

No Future for AI Agents: The Neural OS Will Replace Them All

[Assem Hijazi – Experience, Knowledge & Ideas](#)

I go everywhere these days — events, articles, podcasts, online talks — and everyone’s talking about Generative AI 🧐

And what’s the hottest buzzword?

Yep... **AI Agents** 🤖

People are obsessed.

“How do we build the perfect AI Agent?”

“Let’s connect our tools to the LLM!”

“Every department should create its own AI Agent!”

Even the big players — you know who I mean, no names here 😊 — are *officially guiding* companies to build multiple AI Agents, one per department.

Like seriously?

That’s the big vision? Let every team go build their own AI intern? 🧐

Meanwhile, while everyone’s busy designing their “dream agent,” I see **something way bigger** happening:

🚀 **Hardware is moving FAST.**

- Quantum computing is no longer a lab toy — it’s coming to *devices*.
- In-memory apps are no longer a theory — they’re the future.
- AI chips are landing inside your phone and laptop — and yes, some with LLMs embedded. (Hi Apple 🍏 🙌)

And that’s when I asked myself:

🧠 **What’s the future of operating systems?**

If AI Agents are running the show, do we even need operating systems?

Or... is the real future something much deeper — where the OS *becomes* the AI itself?

That’s where this story starts.

Let’s flip the narrative. Because trust me, the future won’t look anything like today’s AI Agent trend.

The Mission of Today’s AI Agent

Right now, AI Agents are everywhere. They’re like the “middle managers” of the GenAI revolution — sitting between the LLM (aka the brain 🧠) and the tools (aka the hands 🛠️).

Here’s what they do in a nutshell:

Role	What It Means
Interpret tasks	Turn your prompt (“Find me a candidate”) into structured steps.
Connect tools	Call APIs to interact with calendars, CRMs, emails, documents, and more.
Remember stuff	Keep context alive across multiple steps and interactions.
Orchestrate workflows	Coordinate between different tools or even other agents to get things done.

So far, so good.
They’re useful. They’re helpful. They’re smart.

But — and it’s a big **but** — this whole system assumes one thing:
That the **LLM is in the cloud**, the **apps are external**, and the **agent needs to glue them together**.

That’s the entire model.

It’s the digital version of walking across the room to pass a note between two people. It works, but...
come on. It’s 2025. You’ll see what I mean in a second 👁️

Neural OS – Redefining the Operating System

Here's where the real shift happens.

We're not just moving from the cloud to the chip — we're **rebuilding the concept of the operating system itself**.

Think about it: what does an OS do today?

- It runs apps
- It manages memory
- It talks to hardware
- It acts as the environment for user interaction

Now imagine this:

 **The AI chip doesn't just support the OS — it *becomes* the OS.**

The AI Chip = The New Digital Brain

This chip isn't just running isolated tasks. It's:

- Holding a **full version** of the LLM or something even more advanced
- Acting as the **central digital brain** of the entire system
- Serving as the **communication hub** between apps, tools, the OS, and external updates

The apps talk to the chip.

The OS talks to the chip.

The chip *is* the AI — and everything else is just context.

Even the traditional CPU, GPU, RAM, and storage are controlled by this AI chip — it's the **orchestrator**, the **logic**, the **reasoning unit**, and the **manager** of your entire device.

Embedded Apps? Already Inside.

Not only is the chip the digital brain, but it also:

- Hosts **embedded apps** (mini-tools, workflows, memory functions)
 - Interacts **natively with the OS UI and system tools**
 - **Learns and evolves**, syncing with the main LLM (e.g., OpenAI) for periodic updates, just like a Windows update — except now it's your AI getting smarter.
-

And Why Is This Even Possible?

Simple:

Because **hardware evolution is outpacing our expectations**:

- **Quantum computing** is becoming portable
- **In-memory applications**
- **Local LLM compression** and optimization are accelerating
- And AI chips are **no longer assistants — they are control centers**

This isn't just an evolution of AI agents — it's a whole new species of computing.
The OS won't run beside the AI — it will be built *around* it.

AI-Managed Apps – The End of APIs

Here's a question that changes everything:

If the AI chip is already managing all apps, handling storage, and orchestrating everything across the OS... why do we still need APIs?

That's the magic of where we're heading.

From APIs to Awareness

Today, APIs are necessary because:

- Apps live in silos
- They don't "know" about each other
- They need translators to talk to each other and to the system

But when the **AI chip** is the central brain:

- It **knows every app**, system tool, file, process, and preference
- It **doesn't need to be told** how to connect A to B — it already understands the intent and context
- It can **connect apps directly**, in real time, without REST calls or JSON payloads

The chip becomes a **context-aware router**, not a passive listener. Every app is "seen" by the AI chip — and managed through understanding, not hard-coded interfaces.

The App Layer Is No Longer a Bunch of Boxes

In the current world:

- Apps are isolated systems
- APIs are bridges we build between them

In the Neural OS world:

- Apps are **native thoughts** in the mind of the AI chip
- No bridge needed. The brain just "thinks" the connection into existence

For example:

- Want to send an invoice from your accounting app to your CRM?
- No API call.

- The AI chip knows both apps, understands your intent, and executes the task **natively across apps, without the apps needing to talk directly.**

That's not middleware.

That's intelligence.

Smooth Hardware Integration

Because apps are now:

- Built for AI-native environments
- Aware of the AI chip as the control layer
- Designed to **leverage chip-level performance and context**

You get:

- Ultra-smooth UX
- Instant cross-app actions
- No network delay, no protocol mismatch

It's not integration. It's **unification.**

So now ask again:

Do we need APIs when apps are integrated into the AI chip's brain and the OS itself is AI-driven?

Probably not.

Not in the future we're heading toward. 🚀

Why On-Premise LLMs Will Rise Again

Let's take a step back and look at the bigger picture.

Right now, everything is **cloud-obsessed**.

- LLMs? Hosted on massive cloud servers.
- AI agents? Tied to internet access and server-side APIs.
- Your data? Floating around in a thousand rented silos.

The message is: “You can't do real AI unless you're in the cloud.”

But here's the truth nobody's admitting...

The future belongs to the device.

And the LLM will live in your hands.

⚡ Why On-Premise Is Coming Back (With a Vengeance)

Let's connect the dots:

- Devices are getting smarter
- AI chips are getting smaller
- Hardware is becoming powerful enough to host models **locally**
- Quantum performance + compressed LLMs = 🌟
- AI chips are becoming the **brain of the device** — not just a co-pilot, but the pilot

So what happens when:

- You don't need to call OpenAI or Anthropic to process a prompt?
- Your **own device** handles it instantly?
- You have a **personal digital brain** running your life — securely, locally, and fast?

That's **on-premise LLMs**.

Not just for enterprises. For **every user**.

And yes — you already wrote about this here:

👉 [On-Premises Personal LLMs for Each User](#)

The AI Chip = Your Local LLM

This chip:

- Hosts the full (or partial) LLM
- Stores your memory, your apps, your preferences
- Knows your patterns, habits, and context
- Connects only occasionally to the cloud for updates — just like your OS does now

It's fast.

It's personal.

It's **private**.

And it completely rewrites the story of cloud dependency.

So... Is Cloud Dead?

No. But it's changing roles.

The cloud becomes the **sync point** — not the core.

It's where your chip goes to update, learn, and evolve — but not where it lives.

This saves:

-  Cost (goodbye massive cloud bills)
-  Energy (localized efficiency > server farms)
-  Privacy (your data stays with you)

And best of all? You get **instant, real-time intelligence** — not laggy, network-dependent AI.

So when we say “the LLM moves on-premise,” we're not just upgrading architecture — we're giving users real control and real intelligence that doesn't depend on a data center.

Conclusion – Let the Devices Speak

We started with AI Agents.

We talked about how they connect tools, how they work with the cloud, and how they've become the trend of the moment.

But trends fade.

And what's coming isn't just a new tool — it's a **paradigm shift**.

The Device Is the Platform

We're entering an era where:

- **The AI chip becomes the brain** of your device
- **The OS becomes a Neural OS**, built around intelligence, not just execution
- **The apps become native thoughts**, not separate entities
- **The cloud becomes the update server**, not the host of your life

No more APIs.

No more bridges.

No more middle layers.

Everything becomes **one unified, real-time ecosystem** — running locally, thinking fast, learning continuously.

It's not about killing the AI Agent.

It's about **evolving past it**.

The Real Future?

When you power on your device —
you're not booting into an operating system.

You're waking up your **digital brain**.

One that runs your apps.

Understands your intent.

Learns with you.

Protects your privacy.

And evolves **right by your side**.

This is the **future of computing**.

And it's not in the cloud.

It's in your pocket.

Let the devices speak.

The Neural OS is coming. And it's already thinking.

  — **Assem Hijazi**

[!\[\]\(6e934896f25e6ce1b0dbb50c23abc197_img.jpg\) Read my full vision on On-Premise LLMs](#)

Everyone's talking about AI agents lately...

Build an agent.

Connect it to tools.

Let every department have their own AI assistant.

Cool idea, right?

Well... maybe for *now*.

But what if I told you this is just a **temporary phase**?

The real future? It's already taking shape — and it's **not in the cloud**.

It's **in your device**.

We're entering an era where:

- 🧠 Your device will have its own **AI chip**
- 🧠 The AI chip *is* the operating system's brain
- 🧠 Your apps won't need APIs — they'll just “talk” to the chip directly
- 🧠 The OS itself will become a **Neural Operating System**

No more bridges.

No more cloud bottlenecks.

No more over-complicated agents.

When you power up your device...

You're not booting an OS.

You're waking up your **digital brain**.

Let the devices speak.

The future isn't about connecting agents.

It's about embedding intelligence directly into the core of your system.

— Assem Hijazi

#AI #GenerativeAI #ArtificialIntelligence #FutureTech #QuantumComputing #NeuralOS #DigitalBrain
#LLM #AIChip #OnPremiseAI #AIArchitecture #EdgeComputing #TechTrends #OperatingSystem
#AIAgents #NextGenAI #AIHardware #InMemoryComputing #FutureOfDevices #AssemHijazi