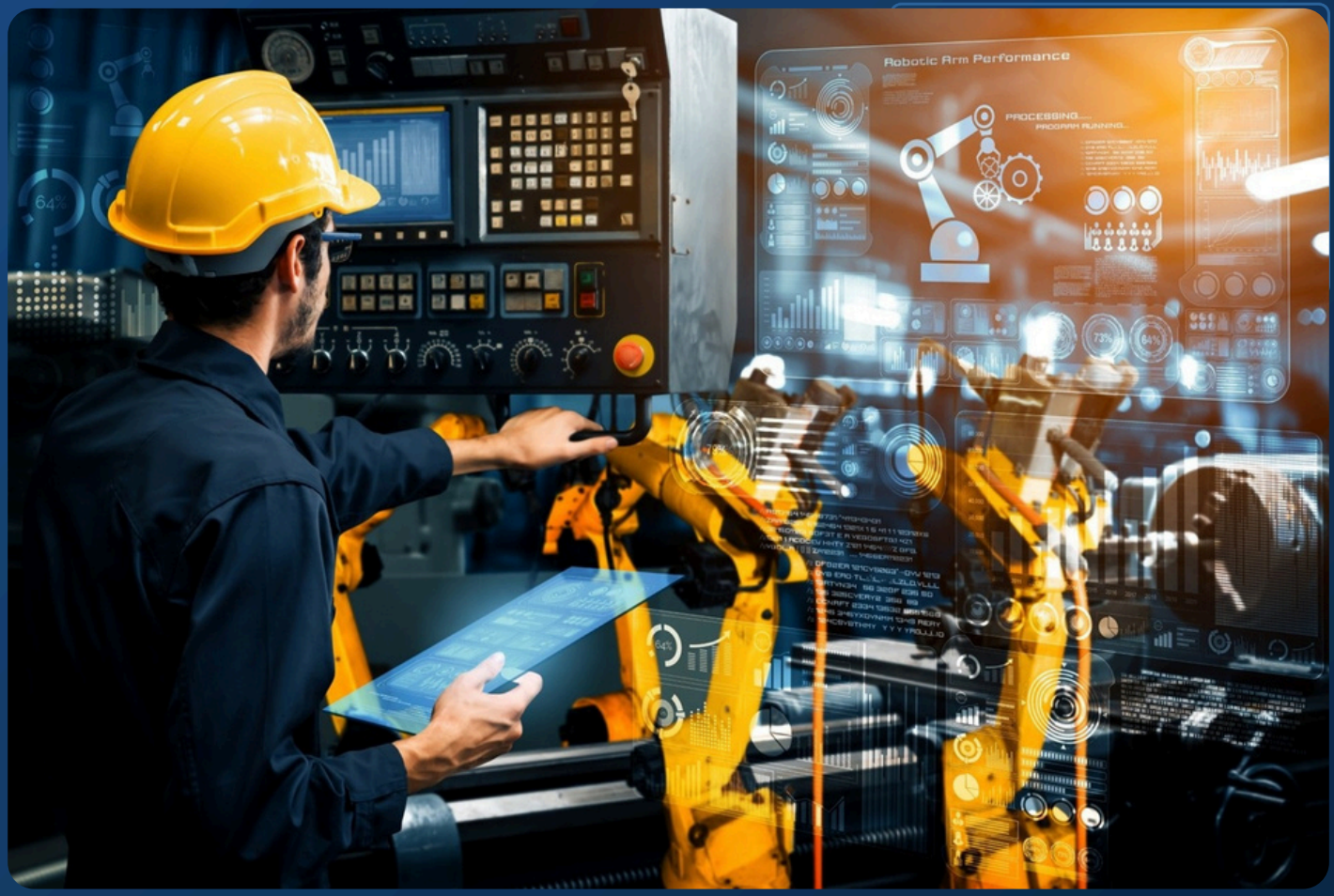
DEVICE AUTHORITY



When AI Agents Control the Physical World

Why device trust must become part of the AI trust model.

Device Authority | **September 2026**

Introduction: AI is acquiring hands

Much of the discussion about AI security has understandably focused on the AI itself.
Can we trust the model? Is the agent authorized to perform an action? Can its reasoning be constrained? Can sensitive data be protected? Can we prevent an agent from accessing tools or information beyond its intended permissions?

These questions remain important. But as AI moves beyond generating information and begins interacting directly with the physical world, another trust question becomes equally important:

> **Can the AI trust the machine it is talking to?**

Anthropic's recently announced Model Hardware Standard (MHS) brings this question sharply into focus.

MHS is a research-preview specification designed to create a common interface between AI agents and physical equipment. Developed initially by Anthropic and HHMI Janelia Research Campus, it is being explored across scientific research, robotics, electronics and advanced manufacturing.

The potential is significant. Instead of engineers spending weeks building bespoke integrations between instruments, MHS provides a standardized way for agents to discover equipment, understand what it can do and issue commands. An AI agent could orchestrate microscopes, liquid handlers, robotic arms, sensors and other equipment as part of a single workflow — an important step toward autonomous laboratories, factories and other cyber-physical environments.

It also changes the security equation. When AI gains the ability to act in the physical world, trusting the intelligence issuing the instruction is only half of the problem. We must also establish trust in the machine receiving it.

From tool use to physical agency

The Model Context Protocol and similar approaches have helped establish a model in which AI agents can interact with external software tools and data. MHS extends this principle into hardware.

A standardized driver allows equipment to expose its capabilities and state through a common interface. An agent can therefore discover an instrument, understand its characteristics and operating constraints, read information from it, and issue instructions without requiring a bespoke integration for every device.

The implications go beyond convenience. Anthropic describes environments in which agents can coordinate multiple instruments, monitor results, adjust parameters as conditions change, and create deterministic sequences of hardware commands for long-running operations. This creates a progression from:

AI that recommends → AI that acts in software → AI that acts on physical systems

Each transition increases the importance of identity and trust. A software agent accessing the wrong database can cause a security incident. An agent controlling the wrong robotic arm, laboratory instrument, actuator, or piece of industrial machinery can potentially cause a physical one.

The missing half of the trust relationship

Consider a simple scenario. An AI agent has been authorized to adjust the temperature of a particular laboratory instrument. The agent itself is legitimate. Its instruction is permitted. The requested temperature is within the operating limits described by the hardware interface. From the agent's perspective, everything appears valid.

But what proves that the endpoint responding as that instrument actually **is** that instrument?
The endpoint could have been substituted. A legitimate device could have been cloned or impersonated. Its software could have

been compromised. Its credentials could have been copied. The device could have been replaced without the agent's trust information being updated. Or the device could be genuine but no longer in a sufficiently trusted operational state to execute the requested action.

In conventional IT environments these are familiar identity questions. In an AI-controlled physical environment, however, their consequences change. The question is no longer simply "Is this agent authorized to issue this command?" It becomes:

> **Is this agent authorized to issue this command to this specific, authenticated and sufficiently trusted physical machine, in its current state?**

Discovery is not identity

One important security principle for emerging AI-to-hardware architectures is that **discoverability should not automatically imply trustworthiness**. A device being visible to an agent does not prove its identity. A device accurately describing its capabilities does not necessarily prove its identity. An endpoint responding at an expected network address does not prove its identity. Even possession of a credential may not be sufficient if that credential can be extracted and copied to another machine.

MHS is designed to help an agent find equipment, understand what that equipment can do, and interact with it consistently — solving an important interoperability problem. Machine identity security addresses a different problem: establishing cryptographic evidence that the physical machine behind that interface is the machine it claims to be.

As AI-to-hardware architectures mature, these two functions should complement one another. The agent needs to understand **what** a machine can do. It also needs confidence in **which machine** it is talking to.

Moving from identity to attestation

For low-risk applications, conventional authentication may provide sufficient assurance. For higher-risk cyber-physical environments, a stronger model is likely to be necessary.

Rather than simply asking whether a device possesses a valid credential, an agent may need evidence tied to the physical characteristics or hardware root of trust of the machine itself — the distinction between authentication and stronger forms of device attestation.

The objective is not merely to establish "this endpoint possesses credential X." It is to establish something closer to: "This is physical device X, its identity can be cryptographically verified, and sufficient evidence exists to determine whether it is currently trusted for this operation."

This becomes particularly important as AI agents become more autonomous. A human technician operating a laboratory or factory may notice that a machine has been replaced, moved, reconfigured, or is behaving unexpectedly. An autonomous agent operating hundreds of devices continuously may not have that contextual awareness unless the underlying architecture supplies it. Trust therefore needs to become machine-verifiable.

Trust should be contextual, not binary

Identity alone is also unlikely to be enough. A machine can be genuine and still be unsafe to use. That suggests the AI-to-hardware trust model needs to evolve beyond binary authentication toward **continuous device trust**.

Before an agent performs a sensitive action, several questions may ultimately need answering:

- Is this the expected physical device?
- Can its identity be anchored to hardware rather than simply software?
- Has the device been cloned or substituted?

- Is its credential valid?
- Is its software or firmware in an acceptable state?
- Does its current configuration satisfy policy?
- Is there security or operational context that makes this particular action unsafe?
- Is the agent itself authenticated and authorized to control this device?

The result is a two-sided trust relationship. **Agent trust** establishes who or what is requesting an action and whether it has permission to do so. **Device trust** establishes which physical machine will execute the action and whether that machine is sufficiently trustworthy to participate. Only when both conditions are satisfied should high-impact actions proceed.

The machine identity problem becomes a safety problem

This distinction matters because MHS is aimed at physical equipment. Anthropic's examples include robotic arms, liquid handlers, microscopes and laser systems. The same architectural concept could ultimately apply much more broadly across manufacturing, healthcare, energy, automotive and other operational environments. In these environments, cyber risk and physical risk increasingly converge.

If an agent is authenticated but the device is not, a command to move a robotic arm could be routed to an impersonated or incorrectly identified system. If an instrument has been substituted or compromised, an AI-controlled laboratory may make subsequent decisions using data from an untrusted source. And in an autonomous manufacturing workflow involving dozens of interconnected machines, if the agent assumes every discovered endpoint is trustworthy, one compromised machine could influence decisions throughout the workflow.

This means machine identity becomes part of the safety architecture. A useful principle for physical agentic AI may therefore be:

> **No consequential action without mutual, verifiable trust.**

Trusting what comes back matters too

There is another side to this problem that is easily overlooked. Agents do not simply send commands to machines — they increasingly make decisions based on information returned by those machines. Anthropic describes closed-loop scenarios in which an agent observes results and adjusts subsequent actions accordingly. That creates a chain of machine-generated evidence:

Device → data → AI interpretation → decision → device command

If the provenance of the original data cannot be trusted, every subsequent decision inherits that uncertainty. Device identity therefore contributes not only to command authorization but also to **data provenance**. The agent needs confidence that a temperature reading came from the expected sensor, an image came from the intended microscope, or a measurement came from the correct laboratory instrument. Machine identity becomes the anchor connecting digital information back to its physical source.

A Zero Trust model for physical agentic AI

The security industry has spent years applying Zero Trust principles to users, applications, workloads and networks. Agentic AI controlling physical systems suggests those principles need to extend further. Instead of assuming trust because a machine is connected, discovered, or located on an approved network, trust should be explicitly established and continuously evaluated. A future AI-to-machine transaction could therefore resemble:

Discover → Identify → Attest → Authorize → Act → Verify

MHS addresses important elements around discovery, machine description and interaction. A complementary machine identity and attestation layer could establish the identity and trustworthiness of the physical endpoint before consequential actions are

authorized.

This separation is useful architecturally. An interoperability standard should not necessarily need to become a full identity platform. But secure deployments will need a mechanism through which trusted identity and attestation evidence can inform the agent's decision-making process. In other words, the hardware standard can define **how AI communicates with machines**, while a machine trust layer determines **which machines should be trusted to participate**.

Designing trust into the standard early

MHS is currently a research preview, which makes these questions particularly timely. Anthropic is inviting participants to help develop safety evaluations and best practices before the specification becomes open source. Security controls are generally easier to incorporate while architectures are being defined than after ecosystems have scaled.

Several principles are worth considering as MHS and similar approaches mature: device identity should be first-class information rather than an implicit property of a connection; attestation should be available for environments requiring stronger proof of physical device identity; authorization decisions should be able to consume real-time device trust context; agent identity and device identity should be mutually verifiable; device-generated data should have trustworthy provenance where it informs autonomous decisions; and trust should be continuously reassessed rather than established once during onboarding.

These concepts are not unique to MHS. They represent broader requirements likely to emerge wherever AI agents interact autonomously with cyber-physical systems.

The bigger picture: a new class of machine-to-machine identity

Enterprise identity architectures were originally designed around people. Cloud computing expanded the problem to workloads, applications, containers and services. Connected products and operational technology expanded it again to machines and physical devices. Agentic AI introduces another participant: an autonomous digital entity capable of discovering machines, making decisions, and instructing them to act.

The resulting environment is therefore not simply human-to-machine or application-to-machine. It is increasingly **agent-to-machine**. Both sides are non-human identities. Both require authentication. Both require authorization. Both require lifecycle management. And when those identities interact in the physical world, the consequences of misplaced trust become considerably greater.

This is why machine identity security should become part of the emerging discussion around safe physical AI.

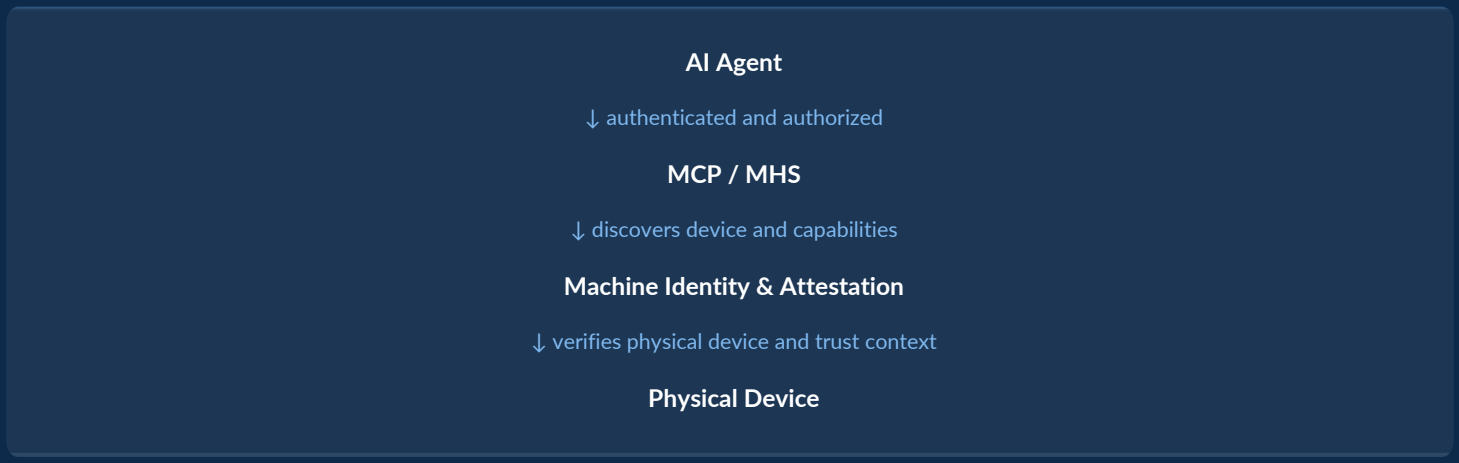
APPLYING THE PRINCIPLE

Hardware-rooted trust with KeyScaler

The security model described above is technology-neutral. One example of how it could be implemented is through hardware-rooted device identity and attestation technologies such as Device Authority's KeyScaler and Dynamic Device Key Generation (DDKG).

DDKG was developed to establish strong identity for connected physical devices. Rather than relying solely on a credential presented by software, it can use hardware-derived characteristics to establish cryptographic device identity and help determine whether a device is genuine rather than a cloned or substituted endpoint.

Applied to an MHS architecture, this could introduce an explicit device-trust step between discovery and control. An agent might discover a device through MHS and learn that it is a particular robotic arm with a defined set of capabilities. Before issuing a consequential command, however, the system could request attestation. KeyScaler could verify the machine's identity and relevant trust context, policy could then determine whether the device is authorized to participate, and the agent could subsequently receive an appropriate credential enabling authenticated communication with that verified physical machine.



This would create mutual trust rather than relying solely on the agent discovering an endpoint and accepting its presented identity. Device Authority's analysis identifies three practical benefits from such an architecture:

01 PHYSICAL SAFETY
An agent can be prevented from issuing consequential commands to an unverified, cloned, or compromised device.

02 OPERATIONAL RESILIENCE
Strong device identity reduces the risk of an autonomous workflow controlling the wrong machine among many similar instruments.

03 DATA INTEGRITY
Information consumed by the agent can be cryptographically associated with the physical device expected to have generated it.

The broader point is not that MHS should depend on any single security technology. It is that **device identity and attestation should become explicit components of the trust architecture around physical agentic AI**.

DEVICE AUTHORITY

Trusting the intelligence issuing the command is essential.
Trusting the machine receiving it is equally important.

Interested in what hardware-rooted device attestation looks like for agentic AI in practice? Let's talk.

sales@deviceauthority.com