cognizant

Physical AI: Engineering intelligence in the real world

What it takes to move intelligent systems from pilot success to production-scale reality.

Contents

Executive summary

Physical AI is beginning to transform the world around us, at ever greater speed, moving from experimentation into execution. Across industries, intelligence is no longer confined to digital workflows or cloud-based analytics. It is increasingly being engineered into physical systems, from vehicles and factories to intelligent products and critical infrastructure—where AI must act in real time, operate safely and perform reliably under real-world constraints.

While physical AI is already driving higher levels of productivity, performance and putting new levels of intelligence at our fingertips, its adoption is far from straightforward. To function effectively and reliably in the real world, enterprises must overcome a set of persistent challenges spanning complex technical requirements, cost concerns and public trust.

To help companies understand how to resolve these challenges, Cognizant has created this guide, which examines the major barriers and pitfalls enterprises face when implementing and using physical AI, including:

- **AI not being embedded at the outset**
- **Practical constraints related to hardware, latency and safety**
- **Rising cost pressures across compute, infrastructure and operations**
- **Low confidence in outcomes and the ability to scale beyond pilots**
- **Organizational transformation required to operationalize intelligence at scale**

The paper examines these barriers in depth and explores how enterprises can move beyond pilots and fragmented initiatives to industrialize AI in the real world.

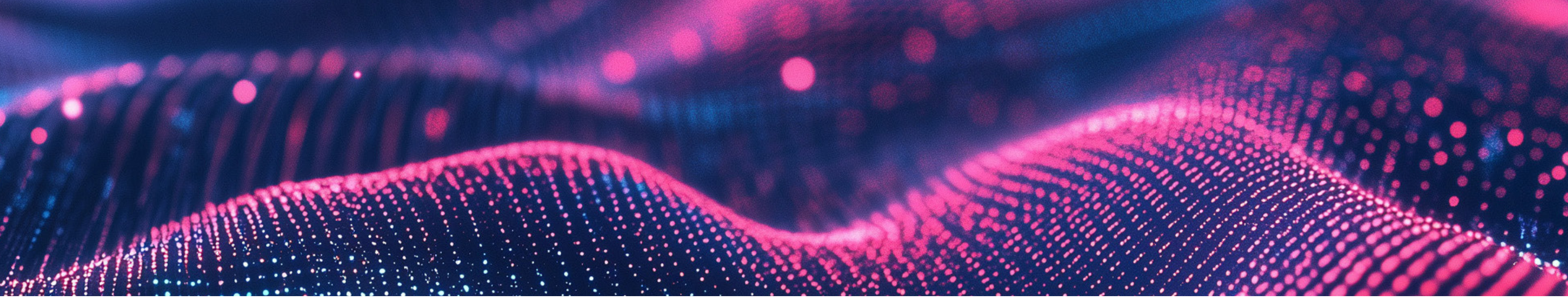
Drawing on real-world experience across multiple industries, it outlines pragmatic principles that help organizations bridge the gap between vision and execution—where the true bottleneck of AI adoption now lies. These insights are grounded in hands-on physical AI programs across manufacturing, automotive, semiconductor and life sciences environments—reflecting the practical realities of deploying intelligence in production systems, not just pilots or proof of concepts.

To help businesses and technology leaders unlock value from physical AI, the paper also highlights five critical steps organizations should take:

- **Unlock the potential of edge engineering**
- **Using high-fidelity simulations to derisk decisions**
- **Identifying operational efficiencies**
- **Taking an incremental and scalable approach to adoption**
- **Managing the organizational changes**

Finally, the paper illustrates how these principles are applied in practice across key sectors through real-world examples. Together, they demonstrate what organizations can unlock when intelligence
is engineered into physical systems: higher productivity, reduced technical debt, more resilient operations and scalable AI-enabled products and services.

This paper also describes how Cognizant works with enterprises in these industries, drawing on its engineering expertise, applied data capabilities and global operating experience to help organizations move from AI ambition to practical execution.



Introduction

We're at a pivotal moment as AI steps out of the cloud and into the physical world. Increasingly, AI-enabled machines and systems are no longer limited to analyzing data after the fact. They can perceive what is happening around them, interpret context and act autonomously in physical environments.

This shift is already visible across industries. Automation is moving deeper into factories and warehouses. Autonomous and semiautonomous vehicles are becoming part of everyday transport. Intelligent consumer devices are embedded in homes. In healthcare, robotic systems are supporting complex medical procedures. Across these environments, AI is no longer a layer on top of operations, it is becoming integral to how physical systems operate.

The scale and momentum behind this shift are significant. Gartner has identified physical AI as one of the top strategic technology trends expected to shape enterprise priorities over the next five years.[1]

> The humanoid sector of the robotics market is already projected to reach **$200** billion by **2035** as physical AI adoption soars in labor-intensive sectors.[2]

The reality is that these robots, along with autonomous vehicles and quadrupeds, are already among us. They have entered workspaces and are increasingly operating in the world around us.

Their adoption is slowly transforming industries and physical environments as companies bring full automation into factories and warehouses; autonomous automobiles become the norm on our streets; intelligent consumer devices permeate domestic homes; and the use of robotic devices for medical procedures grows.

Organizations are already rethinking their business models as AI becomes fundamental to operational strategies and investment planning. It is leading businesses to place huge bets on smart devices and hyperconnected industrial sites, which are enabled by digital twins.

As physical AI moves from theory to reality, however, organizations will need to manage many practical considerations—such as the physical constraints on edge devices, energy consumption, regulatory compliance, environmental considerations and more. As AI workloads move closer to the edge, energy efficiency and carbon impact become primary engineering considerations that must be optimized alongside performance, safety and scale.

These constraints show up differently by industry, but the pattern is consistent. Leaders face a familiar set of questions:

- **Where should intelligence run: cloud, edge or embedded?**
- **How do we prove decisions are safe before they touch operations?**
- **How do we move beyond pilots to repeatable impact?**
- **How do we build trust, not just in the model, but in the outcome?**
- **And how do we scale adoption without creating new technical debt, operational fragility or unsustainable energy demand?**

This is why a cross-industry perspective matters. Physical AI is not a single technology decision. It is a system engineering challenge that blends embedded and edge engineering, applied data, simulation and digital twins, and the operating model required to run AI in production. It also depends on ecosystems: Successful deployments combine capabilities across hardware, platforms and domain solutions, with orchestration and integration that make the whole system work in practice.

Innovative solutions will need to be developed to negotiate the shift from traditional practices and legacy technologies to this new operating model. The transformative impact of physical AI will also reshape workforces. New skill sets will be sought to service and work alongside intelligent machines that are capable of operating independently.

This is a transition that needs to be carefully planned and executed. If physical AI is to be used across organizations, deployments need to be proven to be safe, their value broadly understood and usage accepted by the workforce. Investments, which can be significant, will need to demonstrate clear and compelling returns.

This paper looks at how businesses can successfully manage this transition. The barriers and pitfalls they will face at both a device and operational level, and the solutions needed to overcome them. We hope you find the guidance and examples helpful as you embark on your journey into physical AI.

Defining physical AI

Physical AI refers to machines with the ability to perceive, understand, reason and act autonomously in real-world environments. The examples span robots, drones, self-driving vehicles, intelligent medical devices and industrial automation. Digital twins, simulations and synthetic data help train, test and derisk these systems before deployment.

The major difference with traditional AI, where humans are still required to act on insights in real-world scenarios, is that physical AI empowers the machines to perform these tasks independently. This requires devices operating at the edge to grasp the laws of nature. For example, they need to compensate for forces such as gravity, friction or inertia. So, in addition to inferring what's happening in the local environment—using cameras, sensors and sophisticated edge computing— physical AI devices need to be highly trained in cause and effect.

Deployments typically integrate multiple computing layers, including edge computing for fast local decisions, cloud environments for training models and enabling continuous improvement, and simulation to validate performance before rollout—with physical technologies that enable machines to make fast, accurate decisions within dynamic real-world scenarios.

In practice, leaders are increasingly designing distributed architectures that balance where models are trained, where decisions are made and how systems are monitored and improved over time.

Examples of physical AI in real-world deployment

Robots

AI-powered systems that perceive, grasp and act on physical objects in dynamic environments.

Drones

Autonomous aerial systems that use AI for navigation, surveillance, inspection, and mission execution in live operational settings.

Self-driving vehicles

Intelligent mobility systems that continuously sense, process and respond to live traffic conditions.

Intelligent medical devices

AI-enabled clinical systems that monitor, diagnose, and support treatment decisions within regulated healthcare environments.

Industrial automation

AI-enhanced machinery and control systems that optimize manufacturing and operational processes through real-time sensing and autonomous adjustments.

For technology leaders, the defining challenge is not experimentation, but engineering these systems to operate safely, cost effectively and at scale in production.

Barriers and pitfalls

Physical AI can unlock major gains in productivity, safety and service performance, but it typically requires meaningful investment such as in equipment, connected infrastructure, plants, data foundations and the engineering effort to run AI reliably in production. Given the scale of commitment, organizations can't afford initiatives that stall at pilot stage or create new operational risks.

Understanding the likely complications is therefore essential at the outset. If projects hit any obstacles, progress will be curtailed. Pilots can get stuck in purgatory, rollouts will be encumbered and future market competitiveness compromised.

The reality is that physical AI fails less often because the vision is wrong, and more often because execution is underestimated. Real-world deployments introduce constraints that don't show up in digital pilots: safety and compliance, latency and connectivity, legacy integration, cost-to-operate, workforce adoption and the ability to sustain models over time.

To avoid fragmented initiatives and "pilot fatigue," leaders should watch for five common barriers and pitfalls:

1. AI not being embedded at the outset

Physical AI is still emerging, so it is inevitable that many organizations are trying to lay a new intelligence onto products, plans and operations that were designed before real-time autonomy was a requirement. Such legacy architecture—often centralized, siloed and not designed for real-time autonomy, can prove to be a major obstacle to physical AI adoption. Physical AI, by contrast, requires a distributed architecture that connects edge, cloud and physical assets as one cohesive system.

Organizations in the design and development phase should consider physical AI requirements early. Doing so helps them avoid fragmented engineering across hardware, firmware, cloud and applications, and ensures assets and operations are connected so that insights can flow without interruption.

To embed physical AI from the outset, leaders should design for:

- End-to-end system integration across devices, edge and cloud, with clear ownership and interfaces to ensure cohesive system performance.
- Connected operations and strong data foundations that eliminate silos and enable real time, usable insights.
- Safety, trust and sustainability requirements treated as core design constraints rather than after-the-fact checks. This includes engineering efficiency and carbon-aware AI workloads across training and inference, especially as intelligence moves to distributed edge environments.
- Lifecycle readiness, including model monitoring, updates and governance over time to ensure long-term reliability and compliance.

This is already evident in industries modernizing architectures to move faster. For example, in the automotive industry, where new-age manufacturers—with software-defined architectures—are able to deploy new solutions faster.

The lesson is transferable: The earlier physical AI requirements shape design choices, the easier it becomes to scale without creating new technical and operational debt.

If companies don't take these steps it can lead to disjointed workflows, operational bottlenecks and suboptimal performance. At product level, it also can mean technical debt, with products needing to be rearchitected to meet new commercial opportunities.

Gartner has also coined the term "AI debt," which it says, if left unaddressed, will impact an organization's ability to pivot, scale and innovate. It forecasts that businesses taking a more proactive approach to AI debt will mature **500% faster** over the next three years.[3]

2. Practical constraints

When machines operate in the real world, there are several practicalities that need to be addressed. For example, when quadrupeds run around a warehouse or autonomous vehicles move around streets, constant connectivity with the cloud is not guaranteed—they can hit network blind spots. This is not acceptable in safety critical situations, such as when a pedestrian steps in front of an autonomous vehicle. In these environments, milliseconds matter. Decisions must be deterministic, predictable and provably safe—not dependent on round-trip latency to the cloud.

To reduce latency, decisions need to be made at the edge. Without the scalable computational power of the cloud, however, trade-offs are inevitable—requiring robust edge engineering to balance accuracy, responsiveness, safety and cost so AI can make reliable decisions locally that are accurate enough to fulfil the intended use case safely.

Edge devices also face hard constraints around compute capacity, memory, power consumption, thermal limits and form factor. Because physical AI depends on real-time inference at the edge, these constraints force deliberate trade-offs in model size, update frequency, hardware selection and inference strategy.

3. Cost concerns

While the capital costs of physical AI are substantial, the operational costs can be too. For instance, a computer vision solution that is reliant on the cloud will incur significant bandwidth costs and API call fees. These costs can mount quickly and make projects cost-prohibitive as they scale. When designing greenfield industrial sites or upgrading brownfield operations with physical AI, companies also need to consider energy costs, and the associated carbon emissions. What often catches organizations off guard is not the initial cost, but how quickly costs escalate as deployments scale. Bandwidth usage, inference frequency, model updates, monitoring and support overhead can increase nonlinearly once solutions move from a handful of devices to hundreds or thousands in production. Without clear visibility into these cost drivers early on, organizations risk building solutions that are technically viable but economically unsustainable.

These factors can determine whether projects ever move beyond the pilot phase. It's worth noting that industry estimates suggest

78%

of AI enterprise pilot projects fail at the pilot stage, so careful planning and cost projections are crucial.

4. Low confidence

It's not always because of the technology challenges and cost that pilots fail. If stakeholders still have concerns around factors such as cybersecurity, safety standards, data privacy, transparency and compliance regulations, projects will stall. Even where they do not fail completely, hesitancy can significantly limit realized outcomes. For example, if uncertainties remain around the deployment of AGVs on a warehouse floor, and they are reduced to working in cages, this will reduce the return on investment (ROI).

Confidence is also shaped by how predictable and explainable systems are in real operating conditions. If teams cannot understand why a system made a decision, intervene when needed or reassure regulators and stakeholders that behavior remains safe as models evolve; adoption will slow even if technical performance is strong—which is why explainability and transparency are becoming core requirements in emerging AI regulations (e.g., the EU AI Act).

Gartner predicts, however, that if businesses focus on transparency, trust and security in their AI projects, they will achieve a

50%

improvement in terms of adoption, business goals and user acceptance.[4]

5. Organizational transformation

Physical AI projects will also struggle if the workforce fails to come to terms with the transition to autonomous machinery. Organizations should be aware that they may need to acquire new skill sets to manage AI-enabled devices at the edge, as moving AI from the cloud to the edge will require software engineers capable of working in different programming languages.

Beyond skills, organizations often struggle with ownership and accountability in their operating models. Physical AI dissolves traditional boundaries between IT, OT, engineering and operations—requiring a new governance model that aligns incentives, funding and decision rights across these domains. Success depends not just on technical integration, but on redefining how teams collaborate, share accountability and manage systems continuously, rather than as one-off deployments. Without this shift, IT and OT priorities can diverge—slowing adoption and undermining scale.

It is not always about a skills gap, however. If workers on the factory floor struggle to come to terms with new technology, it can slow progress. This can happen if they don't understand how autonomous systems will support their work, when human judgment takes precedence and how accountability is shared.

Unlocking potential

For physical AI projects to succeed, organizations must prove that autonomous machines can operate safely, efficiently and sustainably in real-world conditions while delivering measurable returns on investment. This is not about experimentation or isolated pilots. It requires engineered systems that can perform reliably in production environments, adapt to physical constraints and continue to deliver value over time.

Leading organizations are addressing this challenge by taking a more disciplined, system-level approach to physical AI. Simulation is used to validate safety, performance and edge constraints before deployment, while staged rollouts reduce risk and build confidence as solutions scale. At the same time, success depends on more than technology alone. Clear ownership, governance and accountability are needed to manage the transition carefully, from the boardroom to the shop floor, ensuring that physical AI is not only deployed, but adopted, trusted and operated effectively across the organization.

The businesses that succeed will follow these five steps:

01 Unlock the potential of edge engineering

02 Run high-fidelity simulations

03 Identify efficiencies

04 Take an incremental approach

05 Manage organizational change

01

Unlock the potential of edge engineering

While physical constraints can be a challenge, organizations should avoid underestimating what is now possible with physical AI at the edge. Advances in edge engineering—including more capable low-power GPUs, specialized AI accelerators and model optimization techniques—are rapidly expanding the range of workloads that can be executed locally. At the same time, the cost and energy efficiency of these devices continue to improve, making edge-deployed intelligence increasingly viable at scale.

Even where legacy products or platforms appear to limit what can be achieved, physical AI may still be feasible. Techniques such as model compression and quantization can significantly reduce computational demands while maintaining acceptable performance. In more constrained environments, distributed edge architectures can offload specific tasks across nearby devices, enabling real-time decision-making without relying on continuous cloud connectivity. In practice, this can mean augmenting existing systems, such as domestic models or in vehicle platforms, or selectively leveraging nearby devices like mobile phones, rather than waiting for full platform redesigns.

What matters most is not where intelligence runs, but how deliberately edge constraints are engineered into the solution from the start, including latency, safety, power consumption and operational resilience. This shifts the conversation from infrastructure limitations to engineering intent.

Cognizant and Gentherm—Bringing physical AI to in-cabin comfort

Cognizant partnered with Gentherm, a global leader in thermal control and pneumatic comfort systems, along with Sensory to create a voice-enabled massage seat designed to reduce driver fatigue discomfort and improve safety by enhancing the customer experience.

As the purpose of the seat was to keep the driver alert, the manufacturer wanted to eliminate any distractions caused by reaching for knobs and buttons when driving. As a result, the voice-enabled solution needed to operate locally within the driver's cab and remain fully functional even in connectivity dead zones.

The solution:

- Cognizant embedded a midsized LLM using Sensory Voice SDK
- The Sensory SDK converts speech to text and extracts intent from the voice messages
- Cognizant's custom python script maps the intent to the seat functions
- Seat functions communicate with the Gentherm controller to activate the requested functions
- This solution enables drivers to control seat settings using natural language

This approach enabled natural language control of seat functions without introducing distraction or latency, while ensuring the system continued to operate reliably in connectivity dead zones. The result was a practical example of physical AI engineered for real-world constraints—where edge intelligence, not cloud dependency, was essential to delivering value.

Benefits to the customers:

- AI-enhanced automotive comfort
- Scalable, multilingual architecture
- Leadership in next-gen vehicles
- Enhanced user experience and safety
- Cost-effective innovation

02

Run high-fidelity simulations

Given the scale, cost and risk associated with physical AI investments, organizations should validate decisions in a simulated environment before deploying technology into live operations. High-fidelity simulations allow enterprises to model complex physical systems, test scenarios and understand second-order impacts without disrupting production, compromising safety or committing capital prematurely.

By leveraging advanced simulation platforms such as NVIDIA's Omniverse, companies can create digital twins of factories, assets and workflows—enabling teams to explore "what-if" scenarios across layout design, automation strategies, energy usage and workforce interactions. These environments make it possible to assess performance trade-offs, identify constraints early and rightsize investments before they materialize in the physical world.

Crucially, simulations are not just visual tools. When connected to real operational data, they become decision-making systems that help leaders derisk capital investments, accelerate planning cycles and align stakeholders around a shared, data-driven view of the future.

Beyond factory automation, the same simulation approaches can be applied to energy- intensive environments. Allowing leaders to test cost and sustainability trade-offs virtually before committing to physical infrastructure.

Global climate solutions manufacturer—Using simulations to derisk capital investments

A leading global climate and industrial solutions manufacturer partnered with Cognizant to modernize its digital factory modeling and simulation capabilities using NVIDIA Omniverse.

Traditionally, creating accurate 3D plant models from point-cloud data could take several months per site, slowing decision-making and limiting the organization's ability to validate automation and layout changes at scale. By implementing a high-fidelity digital twin environment within Omniverse, the manufacturer was able to dramatically accelerate modeling timelines and simulate automation scenarios before deploying them in live facilities.

This approach reduced the time required to convert LiDAR scans into operational plant models from more than six months to approximately six weeks. It also enabled teams to test factory layouts, robotics deployments, and AGV scenarios virtually—significantly reducing the risk of costly rework and deployment errors.

As a result, the organization established a scalable, simulation-driven foundation that supports faster capital planning, improved collaboration across teams, and more confident decision-making before physical changes are implemented. This has helped the manufacturer derisk capital investments, optimize factory layouts, and test automation scenarios in a virtual environment before deploying them in the real world.

"

Our Omniverse-based digital twin application allows us to visualize existing data more efficiently while paving the way for training autonomous and vision-guided vehicles in the digital world, with minimal impact on production.

- A digital factory modeling and simulation leader

03

Identify efficiencies

Beyond determining what is technically possible, organizations must identify where physical AI can drive measurable efficiency gains and cost reductions. Engineering decisions such as where inference runs, how models are optimized and how systems are architected can significantly reduce operational expense. For example, enabling devices to operate independently of the cloud can lower ongoing bandwidth costs and API call fees, while techniques such as quantization can reduce power consumption and semiconductor requirements without compromising performance.

Efficiency also increases when AI capabilities are designed to be fine-tuned and reused across multiple scenarios, rather than built as one-off solutions. Cognizant applies a modular approach to industrial computer vision, with reusable core capabilities that can be adapted and redeployed across different environments. These capabilities have been applied in real-world use cases such as theft prevention in grocery and retail settings, as well as tire degradation analysis at automotive test centers. By reusing core AI components and tailoring them to specific contexts, organizations can accelerate deployment timelines and maximize returns on existing AI investments.

At an operational and strategic level, simulations play a critical role in identifying efficiencies beyond the technology itself. High-fidelity models can help leaders determine whether new investments are truly necessary, or whether greater value can be unlocked by optimizing existing facilities and processes. In some cases, simulation-backed insights may support expansion; in others, they may justify preserving capital and improving current operations—ensuring that physical AI investments are aligned with both business priorities and long-term financial discipline.

04

Take an incremental approach

Taking an incremental approach enables organizations to move with intent rather than hesitation. By progressing in clearly defined stages, companies can turn pilots into sustainable production-ready solutions that are designed to scale, adapt to future requirements and grow with the business rather than stall in isolation.

Simulations play a critical role in this approach by identifying quick wins that demonstrate early ROI and operational value. As milestones are reached, organizations gain concrete evidence of performance, safety and cost impact—creating natural decision points to expand scope with confidence.

Developing an initial proof of concept before large-scale deployment also helps build trust across stakeholders. It allows technical leaders to validate that systems behave as expected in real-world conditions, while giving executives confidence that the technology is reliable, compliant and fit for purpose. This reduces hesitation at scale-up stages and prevents promising initiatives from being delayed by unresolved concerns.

05

Manage organizational change

Physical AI adoption requires organizations to rethink roles, skills and ways of working—particularly across engineering, operations and frontline teams. As intelligence moves from centralized cloud environments to machines operating at the edge, organizations may need deeper expertise in embedded systems, real-time software and lower-level programming languages such as C and C++, alongside traditional IT and data skills. In many cases, this shift demands not just upskilling, but structural changes to how teams are organized and governed.

Beyond skills, successful physical AI programs depend on workforce understanding and trust. When new systems interact directly with physical environments—machines, vehicles or shopfloor operations—employees need clarity on why the technology is being introduced, how it affects safety and performance and what role humans continue to play. Without this alignment, even technically sound deployments can face resistance, slow adoption or underutilization.

A structured organizational change management approach is therefore essential. Early stakeholder engagement, clear communication, targeted training and ongoing support help ensure employees are prepared to work confidently alongside intelligent systems. This not only accelerates adoption but also reduces operational risk and builds the organizational readiness required to scale physical AI initiatives over time.

Organizational change management

The introduction of physical AI has a direct impact on how people interact with machines, systems and physical environments. As a result, organizational change management (OCM) should be embedded into every physical AI initiative, from pilot through scale.

A practical OCM approach should include the following five actions:

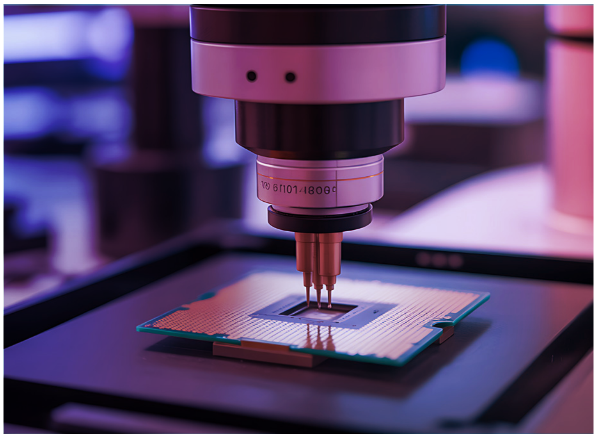
Stakeholder engagement and communication	Engage engineering, operations, safety and frontline teams early to explain the purpose of the initiative, set realistic expectations and address concerns around reliability, safety and job impact. Transparent communication helps build trust in systems that operate autonomously or semiautonomously in real-world conditions.
Advisory framework	Establish cross-functional governance that brings together IT, OT, engineering, safety and business leaders. This ensures physical AI use cases are prioritized based on operational impact, risk and return on investment—and that accountability is clear as solutions move from pilots to production.
Training and adoption	Provide role-specific training that reflects how physical AI systems are actually used in practice—from monitoring and exception handling to system updates and maintenance. Training should focus not only on how technology works, but on how humans collaborate with intelligent machines in day-to-day operations.
OCM and support plan	Develop a deployment playbook covering rollout, support, escalation paths and lifecycle management across sites. This is particularly important for physical AI solutions that must operate reliably in distributed, critical or low-connectivity environments.
Continuous improvement	Treat adoption as an ongoing process rather than a one-time event. Collect feedback from operators and engineers, monitor system performance in real-world conditions and continuously refine both the technology and the operating model throughout the lifecycle.

Impact by sector

To ensure physical AI delivers measurable business and operational outcomes, solutions must be tailored to each organization's environment—its data, physical constraints, operating conditions and workforce. Unlike purely digital AI, physical AI systems interact directly with machines, assets and people, making context critical.

While the underlying technologies may be similar, how physical AI is applied varies significantly by industry. To reflect this, the following examples illustrate how physical AI is shaping outcomes across four sectors:

Automotive



Semiconductors

Manufacturing

Life sciences

1. Automotive

Autonomous vehicles are one of the most visible examples of physical AI in action. However, autonomy is only one manifestation of how intelligence is being embedded into modern vehicles. Physical AI is increasingly shaping in-cabin experiences, driver assistance systems and real-time decision-making—often under strict safety, latency and compute constraints.

Automotive manufacturers are also moving rapidly toward hyper-personalized in-vehicle experiences. When a driver enters a vehicle, settings such as seat position, climate, drive mode and infotainment preferences can adjust automatically. Delivering this seamlessly—especially when multiple drivers share a vehicle—requires real-time inference at the edge, without reliance on cloud connectivity.

In real-world automotive deployments, in-cabin intelligence such as voice-enabled and personalized controls must operate entirely at the edge to avoid latency and driver distraction. In these scenarios, physical AI is embedded directly into vehicle systems to ensure responsiveness and reliability even in connectivity dead zones.

These use cases introduce significant technical challenges. Advanced driver-assistance systems (ADAS) and in-cabin intelligence demand reliable, low-latency computation in safety-critical environments. At the same time, vehicle manufacturers must balance performance with cost, power consumption and thermal limits.

Simply adding more GPUs is rarely viable, forcing teams to optimize models through techniques such as quantization and to carefully select chipsets that can meet both functional and economic requirements.

Across advanced driver assistance programs, these safety-critical requirements place hard constraints on latency, determinism and compute availability—limiting the viability of large, generalized models and forcing teams to optimize models early and validate performance under real-world operating conditions before deployment.

There is also a movement to simplify software applications to reduce the demand on semiconductors within vehicles, to free up computational power and memory for other purposes.

As a result, and as reflected in real-world automotive programs, physical AI is less about deploying the most powerful models and more about engineering intelligence that fits within real-world constraints. Success depends on tight integration between software, hardware, data platforms and vehicle architecture—ensuring end-to-end data flows, scalability and modularity so ensuring intelligent systems perform reliably, safely and efficiently throughout the vehicle lifecycle.

2. Semiconductors

Physical AI places semiconductor manufacturers at the center of a growing tension between model ambition and silicon reality. While AI workloads continue to increase in complexity, advances in transistor density, power efficiency and thermal management are no longer keeping pace. As a result, computational power, memory bandwidth and energy consumption have become defining constraints on what can be deployed at the edge.

To operate within these limits, organizations are increasingly focused on reducing the demands placed on devices through techniques such as model quantization, pruning and software simplification. Rather than assuming continual hardware scaling, physical AI requires tighter coordination between models, software architectures and silicon capabilities—shifting optimization earlier into the design process.

This shift is also influencing semiconductor platform strategies. Instead of relying solely on purpose-built devices with fixed functionality, manufacturers are moving toward more flexible, programmable platforms that can be tuned for different physical AI use cases. These platforms allow organizations to adapt workloads to evolving requirements, accelerating development cycles while avoiding the cost and risk of highly specialized designs.

Where optimization at the chip level alone is insufficient, distributed computing patterns emerge at the edge. In such architectures, AI workloads are partitioned across multiple local processing units, allowing tasks such as perception, speech recognition or sensor fusion to be executed collaboratively. For example, inference workloads may be shared between dedicated accelerators and general-purpose processors within a system, enabling real-time performance without exceeding power or thermal budgets.

As a result, physical AI in semiconductors is less about maximizing raw performance and more about intelligent workload orchestration. Success depends on close integration across hardware, software and system architecture—ensuring AI capabilities can be deployed efficiently, reliably and at scale within real-world physical constraints.

3. Manufacturing

Physical AI has significant potential to automate and optimize manufacturing operations as organizations seek higher performance, resilience and sustainability. In one manufacturing environment, Cognizant partnered with a company where inspections that were previously carried out manually once a month are now performed autonomously on a nightly basis. Acoustically enabled quadrupeds move through production areas, monitoring machinery for abnormal vibrations and early signs of failure. This shift from periodic, human-led inspection to continuous, automated sensing allows issues to be identified sooner, reducing the risk of unplanned downtime and costly disruptions.

However, realizing value from physical AI in manufacturing requires a clear understanding of how these systems impact workflows, productivity and sustainability, as well as how returns on investment are measured. It is not simply a matter of comparing the cost of human labor to automation. Physical AI deployments influence shop floor ergonomics, energy consumption, asset utilization, maintenance strategies and supply chain performance. These broader system-level effects must be evaluated alongside direct efficiency gains.

As physical AI systems begin to make and execute decisions in real-world manufacturing environments, organizations also need visibility into how those decisions propagate across machines, processes and outcomes. Maintaining traceability between data inputs, models, physical actions and operational results helps teams diagnose root causes, support quality and compliance requirements, and continuously improve performance over time.

Without this end-to-end visibility, scaling physical AI can introduce operational and compliance risk rather than resilience. Manufacturing environments also present distinct operational challenges. Most deployments take place in brownfield settings, where legacy equipment, mixed vendor ecosystems and uneven levels of connectivity are the norm. Physical AI systems must be engineered to work incrementally within these constraints, integrating with existing assets and processes rather than assuming clean, fully digitized conditions. Success depends on thoughtful orchestration between machines, software and human operators, not wholesale replacement.

To manage this complexity and derisk investment, manufacturers increasingly rely on digital twins and high-fidelity simulations. By connecting assets, creating virtual representations of production environments and running scenario-based simulations, organizations can evaluate where and how physical AI should be deployed before committing to physical changes. These approaches enable manufacturers to test tolerances, assess process impacts, optimize configurations and make informed adjustments to mitigate operational and safety risks.

Ultimately, physical AI in manufacturing is less about deploying isolated automation and more about creating adaptive systems that learn and improve over time. When intelligence is embedded thoughtfully across sensing, decision-making and execution layers, manufacturers can move beyond detection toward continuous optimization—balancing performance, cost and sustainability within real-world operational constraints.

4. Life sciences

Physical AI is already beginning to reshape life sciences, from pharmaceutical research and drug development to medical training, diagnostics and clinical procedures. Unlike many other industries, however, adoption in this sector is defined not by autonomy, but by trust, validation and patient safety.

Where clinical decision-making is involved, regulations deliberately constrain fully autonomous systems and require a human-in-the-loop (HITL). Far from limiting impact, this requirement is shaping how physical AI is engineered in life sciences—as assistive, decision-supporting systems that augment clinicians rather than replace them. Success depends on designing intelligence that can operate in real-time while remaining transparent, explainable and medically validated.

In practice, some of the most immediate gains from physical AI emerge in training, simulation and clinical support. Digital twins and high-fidelity physiological models are increasingly used to simulate surgical procedures, verify device behavior and support medical training in realistic environments. These systems allow clinicians to train, test and refine procedures safely, reducing both learning time and procedural risk before deployment in live settings.

Example 1:

iEndosurgical system: This system exemplifies physical AI applied to minimally invasive surgery where intelligence is embedded directly into real world clinical workflows. In endoscopic ablation surgery, smoke is a major obstacle that compromises visibility, impedes tool tracking, increases cognitive load, slows workflows and raises safety risks. These challenges are explicitly recognized in the physical AI system description as motivation for real time smoke and plume removal and other perception enhancements designed to support surgeons while preserving human oversight.

The physical AI–enabled ablation surgery system embeds real time intelligence directly into the live endoscopic workflow. It continuously enhances visualization through smoke or plume removal, tracks surgical instruments to maintain situational awareness and detects intraoperative risk events to support timely intervention. In addendum, post surgically, agentic AI capabilities enable rapid retrieval and summarization of procedural video, significantly accelerating report generation and clinical review without manual navigation. The system is designed to operate seamlessly across robotic and conventional ablation platforms, reinforcing HITL governance and preserving clinical accountability.

The solution supports surgeons by enhancing real-time visibility, procedural awareness and risk identification during ablation. It improves intraoperative safety through early event detection and clearer energy delivery visualization. Operational efficiency benefits from a lower cognitive load and quicker post-procedure review. Workflow is streamlined while clinicians retain full oversight and responsibility for patient outcomes.

Physical AI is also driving efficiency in clinical workflows. In one example, agentic models embedded within medical imaging systems generate real-time procedural documentation, significantly reducing the administrative burden on physicians during diagnostic examinations. In another, intelligent imaging platforms support real-time validation during clinical trials, reducing the need for follow-up visits and helping mitigate patient dropout—one of the most persistent challenges in trial execution.

Example 2:

Challenges: Radiology workflows are slowed by manual steps, limited scalability, inconsistent quality checks and time consumption. Clinicians require precise tumor segmentation and prognosis tools, while maintaining compliance and responsibility. Patients face frequent, sometimes unnecessary clinic visits, causing logistical issues and anxiety.

Solution: The AI-powered imaging system uses a human-in-the-loop approach to automate quality checks, tumor segmentation, size estimation and disease prognosis, applying RECIST criteria and producing detailed reports for clinician validation. Integrated feedback keeps the model accurate and compliant, with ongoing monitoring supporting transparency. Real-time review reduces unnecessary patient visits.

Measurable outcomes: Workflow efficiency improves, with faster image reviews and fewer repeat exams. Patient retention rises due to simpler logistics and lower anxiety. Clinicians benefit from advanced tools but remain responsible for care, ensuring safety and reliability through continuous oversight.

Looking further upstream, life sciences experts point to the growing potential of physical AI to materially compress drug discovery and development timelines. By combining simulation- driven experimentation, closed-loop laboratory systems and physician-validated decision support, physical AI can help organizations test hypotheses faster, optimize small-scale manufacturing processes and reduce iteration cycles. While outcomes depend on regulatory context and validation rigor, early applications suggest the potential to significantly shorten discovery lifecycles when physical AI is applied across the full research and development continuum.

Across all these use cases, physician validation remains central. Physical AI systems in life sciences must earn trust not only through performance, but through their ability to integrate clinical expertise into every stage of learning, inference and action. As a result, progress in this sector is less about deploying autonomous intelligence and more about engineering systems that combine AI, simulation and human judgment to deliver safe, scalable impact in real-world clinical environments.

Conclusion

Physical AI represents a fundamental shift in how intelligence is designed, deployed and trusted in the real world. As AI moves beyond digital workflows and analytics into machines, products, factories, vehicles and clinical environments, the challenge is no longer whether intelligent systems are possible, but whether they can operate safely, reliably and at scale under real-world constraints.

Across industries, a consistent pattern is emerging. Organizations that treat physical AI as a bolt-on technology struggle to move beyond pilots. Those that embed intelligence from the outset—engineering it across sensing, decision-making and execution layers—are able to unlock meaningful gains in productivity, resilience and efficiency. In manufacturing, this means shifting from periodic inspection to continuous, autonomous sensing. In automotive and semiconductor engineering, it requires managing complexity across hardware, software and supply chains. In life sciences, it demands HITL systems that balance automation with validation, safety and regulatory trust.

What unites these use cases is not a specific model or platform, but a set of practical principles. Physical AI must be designed with real-world constraints in mind, validated through high-fidelity simulation and testing, introduced incrementally to derisk investment and supported by organizational change that enables sustained adoption. Without this foundation, even the most advanced AI capabilities remain confined to experimentation.

The next phase of AI adoption will be defined less by algorithms breakthroughs and more by excellence in engineering. As intelligence becomes embedded into physical systems, the ability to integrate data, models, hardware, enterprise platforms and human expertise into coherent, adaptive systems will determine which organizations succeed. Physical AI is not about replacing people or processes wholesale, but about augmenting them—creating systems that learn, adapt and improve over time within the realities of the physical world.

For enterprises navigating this transition, the opportunity is significant. Those that approach physical AI with rigor, discipline and a clear understanding of its operational implications can move beyond fragmented initiatives toward scalable impact. The shift from ambition to execution is now underway. The organizations that master it will shape the next era of intelligent, connected and resilient operations.

From vision to execution with Cognizant

Cognizant helps enterprises engineer AI-powered intelligence that works reliably in the physical world. Through its IoT & Engineering business unit, Cognizant combines deep engineering expertise with AI, data and industrial platforms to help organizations move beyond pilots and deploy intelligent systems that operate under real-world constraints—across manufacturing floors, vehicles, medical devices and connected infrastructure.

Cognizant differentiates itself through an end-to-end capability to engineer Physical AI across the full lifecycle—from ideation and design through build, connect, operate, and scale. By combining embedded and digital engineering, edge intelligence, data platforms and AI model acceleration, Cognizant helps clients translate strategic intent into systems that can be deployed, governed and scaled in production environments. This approach focuses not just on intelligence, but on performance, safety, resilience and deep integration across physical assets, digital platforms and existing operational ecosystems—unlocking measurable business value at scale.

Cognizant's global network of specialized engineering and innovation labs underpins this capability. These include industry-specific facilities for smart manufacturing, automotive and software-defined vehicles, medical devices and medtech, life sciences, semiconductor engineering and next-generation connected systems. Located across India, Germany, Japan, Poland and North America, these centers enable clients to design, simulate, test and industrialize physical AI solutions before scaling them into live operations—reducing risk while accelerating time to value.

This engineering depth is strengthened by a curated ecosystem of strategic partners—spanning cloud, silicon, industrial automation and software platforms, including AWS, NVIDIA, OMRON, Siemens, Critical Manufacturing and Microsoft.

In each case, Cognizant acts as the integrator and orchestrator of the physical AI stack, aligning cloud infrastructure, edge computing, silicon, industrial controls and data platforms into cohesive, industry-specific solutions. Together with Cognizant's platforms, accelerators and proven delivery models, these partnerships help industrialize AI—not just prototype it—supporting everything from edge deployment and real-time decision-making to global rollout and continuous optimization.

By combining engineering rigor, AI innovation and end-to-end stack orchestration, Cognizant helps enterprises build intelligent systems that are not only advanced, but operationally viable—turning physical AI from concept into measurable, scalable impact.

References

1. Gartner
2. AlphaSense, Barclays Research
3. Gartner
4. Gartner

Let’s take physical AI from insight to execution

Whether you’re moving beyond AI pilots, engineering intelligence into physical systems, or scaling real-time AI across operations, our experts can help—from strategy through to deployment.

Reach out to continue the conversation:

Deepika Patel | Global Marketing Lead at Cognizant
deepika.patel@cognizant.com

Follow to stay ahead