



From algorithm to action: A framework for safe and secure physical AI

Key considerations for a new era of autonomous systems

kpmg.com



Artificial intelligence (AI) is advancing from digital assistance to physical presence. For decades, AI has operated behind the scenes, optimizing supply chains, producing content, and executing financial trades in various enterprise settings. While significant, much of this impact has been largely virtual. Today, many AI-powered systems can sense, interpret, and physically interact with our world, as well as take control of the everyday machinery that defines modern life. From autonomous vehicles and robotaxis navigating city streets, to collaborative robots on factory floors, automated data centers, smart energy grids and cities, and a host of other systems, physical AI (PAI) continues to reshape the world around us. This rapid proliferation of PAI across both industrial settings and daily life requires companies to redefine what safety, risk, and trust mean when implementing and monitoring these systems.

To appreciate the scale of these operational considerations, it is helpful to situate PAI within the broader landscape of AI and automation. In the case of an AI-based financial algorithm, its immediate effects and direct outputs are confined to the virtual realm. But a flawed algorithm or compromised PAI system can lead to a multiton vehicle swerving into traffic, a factory robot dropping a heavy payload, or an impaired AI-managed data center triggering a cascading failure that disrupts the organizations relying on its services. These life-or-death scenarios elevate safety from one of many considerations to the single most critical and immediate imperative. Since a breach in a PAI system can culminate in a safety failure, the two risks are converging, demanding a new, holistic approach to security.

The three parts of the problem

Effective management of PAI begins with understanding it as a complete, integrated system. To address this challenge, any robust approach must deconstruct the problem into three interconnected domains, each with its own unique failure modes:

01

The brain (the AI model): This is the complex decision-making component that processes sensor data and decides what to do. Failures here are failures of intelligence: For example, a model may misidentify an obstacle and decide not to stop.

02

The nervous system (the control system): This is the signal transmission layer that translates the brain's commands into electronic signals for the hardware. Failures here are failures of execution, such as the brain correctly deciding to stop but a faulty wire preventing the signal from reaching the brakes.

03

The physical body (the asset and ecosystem): This is the hardware itself—the motors, actuators, and joints, and the environment in which it operates. Failures here are mechanical or environmental. For instance, the brain and control system may work perfectly, but the brake pads are worn and fail to stop the vehicle.



01

02

03

04

05

06

A framework to navigate the physical AI journey

Addressing the multifaceted risks of PAI requires moving beyond a static checklist of governance principles siloed across the AI models, the control system, and the asset and ecosystem. Organizations must adopt a prescriptive, hierarchical framework that prioritizes risk based on the system's physical consequence. When a system can exert force, generate motion, or manipulate its environment, certain risks demand precedence.

For years, KPMG has helped organizations navigate the complex risks of AI through our proprietary Trusted AI framework. Grounded in the principles of being values-driven, trustworthy, and human-centric, this thorough approach provides a holistic method for designing, building, deploying, and monitoring AI systems that are not only high-performing but also ethical, explainable, and accountable:



Layer 1 — The foundational pillar: Safety

This is the non-negotiable foundation. Safety must be designed into the framework and assured before any other capability is evaluated. The organizations that most successfully scale PAI will be the ones that can quantify the trade-off between safety investment and operational value.



Layer 2 — The operational integrity pillars: Reliability and data integrity

In the physical world, downtime is not neutral; an unavailable or unreliable asset can create new and severe safety hazards. This layer is therefore focused on ensuring the system is available when needed and behaves predictably. This is underpinned by two core pillars: reliability, which governs whether the system acts correctly and consistently, and data integrity, which helps ensure the quality and trustworthiness of the sensor data the system depends on. A failure in either pillar can directly compromise the system's operational integrity, which is critical for the safety foundation.



Layer 3 — The reinforcing pillars: Accountability, fairness, privacy, sustainability, transparency, and explainability

Once a system is established to be safe and operationally sound, the organization can then effectively address the pillars that govern its broader performance and societal impact. This includes attributes like explainability and transparency that help ensure its decisions are understood, and pillars like Fairness and Privacy that help ensure it operates ethically within its environment.



Omnipresent layer — Security

The final pillar of Trusted AI is not confined to a single layer. It is an omnipresent imperative that serves as the foundation for the entire system's architecture. A security compromise, whether through a poisoned model, a spoofed sensor, or a hijacked control system, immediately threatens both physical safety and operational integrity. Consequently, security must be woven through every layer of the framework.

¹ Autonomy: In the context of AI and robotics, autonomy is the ability of a system to make its own decisions and take independent action to achieve a set of goals, including within dynamic and unpredictable environments. It is distinct from automation, which refers to a system following a predefined set of instructions or programming. The core mechanism of autonomy is the "sense-decide-act" loop.



01

02

03

04

05

06

Illustrative physical AI journey



Layer 1 The foundational pillars

This is the non-negotiable foundation. Safety must be designed into the framework and assured before any other capability is evaluated.



Layer 2 The operational integrity pillars

In the physical world, downtime is not neutral; an unavailable or unreliable asset can create new and severe safety hazards. This layer is therefore focused on ensuring the system is available when needed and behaves predictably. This is underpinned by two core pillars: reliability, which governs whether the system acts correctly and consistently, and data integrity, which ensures the quality and trustworthiness of the sensor data the system depends on. A failure in either pillar can directly compromise the system's operational integrity, which is critical for the safety foundation.



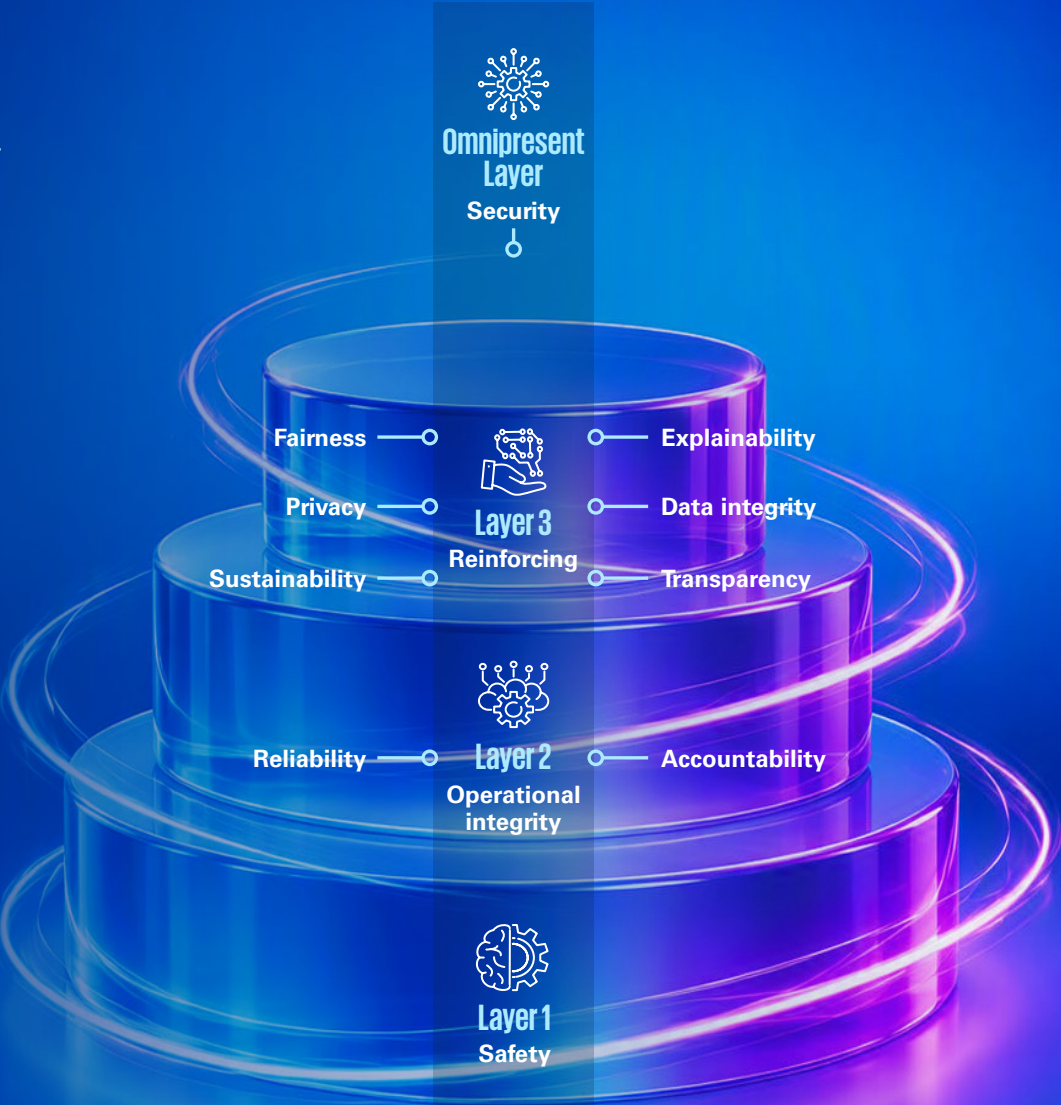
Layer 3 The reinforcing pillars

Once a system is proven to be safe and operationally sound, the organization can then effectively address the pillars that govern its broader performance and societal impact. This includes attributes like explainability and transparency that ensure its decisions are understood, and pillars like Fairness and Privacy that ensure it operates ethically within its environment.



Omnipresent layer Security

The final pillar of Trusted AI is not confined to a single layer. It is an omnipresent imperative that serves as the foundation for the entire system's architecture. A security compromise, whether through a poisoned model, a spoofed sensor, or a hijacked control system, immediately threatens both physical safety and operational integrity. Consequently, security must be woven through every layer of the framework.





Layer 1—The foundational pillar: Safety



The question: If the system fails, can it cause harm to someone?

If a system can cause physical harm, then every other attribute is secondary. Safety must be addressed before any other capability is evaluated. Harm prevention, pre-defined fail-safe states, orderly shutdown, and manual overrides are all bound by an explicit operational design domain (ODD) used to control and operate the automated or autonomous system designed to function safely. The system's ODD should be certified to behave within specified security parameters for such factors as environmental conditions, geographic boundaries, and roadway and traffic conditions.

The maker is accountable for the product's safety and intended use, with tested hardware fail-safes built in and the ODD documented. The adopter is accountable for operating strictly within the ODD parameters, maintaining specifications and integrating the asset safely alongside employees' obligations that are concentrated within system's operational domain.

For the chief information security officer (CISO) and chief technology officer (CTO):

The decisive architectural question is whether the brain can influence the safety features of the control system and supporting assets. In a defensible design system, the brain should not have the capability to modify the safety functions of the control system and supporting assets. The protective safety features are embedded in the control system and supporting asset components (e.g., a separately rated controller with an emergency stop that is hardwired rather than carried on the bus the autonomy stack uses). The brain should have neither write privileges nor a physical path to these controls; otherwise, the safety case depends on a probabilistic component that cannot be fully assured. ODD parameters define the control system's certified operating envelope, so changes should be signed, version-controlled, and independently detected outside the vendor management console. ODD compliance should also be confirmed out of band, rather than self-reported by the same brain that could be compromised.

Critical questions:

- Who verified the control system's AI model deployment?
- When was the system's AI model deployment last verified?
- Where there any modifications to the AI model since deployment?
- Has the manufacturer certified the ODD parameters used by the control system? What parameters have been certified?

Potential solutions:

- Assess the maturity, capability, and functionality of the control system's defined safety features.
- Test and evaluate the designed safety features that the maker (i.e., vendor) provides.
- Verify the operating procedures used by the adopter (i.e., consumer) within a defined environment.



Layer 2 — The operational integrity pillars: Reliability and data integrity



The question: Is the system's operational response fully predictable and repeatable when supplied with trusted telemetry?

Downtime in the physical world is not neutral. An unavailable or unreliable asset can create new safety hazards, rather than accounting for a merely pausing value (e.g., a vehicle that loses its control link mid-maneuver, a robot that halts unpredictably on a floor shared with humans). The failure of an unavailable or unreliable asset becomes a safety failure within seconds:

■ **Reliability** governs whether the system acts correctly and consistently, and each domain contributes a distinct failure. The brain must produce stable decisions and remain accessible. For example, off-device inference introduces a cloud dependency into the control loop—meaning that a simple network outage can immediately halt a production line or leave moving machinery operating without oversight. The nervous system must deliver commands within a bounded time, since a signal that could potentially arrive late is functionally a signal that did not arrive. The body must perform within its published envelope as components age. The question is never whether one of the asset's components can degrade, but what and how the asset will operate in the seconds afterward. These questions make predictable degraded-mode behavior a safety requirement rather than a performance one.

■ **Data integrity** is where body and brain meet and coordinate. Sensors function like the system's sensory organs and everything the brain concludes is downstream of what the sensors report. Calibration drift, occlusion, and environmental degradation are not typical data-reporting errors. Instead, they represent a sensor-to-processor failure cascade, where physical abnormalities mislead a healthy central controller while every component falsely reports normal status.

The maker's responsibility is to publish operating envelopes, design degraded-mode and loss-of-link behavior, provide secure and reversible update mechanisms, and specify a maintenance regime. The adopter builds operational governance based on the maker's design, which can include but is not limited to: operating within the asset's designed constraints, training staff for degraded and failure scenarios, managing change-window risks, and reporting maintenance.

For the CISO and CTO:

Three distinct paradigms separate AI-enabled operational technology (OT) from traditional information technology (IT) in this layer, mapped across the brain (orchestration and secure edge updates), the body (sensor trust and telemetry integrity), and the nervous system (temporal synchronization).

- **The brain:** Over-the-air (OTA) updates to edge-deployed AI decision models represent remote code execution with direct, kinetic consequences. To mitigate this risk, deployment architectures must enforce signed images with anchored rollback targets, staged canary rollouts to isolated fleet subsets, and physical interlocks that block updates while assets are in motion or under load. Crucially, simultaneous fleet-wide updates must be barred, as they risk converting a faulty build into a synchronized, systemic incident.
- **The body:** Physical telemetry reliability cannot be assumed. Sensor calibration states must be continuously attested and logged. Cross-modal telemetry divergence (e.g., disagreement between LiDAR, camera, and radar) must be treated as an active threat indicator. This divergence is often the earliest signal of physical occlusion, calibration drift, or sensor spoofing.
- **The nervous system:** Effective sensor fusion relies entirely on synchronized, sub-millisecond clock sources. Manipulating the timing source degrades AI perception and decision-making without needing to compromise the physical sensors themselves. Consequently, time-source integrity is a foundational perception control, not a passive logging function.

Potential solutions:

- Evaluate contingency plans, failure-prevention controls, and fail-safe procedures against National Institute of Standards and Technology Special Publication 800-82 (NIST SP 800-82) guidance so operational continuity is maintained in accordance with safety controls.



01

02

03

04

05

06



Layer 3 — The reinforcing pillars: Accountability, fairness, privacy, sustainability, transparency, and explainability



The question: Once operational safety and soundness are established, is the system’s decision-making fully auditable, defensible, and compliant with regulatory standards?

Each of the following pillars applies across all three domains (the brain, the body, and the nervous system) and each acquires a physical dimension extension. Examples include:

Pillars	Physical extension
Accountability	Accountability and responsibility grid, liability mapping for physical harm across maker, integrator, and operator
Fairness	Perception performance across skin tones, mobility aids, low light, glare, weather—a detection gap is a safety gap for a specific group of people
Privacy	Bystanders who never chose to interact with the system could be captured incidentally
Transparency	Whether people sharing the space can tell what the machine is about to perform—intent signaling as a safety control
Explainability	Reconstructing what was perceived, what was decided, and what was commanded
Sustainability	Asset lifecycle, battery and materials, end of life

Two pillars can carry a disproportionate enterprise weight. Accountability can become contractual, and the three-domain structure can also make accountability complex. Explainability can become forensic and spans the three domains by necessity. Fairness and transparency are not purely ethical concerns in this scenario. A perception system that does not detect a category of pedestrian, or a machine that gives no indication of its next move, produces unequal physical risk. Layer 3 is where safety stops being a system property and becomes a question of who bears the residual risk.

For the CISO and CTO:

- **The body:** In AI-driven OT, the novel attack surface is the body—the system’s physical sensors, which can be manipulated entirely without network access. Adversarial patches, projected markings, or spoofed positioning signals directly target the AI’s input layer. To enforce strict accountability, physical-world adversarial testing must be brought into red-team scope alongside network penetration testing, as traditional IT telemetry remains blind to these physical intrusions.
- **The brain:** Designing forensic capability is an early architectural mandate. To achieve true explainability and transparency, incident reconstruction must capture the body’s raw perception inputs

alongside the brain’s (the AI model’s) decision outputs. Because this raises data volumes by orders of magnitude, personnel must make deliberate tradeoffs: balancing privacy (strict data minimization) and sustainability (managing the environmental and cost footprint of massive storage) against the fidelity required for rigorous oversight and investigation.

- **The nervous system:** Standard security information and event management retention windows expire long before a liability claim matures. To guarantee fairness in post-incident investigations, retention policies must coincide with long-term litigation and regulatory timelines. Furthermore, within the nervous system, the timing and correlation infrastructure connecting the system must produce tamper-evident, flawlessly synchronized records. Logs that lack strict sequence correlation across the fleet or that an operator could potentially alter undermine accountability and are fundamentally worthless to insurers and regulators.

Critical questions:

- Are the PAI assets actively tested to ensure that their sensors can’t be influenced or manipulated in the real world?
- If a PAI asset causes physical damage because its sensors were fooled by a real-world adversarial attack, then who bears the liability—the organization that adopts the asset, the AI developer, or the sensor manufacturer?
- Given the massive amounts of environmental data that the PAI asset is leveraging, how is forensic necessity balanced against corporate privacy mandates and data minimization rules?
- Is perception data blindly hoarded to mitigate liability, and if so, how does the massive data center footprint required to store this align with sustainability goals?
- If a regulatory claim arises three years after an incident, will corporate logs still exist, or are they being prematurely wiped by default IT retention policies?

Potential solutions:

For manufacturers, it’s essential to ensure that controls are properly designed and embedded. Adopters should integrate PAI risk into existing enterprise functions such as supply chain verification, asset management, monitoring, and incident response.



01

02

03

04

05

06

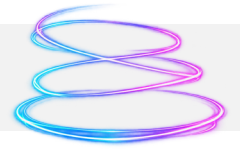
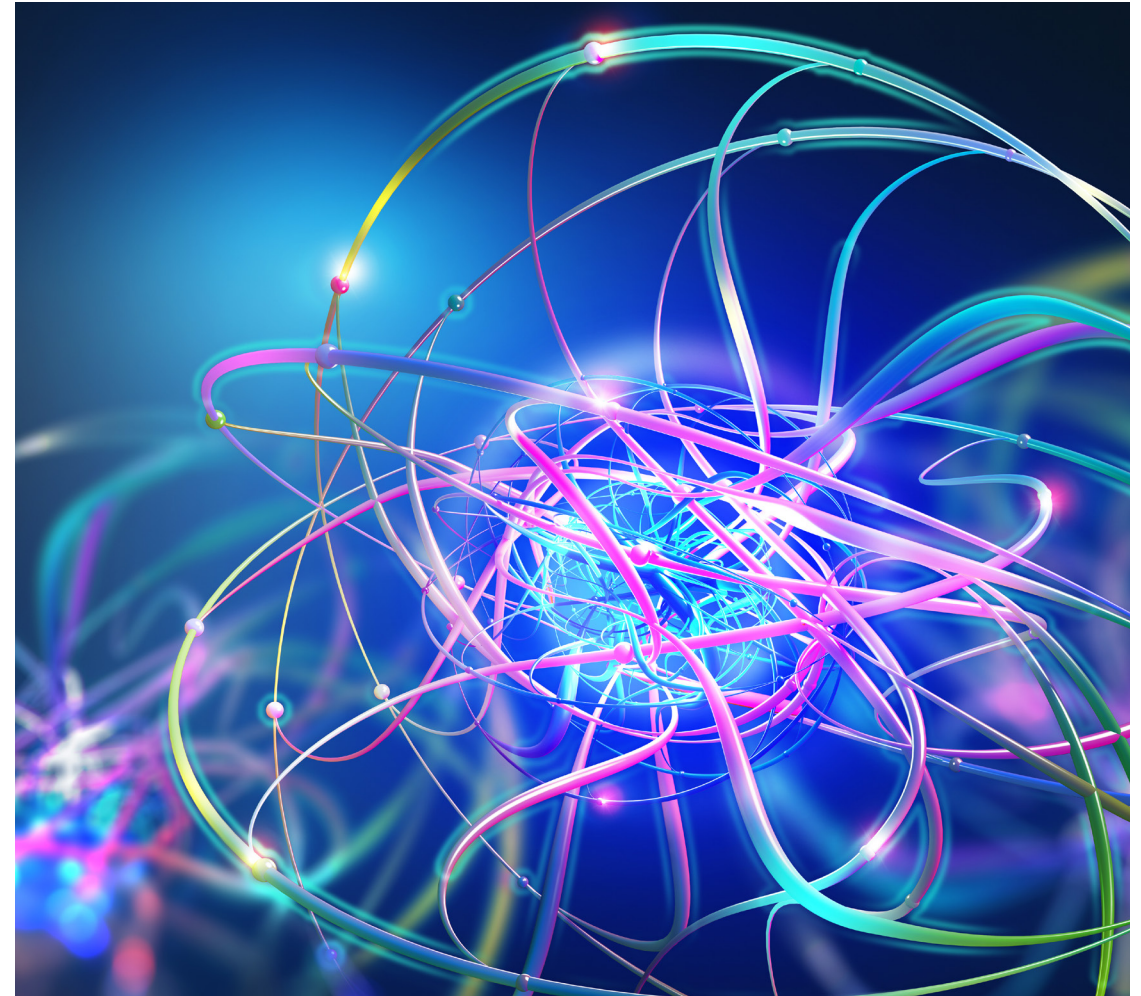
Omnipresent layer: Security

For PAI assets, security cannot be confined to a single layer because the threat landscape spans every domain; it is an omnipresent architectural discipline. The three canonical threat vectors map directly to this architecture—a poisoned AI model compromises the brain, a hijacked control loop subverts the nervous system, and a spoofed sensor deceives the body. A breach in any of these instantly threatens the physical safety and operational integrity of the PAI asset.

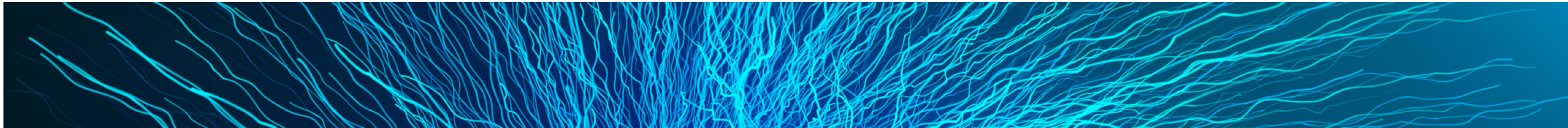
To defend this environment, we must strictly separate two concepts. Safety asks whether the system causes harm when it malfunctions; security asks whether an adversary can force it to malfunction. For example, a geofence that restricts a machine's movement is a safety control, but the authentication protecting that geofence's configuration is a security control. Frameworks that conflate the two often build robust safety mechanisms while dangerously assuming the environment is secure.

This distinction is critical because of the system's cascade effect. In an interconnected AI-enabled OT environment, failure propagates across domains, meaning an adversary only needs to exploit the weakest link or a vulnerability that can generate cascading effects. An AI model secured to the modern cryptographic standard offers zero protection if its physical sensors can be spoofed in the real world or if malicious commands can be injected downstream.

Layer	What security protects	Consequence if it fails
Safety	ODD parameters, failsafe logic, emergency stop paths	Certified safety envelope silently redefined
Operational integrity	Update channels, control links, sensor and timing inputs	Reliable behavior on corrupted perception
Reinforcing	Audit logs, decision records, telemetry	No defensible account of what happened, or understanding of who is liable



Understanding the value on the table



Physical AI is moving from pilots to production.

The organizations that treat trust as an enabler rather than a checkpoint are the ones that will scale it.



Deploy PAI with adequate caution and confidence.

A clear view of what is safe, who is accountable, and where the limits sit allows an organization to move faster because the hard questions are answered before they become roadblocks.



Turn trust into a commercial advantage.

Buyers now arrive with mature security questions. Being able to answer them with evidence shortens sales cycles and wins the deals where trust is the deciding factor.



Protect the value being unlocked.

A single physical incident can erase the operational gains and the reputation behind them. Getting this right protects both the upside and the license to keep operating.



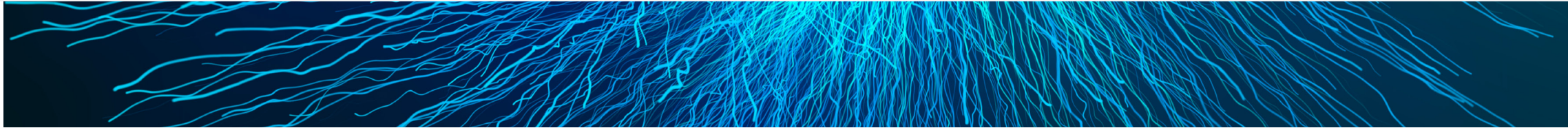
Build once, scale everywhere.

A common framework across the fleet, the supplier base, and every site removes the need to re-litigate trust with each new deployment.



Securing trust is the license to operate.

When a physical AI system fails, the consequences are not digital. They are measured in physical damage, operational downtime, or direct harm to a person. That difference is why safety and trust are the preconditions for operating AI in the physical world.



01

02

03

04

05

06

Authors



Vijay Subramanyam

Principal, Advisory, Technology, Media, and Telecommunications (TMT) Consulting Industry Leader (US), KPMG LLP
E: vijaysubramanyam@kpmg.com

Vijay leads the firm's consulting strategy and delivery for clients across the TMT sector, helping them navigate disruption, accelerate digital transformation, and unlock sustainable growth. As the OpenAI Platform Leader for KPMG in the US, he helps clients realize the value of AI-native enterprise transformations using OpenAI platforms.



Brian Higgins

US and Americas Sector Leader, KPMG LLP
E: bhiggins@kpmg.com

Brian's experience encompasses competitive strategy and operational design. He specializes in identifying and unlocking value within complex global supply chain environments, with a special emphasis on industrial manufacturing challenges and opportunities.



Lisa Rawls

Principal, Advisory, TMT, Cyber and Technology Risk (US), KPMG LLP
E: lisarawls@kpmg.com

Lisa has more than 20 years of experience leading transformative risk, compliance, and security initiatives that leverage modern technology for global companies. Lisa is the co-founder of the Women of Risk and Cyber Community and the co-chair of the KPMG Network of Women.



Brad Raiford

Managing Director, Physical AI and OT Cyber Leader (US), KPMG LLP
E: braiford@kpmg.com

Brad is the US lead for IoT/OT applied cyber engineering, helping clients transform their security footprint from shop floors to consumer devices. His areas of experience include IoT/OT zero trust defense, macro/micro segmentation, software defined IoT/OT networks, trusted autonomy, and cyber program and framework design.



Nana Amonoo-Neizer

Director, Cyber and Technology Risk (US), KPMG LLP
E: namonooneizer@kpmg.com

Nana leads the KPMG US Trusted AI team, helping clients balance AI innovation with effective risk management. He specializes in AI governance, security, risk management, and regulatory compliance for Generative AI and Agentic AI systems.

Core Contributors



Stephanie Kim

Associate Managing Director, Client Technology and Innovation (US), KPMG LLP



Priya Roy

Manager, Client Technology and Innovation (US), KPMG LLP



Cedric Carter

Manager, Cybersecurity and Tech Risk (US), KPMG LLP

Editorial and Design Team: Rama Ramaswami, Whitney La Bounty

How KPMG can help

Succeeding with PAI means managing a complete system: a brain (the AI model), a nervous system (the control architecture), and a physical body (the asset and its environment). The most dangerous failures live in the seams between these domains, and between the organization that builds the system and the one that operates it. A PAI system is only as trustworthy as its weakest component and its least-defined handoff.

The ability to provide verifiable evidence of safety at every stage is what can unlock regulatory approval, public trust, and ultimately, market leadership. This demands a deliberate set of actions: Manage the risks unique to each domain, establish clear responsibility models that close the maker-to-adopter gap before an incident exposes it, and converge learning models with established industrial safety standards.

This is where KPMG is built to help. By unifying our market-leading Trusted AI framework with deep industrial OT security experience, we help organizations bridge the gap between algorithmic trust and physical safety, from designing resilient control architectures to implementing their governance.

The principles are universal, but the pressures and priorities shift sharply by use case. Our sector spotlights show how these challenges play out in practice, from the open public roads of autonomous transportation to the high-stakes, interconnected world of vital infrastructure.

Beyond the core technological and operational considerations, organizations must also address a broader set of workforce-related factors that are critical to sustaining trust, adoption, and long-term value realization. Failure to thoughtfully navigate these issues can undermine confidence in AI-enabled operations and human-machine collaboration. Our future thought leadership will explore several of these areas in greater detail, including:

- Human-machine teaming and collaboration models
- Decision rights and accountability frameworks
- Escalation authority and exception management protocols
- Labor, workforce, and union-related considerations



01

02

03

04

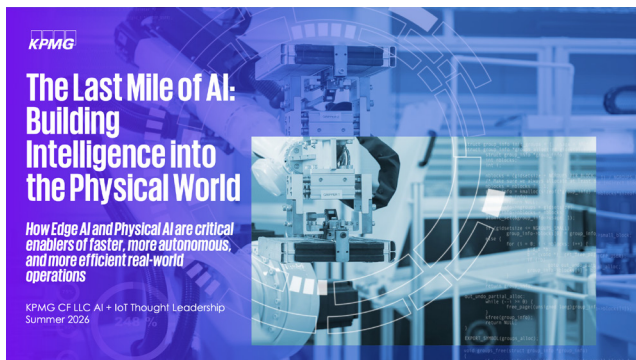
05

06

Related thought leadership and resources



**KPMG 2026 Frontiers Report:
AI as the frontier multiplier**



**The Last Mile of AI: Building Intelligence into the
Physical World**



Digital twin and spatial computing practice

Some or all of the services described herein may not be permissible for KPMG audit clients and their affiliates or related entities.

Please visit us:



[kpmg.com](https://www.kpmg.com)



Subscribe

The information contained herein is of a general nature and is not intended to address the circumstances of any particular individual or entity. Although we endeavor to provide accurate and timely information, there can be no guarantee that such information is accurate as of the date it is received or that it will continue to be accurate in the future.

No one should act upon such information without appropriate professional advice after a thorough examination of the particular situation.

© 2026 KPMG LLP, a Delaware limited liability partnership, and its subsidiaries are part of the KPMG global organization of independent member firms affiliated with KPMG International Limited, a private English company limited by guarantee. All rights reserved.

The KPMG name and logo are trademarks used under license by the independent member firms of the KPMG global organization.

DASD-2026-20766



01

02

03

04

05

06