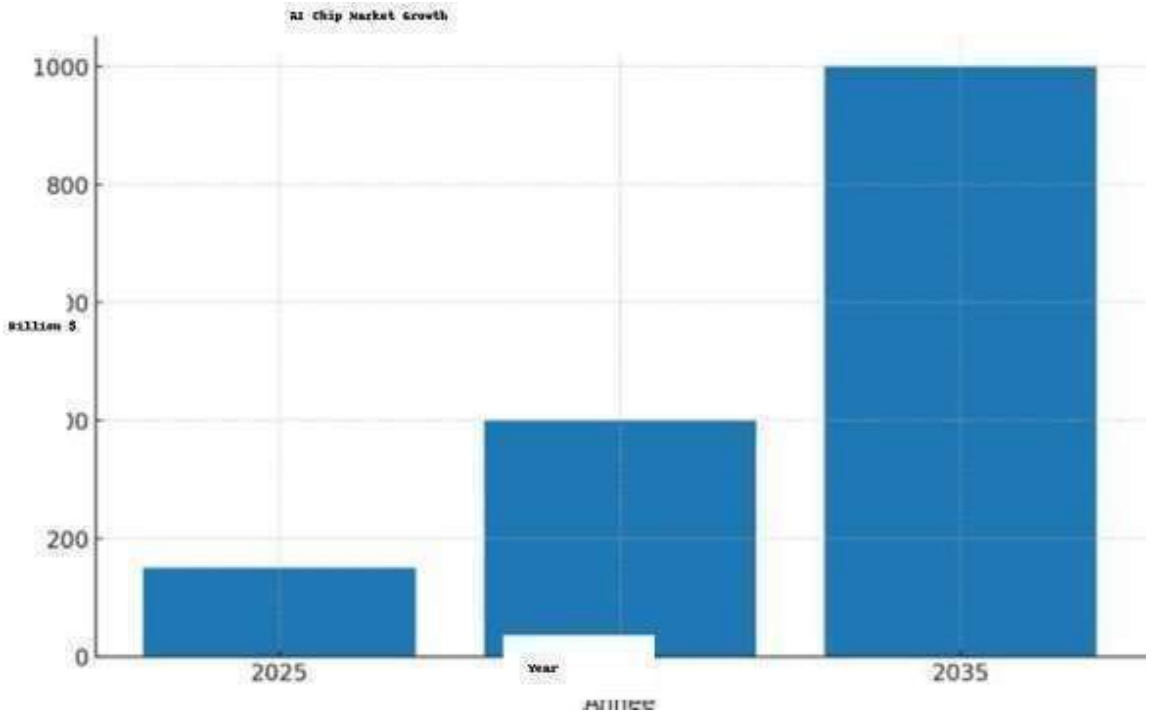
→ **Yet despite this momentum, adoption of these innovations remains limited. Why? Because dependence on CUDA and software fragmentation create a massive barrier.**

The Opportunity: A Universal OS for AI

This is where the strategic opening lies.

- Even a modest **2% penetration** of the global market would already represent **$20 billion** in annual revenue.
- Adoption of such an OS could be swift, driven by:
  - Companies seeking to diversify their suppliers and move away from exclusive dependence on Nvidia.
  - Hardware startups in need of a neutral OS to make their innovations truly usable.
  - The growing demand from hyperscalers (AWS, Azure, Google Cloud) for solutions capable of operating across multiple hardware architectures.
  - Governments and regulated industries demanding **verifiable, traceable, and sovereign AI computations** for compliance and security reasons.

AXIR, the intermediate language we are developing, is the key to this rapid adoption.
It acts as a “lingua franca” between all AI frameworks (*PyTorch, TensorFlow, ONNX, etc.*) and all hardware architectures (*GPU, TPU, photonic, neuromorphic, edge*), while embedding a cryptographic verification layer that ensures every computation is authentic, traceable, and verifiable across any device.


AI Chip Market Growth
1000
800
200
0
2025
Year
2035

Business Model

Our economic model is built on three main revenue streams, all highly scalable and complementary:

1. SaaS Licenses (Software-as-a-Service)

We charge a percentage (between 1% and 5%) of companies' AI expenditures.

- *Example*: a bank spending $100M per year on AI computing would pay about $3M to use our OS.
- *Advantage*: a recurring SaaS model, with predictable and rapidly growing ARR.

2. Cloud Commissions

AXIR acts as a universal gateway; every compute hour flows through this layer, enabling us to capture a transparent and systematic margin. Our OS integrates directly with major cloud providers (AWS, Azure, GCP).

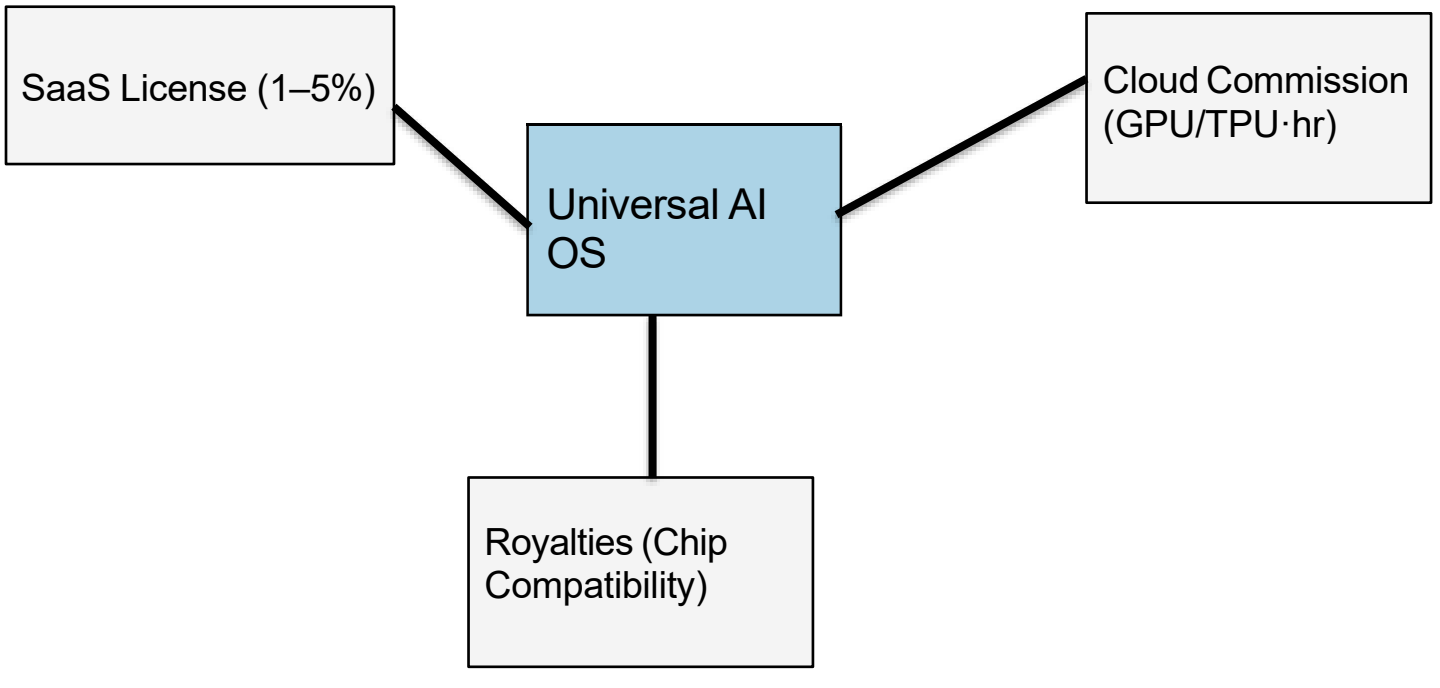
- We collect a margin on each compute hour (GPU-hour, TPU-hour, Trainium-hour, etc.) executed via our OS.
- *Advantage*: growth directly tied to the explosion of AI cloud computing.

3. Royalties (the "invisible tax")

To access the market, chip manufacturers must guarantee compatibility with AXIR. Like ARM for processors, we become an unavoidable technical standard. Every chip manufacturer seeking compatibility with our OS pays a licensing royalty.

- This includes new generations of processors: photonic, neuromorphic, ARM-based.
- *Advantage*: a model similar to ARM, generating revenue from every chip sold within the ecosystem.

With this combination, we capture value at every level of the chain: enterprise users, cloud providers, and hardware manufacturers



Simple Projection Example

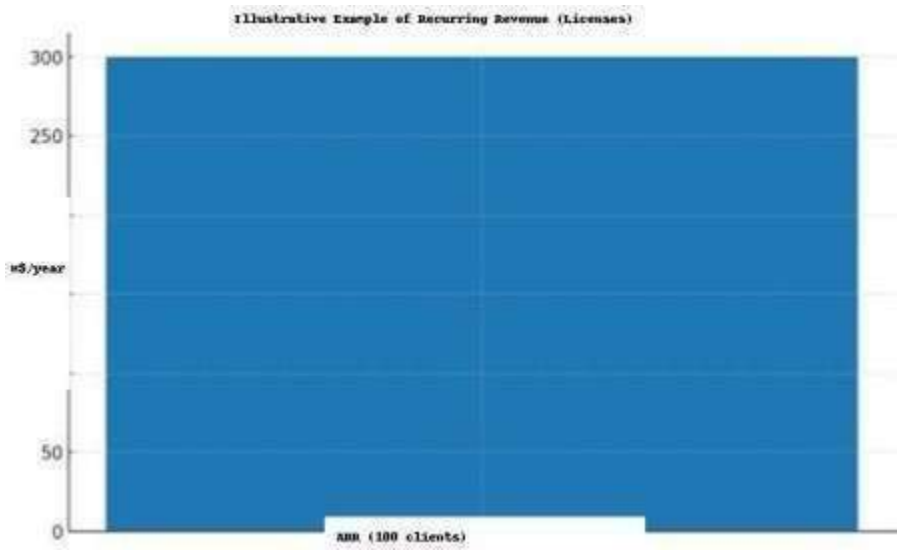
Let's imagine an initial base of **100 strategic clients** (banks, insurance companies, pharmaceutical labs, major technology platforms).
Each spends on average **$100M per year** on AI computing.

- With an average license fee set at **3%**, this represents about **$3M per client per year**.
- That equals **$300M in recurring annual revenue** from just the first 100 clients.

On top of that:

- **Cloud commissions**, directly tied to the explosive adoption of AI on AWS, Azure, and GCP.
- **Royalties from chip manufacturers**, with every new architecture becoming an additional revenue source.

In the long term, this model enables us to target **over $1B in annual revenue**, with **gross margins of 70– 80%**, typical of a highly scalable software model.



Illustrative Example: 100 clients × $3M ARR (licenses) = $300M/year, excluding commissions and royalties

Roadmap (2025–2030)

Phase 1 – MVP & Validation (2025–2026)

- Development of a first prototype: running a PyTorch model on both GPU and TPU through a single translation layer.
- Implementation of **AXIR v1**, the minimal version of our intermediate representation, serving as a pivot between CUDA and other backends.
- Integration of a **first cryptographic verification layer** to sign and authenticate model executions, ensuring traceable and tamper-proof results.
- Formation of the initial team (10–15 engineers specialized in compilers, GPU runtimes, and distributed systems).

- Objectives: demonstrate technological feasibility, publish a first benchmark, and launch academic partnerships (EPFL, ETH Zurich).

Phase 2 – Multi-Architecture Expansion (2026–2027)

- Extending support to five key architectures: Nvidia GPU, Google TPU, AWS Trainium, neuromorphic processors, and photonic processors.
- Evolution to **AXIR v2**: adding specialized dialects for hardware optimizations (Tensor Cores, systolic arrays, etc.).
- Enhancement of the **cryptographic trust layer** to include decentralized verification and secure proof-of-computation logs.
- Launch of first cloud partnerships with AWS, GCP, and Azure.
- Start of integration into AI pipelines of major labs and enterprises.

Phase 3 – Global Standardization (2027–2030)

- Adoption of **AXIOMOS** as the de facto standard in data centers and hyperscalers.
- Rollout of the royalties model: every new chip manufacturer implements AXIR compatibility to access the market.
- Publication of **AXIR v3**, stable and open, as a **neutral and cryptographically verifiable standard** — similar in role to ARM or POSIX.
- Final goal: to become the **Windows / ARM of AI** — an invisible yet indispensable layer enabling **trusted, universal computation**.

Summary

- Recurring revenues through SaaS licenses.
- Exponential growth via cloud commissions.
- Royalties as an “invisible tax” on every compatible chip.
- A clear roadmap:
  - MVP validated within 1 year,
  - Multi-architecture expansion within 2 years,
  - Global standardization (with cryptographic verification) within 5 years.

Founding Team

CEO – Pierre Seck

Founder of Axiomos. Built a working PoC in two weeks proving CUDA/HIP portability to Intel GPUs. Responsible for
strategic vision, fundraising, recruiting top talent, and building the AXIR ecosystem.”

CTO (To be recruited)

Desired profile: a recognized expert in distributed systems, compilers, and GPU/AI runtimes. Role: lead the technical roadmap, from MVP to global standardization.

Advisors – In progress

- Researchers from **EPFL** and **ETH Zurich**, two centers of excellence in AI and

systems.
- Former engineers from **Nvidia, Google Brain, and DeepMind**.
- Experts in next-generation architectures: neuromorphic and photonic.

Our team aims to be **global, interdisciplinary, and neutral**, with both an academic and entrepreneurial DNA.

Final Vision

The universal Layer for AI is not just another piece of software.

It is the future invisible infrastructure of global artificial intelligence.

Today:

- 70% of AI computing relies on CUDA/Nvidia.
- Every migration to a new chip costs millions of dollars and takes months of work.

Tomorrow:

- All AI models will be able to run on any hardware (GPU, TPU, Trainium, neuromorphic, photonic, edge devices).
- Enterprises will regain their freedom of choice: speed, cost, energy.

- AXIR will become a **global standard**, as invisible yet indispensable as Windows for PCs or ARM for smartphones.

Our Ultimate Goal

- Not a product: a **universal standard**.
- Invisible to the public, but **essential for the future**.
- The neutral layer that will allow AI to deploy everywhere: from cloud to connected devices, from healthcare to space.

We are not just creating software.

We are building the **invisible infrastructure of global artificial intelligence.**