2 SEPTEMBER 2026

From AI factory to AI production

The Netherlands is rightly investing in a national AI factory. At the same time, new research by ABN AMRO shows that traditional data centre development is increasingly clashing with limits to power, grid capacity, water, space and public support. It is precisely between these two developments that there is room for a different approach: central AI development where scale is needed, and modular production inference where data, energy and user come together.

ABN AMRO's Group Economics exposes a fundamental problem in its analysis of 27 August 2026. The demand for data center capacity continues to grow strongly due to AI, but the traditional way in which we realize that capacity is increasingly clashing with limits to electricity, grid capacity, space, water, materials, permits and public support.

ABN AMRO expects global data centre capacity to almost double between 2025 and 2030. At the same time, the report states that the strategic challenge for Europe is not primarily to simply build more data centers, but to combine digital growth with energy security, social acceptance and competitiveness.

I think that's an important starting point. Because at the same time, the Netherlands is rightly investing heavily in its own AI capacity. The Dutch AI Factory in Groningen will create a national infrastructure for sovereign AI, with an AI supercomputer, a centre of expertise and support for companies, knowledge institutions and public organisations.

These two developments may seem contradictory at first glance. On the one hand, we urgently need more AI computing power. On the other hand, it is becoming increasingly difficult to physically fit large concentrations of computing power.

In my opinion, this is exactly where one of the most important infrastructure questions for the coming years lies: should we organize all future AI compute in the same way?

My answer is no. For training large models, very large, central and highly coupled AI systems are logical and necessary. But after training, another market begins: inference. The daily use of AI in enterprises, hospitals, governments, industry, logistics, energy supply and other operational processes.

Completely different infrastructure principles may apply to that market. That's where I see a possible role for SOLOAI, in combination with Zirrow. Not as a competitor to the Dutch AI factory. Especially as a possible next link.

The Dutch AI factory is more than just a large computer

To properly understand this positioning, we first have to get rid of the image that the Dutch AI factory is primarily a supercomputer project.

The Dutch AI factory consists of two important parts: a specialized AI supercomputer and a center of expertise. The supercomputer is intended for developing, training and fine-tuning advanced AI models. The centre of expertise supports organisations in developing usable and responsible applications.

That is a fundamentally different role than a commercial cloud environment. Access is primarily focused on development, training and R&D. This naturally raises the follow-up question: what happens when an AI model has been developed, trained, tested and validated and then has to become part of the user's business process 24 hours a day?

That's where another infrastructure issue begins.

From AI Factory to AI Production

Suppose a Dutch manufacturing company develops a model for visual quality control via the AI factory. Or a hospital develops its own medical language application. A logistics company builds an AI model for planning. A government organization develops a sovereign document assistant. Or an industrial company develops a model for predictive maintenance based on sensors and historical production data.

In the development phase, central access to powerful AI compute is very valuable. But when the model is successful, the demand changes.

Then it is no longer primarily about: can we train this model? But how do we run this reliably, safely, economically and scalably within our daily operations?
Then latency, availability, cybersecurity, energy consumption, SLAs, ownership, costs, operational data and lifecycle management become important. And for many applications, the question of whether organizations want to permanently send their operational data to an external cloud environment also becomes relevant.
For regulated industries, industry, healthcare, government, financial services and organizations with high-quality intellectual property, this is not just a technological question. It can be a strategic, legal, and procurement question.
This creates a missing physical bridge between AI development and production inference.

NLAIF → SOLOAI → klant

A possible architecture could therefore go much further than just Dutch AI factory → AI model.
The full chain can be: Dutch AI factory → development → training → fine-tuning → validation → SOLOAI staging → deployment → local/private inference → business value.
Within that chain, the Dutch AI factory will continue to do what it was set up to do: provide access to advanced compute, knowledge, expertise, experimentation and development.
SOLOAI then focuses on another part of the lifecycle: making inference capacity physically operational.
That distinction is important. SOLOAI should therefore not be positioned as 'we also have an AI data center' and certainly not as 'we can replace the Dutch AI supercomputer'.
The stronger positioning is: SOLOAI can become the physical deployment layer between a successful AI development and production inference at the organization that actually uses the AI.
Dat is in essentie compute mobility from innovation to production.

Train centrally. Deploy where it makes sense.

This creates a much more interesting European AI architecture.
Not every organization needs to build a training cluster on its own. Large models can be developed and trained centrally within NLAIF, other European AI Factories or commercial AI infrastructure.
But once developed, the model can then run where it makes the most operational sense: at a company, on an industrial site, in a regional AI hub, in a healthcare cluster, next to a large data source, in a sovereign cloud environment or at a location where electricity, fiber optics and a good destination for residual heat are available.
De architectuur wordt daarmee: train centrally → validate → deploy selectively → infer locally or regionally.

Not because local is always better, but because the optimal location for training is not automatically the optimal location for inference.

Inference Changes the Economic Logic of AI Infrastructure

A strong law of scale applies to very large training. Thousands of accelerators need to work together at the same time. Network architecture, interconnects, storage throughput and very high power densities are essential.

Inference is much more varied. An inference workload might include a knowledge assistant that searches through documents, a computer vision model that monitors production lines, agents that automate administrative processes, or an enterprise LLM that supports hundreds or thousands of employees.

Not every workload needs the same GPU, the same latency, or the same power density. Therefore, we must be careful that the infrastructure for the inference market is not designed according to the principle: more GPUs, more MWs and more kW per rack is by definition better.

For SOLOAI, the opposite principle applies: less is more — as long as performance, availability and future-proofing are guaranteed.

The current design philosophy assumes modular configurations in the range of approximately 0.3 to a maximum of 1.2 MW IT load. In addition, 1.2 MW is a maximum workload envelope, not the total grid connection capacity. Very high rack densities should also be seen as a design and qualification envelope and not as an end in itself.

An infrastructure can be prepared for very high densities without every customer having to install that capacity from day one.

Why this modular approach is becoming relevant right now

The Dutch AI factory also shows why this issue is becoming important. The AI hardware market is changing at lightning speed. Hardware is becoming more powerful, more expensive, scarcer and also changes generation faster than the physical infrastructure around it.

This illustrates a structural problem of the entire AI market: compute technology changes much faster than physical infrastructure.

A GPU generation can be economically overtaken within a few years. A well-designed electrical installation, construction, cooling infrastructure and heat interface should last much longer.

It is therefore illogical to let an entire physical asset age along with the installed IT every time.

SOLOAI is developed precisely on that principle: the physical infrastructure layer must be able to survive multiple generations of compute. Rack technology, GPUs, CDUs, direct-to-chip cooling, immersion technology, sensors, and electrical components must be able to be replaced or requalified within controlled interfaces without having to redesign the entire asset.
This does not make the infrastructure hardware-independent in an absolute sense. Every hardware change must be technically validated. But the design can be made accelerator and technology flexible.

From stranded capacity to progressive provisioning

This also has financial consequences. Traditional data center development has the risk of pre-installing large amounts of power, cooling, and electrical equipment while the actual IT load comes later.
With rapidly changing AI hardware, that risk increases. That is why progressive provisioning for inference is interesting.
For example, a SOLOAI can grow from an initial partial load when workloads, customers or contractual capacity actually become available. The fixed infrastructure is being prepared for further growth, but active components do not all need to be pre-installed.
The question then does not become: how many GPUs can we buy today? But: how do we design an infrastructure in which we can always place the most economically suitable compute during the next ten or fifteen years?

This is precisely where the ABN AMRO report fits in

ABN AMRO shows that traditional data centre development is increasingly limited by location.
According to the report, grid congestion is now one of the most important selection criteria. In the European comparison, the Netherlands has a very high risk of grid congestion, while at the same time the country scores exceptionally well on digital infrastructure and digital competitiveness.
That seems like a paradox. The Netherlands has exactly the companies, knowledge, connectivity and digital infrastructure needed to use AI economically. But it is precisely in the places where this economic activity is concentrated that large new electrical capacity becomes more difficult to obtain.
A possible solution is therefore not only to look for larger data centers in locations where tens or hundreds of megawatts are available.
For part of the inference market, we can also reason the other way around: where are available energy, fiber optics, users, data and potential heat consumers — and how much inference capacity is appropriate?
That is a different approach to location development.

SOLOAI does not create grid capacity

There is an important nuance to this. A modular AI infrastructure does not magically solve grid congestion. If electricity is not available at a location, SOLOAI cannot create a megawatt from scratch.

That is not the claim either.

The potential value lies in being able to adapt to different locations and in preventing each AI application from being made dependent on one huge central connection.

SOLOAI can be designed to take into account available grid capacity, local renewable production, battery storage, fiber optics, cooling conditions, and local heat demand.

This shifts the model from: find 100 MW and build around it, to: find the right combination of workload, energy, location and useful output.

For inference, that can become an important difference.

From NLAIF staging to production at the customer

One of the most interesting next steps is the concept of an AI Deployment Zone. Not as an already agreed part of NLAIF, but as a concept that could be explored with the ecosystem.

Suppose an organization has developed a successful AI application through the NLAIF center of expertise. The model has been tested. The security architecture has been determined. The need for inference is well known.

Then a SOLOAI configuration can be staged physically or operationally. There, hardware, software environment, security, workload, cooling, energy profile, monitoring and performance are tested together.

The same standardised infrastructure is then brought to the final production site or built elsewhere according to the same validated configuration.

De gedachte is dan: develop → train → validate → stage → deploy → operate.

This makes AI compute a deployable asset, as it were. The innovation is not only in the movable box. The innovation lies in the standardized process by which an AI workload is brought from a development environment to operational production.

The Dutch AI factory can therefore create more economic impact

This could also be interesting for NLAIF.

Ultimately, the success of an AI Factory should not be measured solely by the number of models trained, the number of GPU hours used, or the number of organizations that have been granted access.

The real economic and social impact only arises when applications actually come into production.

A model that remains on the shelf after a pilot yields little productivity growth. A model that becomes a safe and economical part of a hospital, factory, government service or SME is.
Therefore, in addition to the classic AI Factory, the concept of AI Production Factory could also become interesting. Not as a new building, but as a chain.
The Dutch AI factory creates successful AI applications. A deployment layer brings them to production. The customer uses them. And then we measure whether they actually deliver value.
That is ultimately the step from innovation to economic earning capacity.

Moreover, there remains a clear public and commercial separation

It is precisely the division of roles that can remain relatively pure.
NLAIF has a public and strategic function around AI expertise, access to compute, research, training, fine-tuning, validation, responsible AI and supporting innovative organizations.
SOLOAI can provide a commercial physical infrastructure layer for production, deployment, operation, maintenance, upgrades and SLAs.
The proposition does not have to be: 'NLAIF should buy SOLOAIs.'
A much more interesting question is: can the NLAIF ecosystem develop a standardized pathway that will take successful Dutch AI applications to sovereign production inference after validation?
And can SOLOAI become one of the physical deployment options?
In my opinion, that is a much stronger conversation.

Zirrow can become the evidence layer across the chain

But physical deployment alone is not enough.
We need to avoid building a new generation of AI data centers and ultimately measure how much electricity enters the building again.
PUE remains relevant, but says little about whether IT actually uses the energy productively.
With inference, this becomes increasingly important. Two infrastructures can have the same PUE, while one provides significantly more usable AI output per kWh than the other.
That's where Zirrow comes in.
Zirrow can be positioned as an independent Real-Data Evidence Layer that connects data on energy, water, costs, utilization, governance, and workload performance. The starting point is actual operational data rather than just averages or theoretical nameplate values.
That can become particularly interesting in combination with NLAIF.

From model benchmark to operational benchmark

During development and training, we measure all kinds of AI indicators: accuracy, latency, model quality, memory usage, tokens per second, and GPU performance.

But after deployment, a different set of questions arises. What does this workload cost per day? How much energy is actually used? What is the energy consumption per transaction or per million tokens? How much of the GPU capacity is actually being used? When is hardware largely idle? What is the latency under real business tax?

What happens to performance when power is temporarily capped? How much heat is produced? How much of this is technically absorbed? And how much of it is actually used by an external customer?

This allows Zirrow to provide a continuum of evidence: from technically validated model to demonstrably efficient production.

From PUE to useful AI output

In digital infrastructure, we ultimately need to focus more on useful output per resource used.

For AI, for example, this could mean — where the interfaces allow — that energy and costs are related to workload, transaction, inference, tokens, latency and SLA.

That is an important next step. Because an AI infrastructure that spends little energy on cooling but whose accelerators do nothing useful much of the time can hardly be called optimally sustainable.

An efficient machine that performs inefficient work remains wasteful.

Residual heat must also be proven

ABN AMRO also emphasises the increasing importance of water use, energy, space and local support.

The same applies to residual heat. Many sustainability stories say that heat is 'available for reuse'. But being available and actually being used are two different things.

An evidence layer must therefore be able to distinguish between heat that is captured, heat that is transferred to an external interface and heat that is then demonstrably put to good use.

That seems like a detail. But that is exactly where the difference between sustainability marketing and evidence lies.

This fundamentally changes the choice of location

When we combine SOLOAI and Zirrow with ABN AMRO's learnings, we ultimately create a different method for selecting AI locations.

ABN AMRO assesses locations on electricity availability, grid congestion, electricity price, renewable energy, digital infrastructure, labour market, land, water stress, climate, regulations, data sovereignty and proximity to users, among other things.
For distributed inference, you can add an extra dimension: where does a MW AI compute provide the most economic and social value?
A location where 1 MW is available next to an industrial company with large local AI workloads and a permanent heat demand may be more economically interesting than a location where 10 MW of cheap electricity is available but all data, users and heat consumers are hundreds of kilometres away.
Not for every workload. But for some it is. And that distinction will become increasingly important in the coming years.

The Netherlands has an interesting starting position for this

Despite the very high grid congestion, ABN AMRO places the Netherlands high in its European comparison of location criteria for data centres. The Netherlands combines its problems on the electricity grid with an exceptionally strong digital infrastructure and proximity to users and economic activity.
We therefore do not need to copy Scandinavia. For hyperscale, cheap space and abundant renewable electricity can be a huge advantage. But for part of the inference market, the Netherlands has other strong cards.
We have a very high digital density, strong fiber optic connectivity, universities, SURF, TNO, a large AI ecosystem, high-tech industry, healthcare institutions, financial institutions, logistics, energy clusters, greenhouse horticulture and a lot of economic activity at a relatively short distance from each other.
And soon NLAIF as an important national AI development infrastructure.
Together, this can result in a completely different infrastructure strategy than just trying to build the largest AI campus.

A Dutch sovereign AI stack

The Dutch AI factory should therefore not be judged on whether the Netherlands can beat the United States or China in absolute GPU numbers. That is the wrong benchmark.
The strength lies in the combination. SURF brings infrastructure and supercomputing. TNO brings applied AI and the centre of expertise. AIC4NL connects the ecosystem. Cooperation North connects regional users and economic development. The Dutch AI factory brings dedicated AI compute.
Add to that a private and regional production-inference layer and something much more strategically interesting emerges: a Dutch sovereign AI stack.
Not one computer. But a chain.

From national AI infrastructure to the European model

The Dutch AI Factory is part of the European network of AI Factories. Therefore, when the concept of standardized production deployment works, the same model can theoretically be replicated.

European AI Factory → training and development. SOLOAI deployment infrastructure → productisation and staging. Regional or private location → production inference. Zirrow → independent operational evidence.

NLAIF can then be a Dutch pilot environment for a concept that can ultimately be applied more widely within Europe.

This is much more interesting than presenting SOLOAI only as a transportable data center. The box is the physical product. The real proposition is the transition: from national AI compute to local AI operations.

The competitor of SOLOAI is therefore not NLAIF

That brings me to the most important positioning choice.

SOLOAI should not try to beat the Dutch AI factory. Not even the European AI Factories or Gigafactories. And also not trying to become a smaller hyperscaler.

Those markets have different economies of scale and a different function.

SOLOAI focuses on private, sectoral and regional infrastructure for enterprise inference, RAG, agents, regulated AI, AI-assisted engineering and selected fine-tuning, with a modular physical deployment.

The AI Factory is mainly located upstream. SOLOAI can position itself downstream.

That makes cooperation much more logical than competition.

The real competitor is the old way of thinking

The fundamental competitor may not even be a specific company at all.

The idea is that every new digital application eventually requires a larger building, more land, more contracted capacity, more cooling, more hardware, and that we will then see what happens to the energy and heat.

For the next generation of AI infrastructure, it's better to start exactly the other way around.

What workload do we want to run? What performance is really needed for that? Which dates can go where? How much capacity should be available immediately? What capacity can be added later? What energy is available locally? Which workloads are flexible? Which cooling suits the hardware? Can the heat be made useful locally? Can the physical infrastructure last for multiple generations of IT? And can we then measure all that performance objectively?

Only then do we determine how much infrastructure we really need.

More AI value per available megawatt

This brings me back to ABN AMRO's conclusion. Europe's challenge is not just to build more data centers. We need to combine digital growth with energy security, social acceptance and competitiveness.

For me, that means that we should not only discuss how many MW AI the Netherlands needs, but especially about how much economic, social and digital value we can get from every available MW.

The Dutch AI factory can play an essential role at the beginning of the chain. There, AI can be developed, trained, tested and validated.

SOLOAI can then be explored as a potential physical infrastructure layer that brings successful workloads from that development environment to production.

And Zirrow can demonstrate what actually happens next: how much energy is used, how well hardware is used, what performance is delivered, what the workload actually costs, how much water is needed, how much heat is actually reused, and how much useful AI output ultimately compares to the resources used.

This creates a completely different approach to AI infrastructure.

Niet: train everything centrally and keep adding capacity.

But: develop centrally where scale is needed. Deploy selectively where production makes sense. Measure everything that matters.

Of nog korter: Train centrally. Deploy intelligently. Prove the performance.

In my opinion, it is precisely there that a distinctive position can arise for the Netherlands — and possibly also for SOLOAI and Zirrow.

Not by wanting to build the largest AI factory in Europe. But by developing an infrastructure model that actually, efficiently and demonstrably converts the knowledge and computing power of the AI Factory into production AI at European organizations.