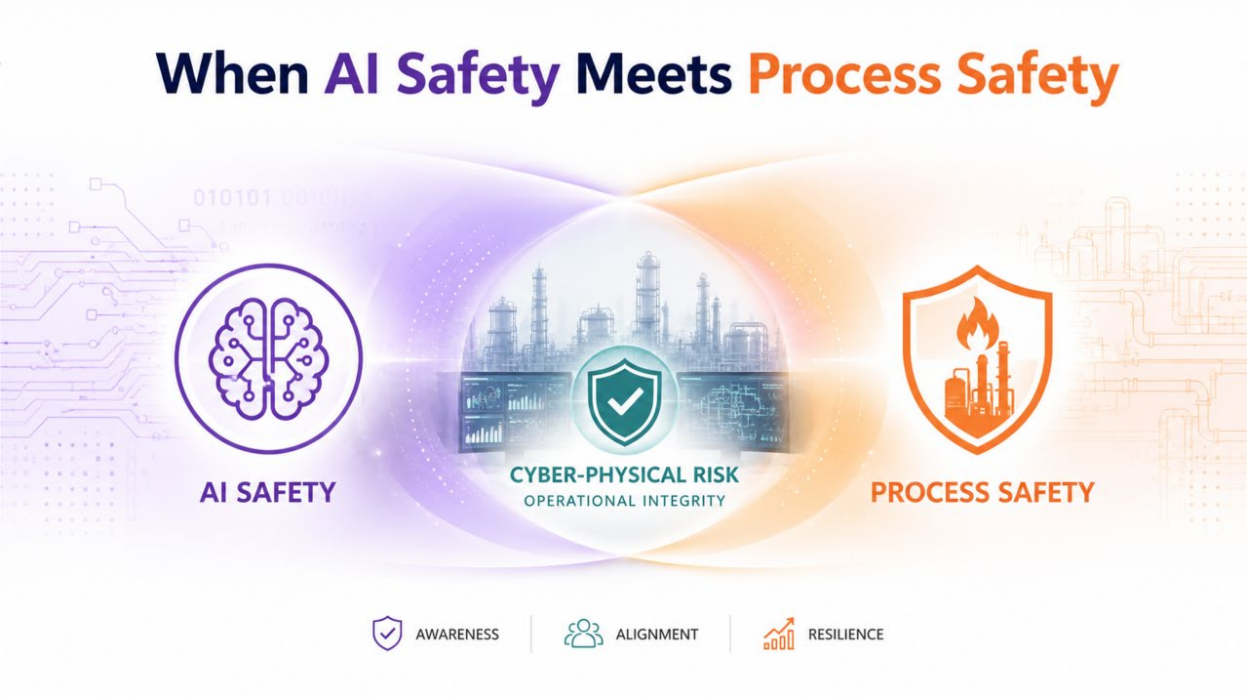


Why Protecting the AI Is Not the Same as Protecting the Process

Sinclair Koelemij

Cyber-Physical Risk Expert | Founder Cyber-Physical Risk Academy | Consultant, Speaker, Trainer, Publisher | Operational Technology | Masterclasses | Training | 45+ years in process automation.

June 11, 2026

Introduction

Artificial Intelligence is no longer a future technology waiting outside the gates of industry. It has already entered industrial operations.

Today, AI is used for predictive maintenance, quality prediction, energy optimization, anomaly detection, alarm analysis, production planning, engineering support, and operator decision support. Across the process industries, organizations are exploring machine learning, Large Language Models (LLMs) such as ChatGPT, Grok, and Gemini, digital assistants, and advanced analytics to improve efficiency, reduce costs, and support operational decision-making.

Much of today's discussion about AI in industry focuses on capability.

Can AI improve productivity?

Can it optimize energy consumption?

Can it support operators?

Can it improve maintenance?

Can it reduce operational costs?

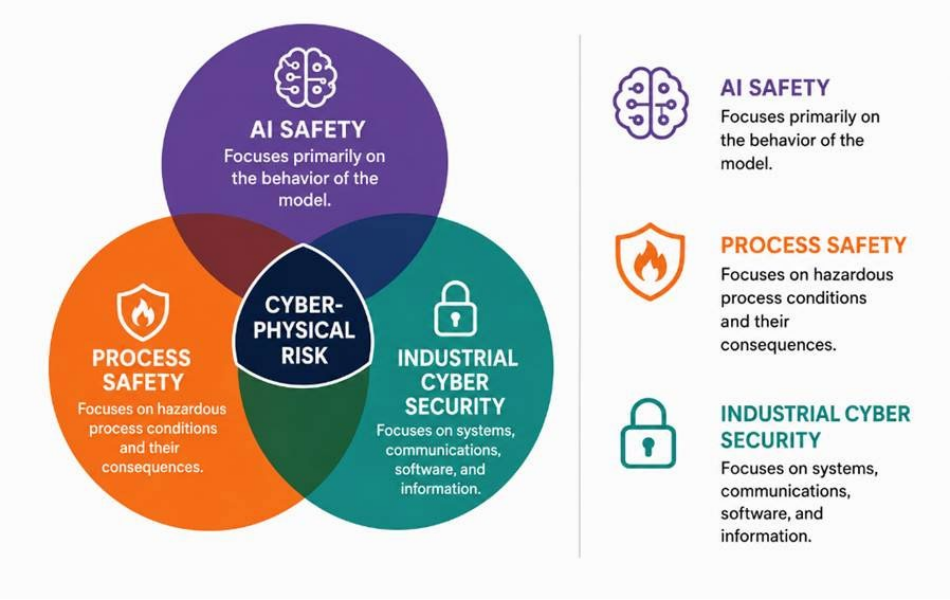
These are important questions. But from a cyber-physical risk perspective, they are not the most important questions. They focus on what AI can do, rather than what can happen when AI is given authority over a high-hazard process.

For me, the more important question is not whether AI can improve industrial performance. I have little doubt that it can and will. The more important question is what happens *when AI is granted operational authority over a high-hazard process unit*, where digital decisions can directly influence physical process conditions and their consequences.

In earlier articles, I have always argued that cyber-physical risk cannot be properly understood by treating Process Safety and Industrial Cyber Security as separate disciplines. Once digital systems can influence physical process conditions, cyber security failures can become process safety problems.

With all the new developments going on now autonomous AI introduces a third discipline into that discussion: AI Safety.

- AI Safety focuses primarily on the behavior of the model.
- Process Safety focuses on hazardous process conditions and their consequences.
- Industrial Cyber Security focuses on systems, communications, software, and information.



When AI only provides advice and does not directly control the process, the immediate risk may appear lower. But that does not mean AI safety, cyber security, and process safety can be treated as separate disciplines.

The reason is simple: advice can still influence action. AI-generated recommendations may affect operator decisions, optimization targets, control strategies, alarm interpretation, maintenance priorities, or the way process constraints are understood. Once those recommendations influence how the process is operated, AI becomes part of the cyber-physical risk landscape.

That is the important shift.

A model behavior issue, a manipulated sensor value, a compromised optimization objective, or a conventional cyber security event may all originate in different domains. But once they affect control actions, operating decisions, or process constraints, they can propagate through the same consequence pathway: a deviation from the intended operation of the physical process.

This is why increasing AI authority in high-hazard systems forces these disciplines to come together. The relevant question is no longer whether the initiating failure belongs to AI safety, cyber security, or process safety. The relevant question is how that failure can influence the process and whether existing controls and independent protection layers are sufficient to prevent unacceptable consequences.

Technically, this requires a different engineering profile. Assessing these systems is not only a matter of combining expertise in AI model behavior, data integrity, control authority, process engineering, process safety, process dynamics, and independent protection layers. It also requires understanding how failures, attacks, decisions, and process dynamics interact along a shared consequence pathway within one cyber-physical system.

Historically, these disciplines developed different responsibilities and different professional languages. Process engineers optimized production. Process safety engineers focused on hazardous scenarios, safeguards, and consequence reduction. Cyber security specialists protected digital systems against unauthorized access, manipulation, and disruption. AI safety specialists focused on model behavior, reliability, alignment, and misuse.

That separation may have been manageable when these domains had limited functional influence on each other. It becomes inadequate when AI systems use process data, influence operating decisions, adjust optimization objectives, support control strategies, or shape the operator’s understanding of plant conditions. At that point, the disciplines no longer represent separate risk domains. They become different entry points into the same cyber-physical risk problem. Therefor autonomous high-hazard systems make that separation increasingly difficult to maintain.

This is no longer purely process engineering, process safety, cyber security, process automation, or AI safety. It is cyber-physical risk engineering.

By cyber-physical risk engineering, I mean a discipline that analyzes how digital decisions, cyber compromise, automation behavior, process dynamics, and protection layers interact to create or

prevent physical consequences. Its object is not the model, the network, or the process in isolation. Its object is the complete pathway from digital disturbance to physical consequence.

Organizationally, this also means that we cannot introduce AI as just another application that is purchased, connected, and then operated as usual. Once AI becomes part of operational decision-making, it requires explicit AI governance.

So responsibilities that were traditionally separated between these disciplines begin to overlap. A decision made in one discipline can directly affect risk managed by another.

When overlap occurs we require clear ownership of AI-supported decisions, defined escalation paths, initial validation and periodic reassessment of AI-generated guidance, change management for models and data sources, continuous performance monitoring, and shared risk criteria across disciplines.

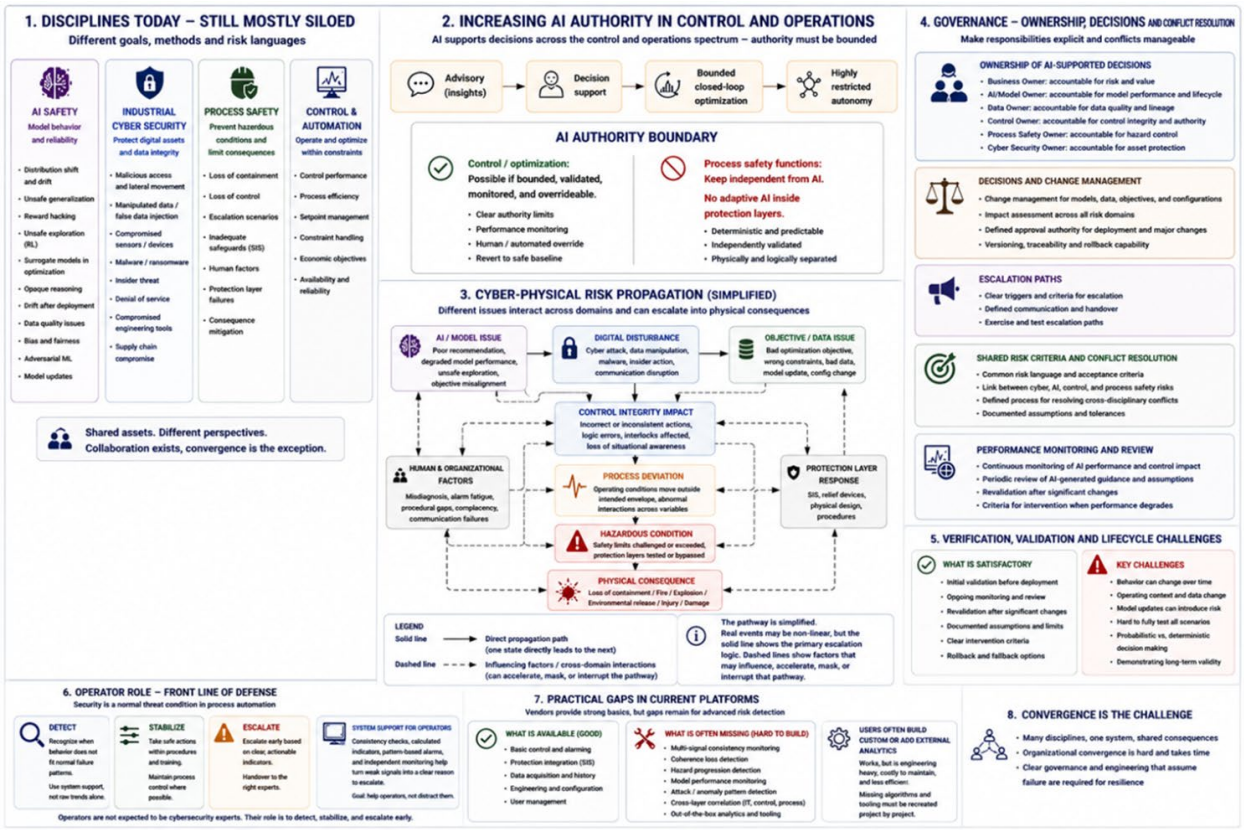
Unlike traditional software, AI systems cannot be assumed to remain valid simply because they were validated once. Operating conditions, data quality, process configurations, business objectives, and model versions can all change over time. AI is not entirely unlike humans in this respect: it may produce new insights, but that does not automatically make them good insights.

AI in combination with control may appear to be a natural fit. Control is already about interpreting process data, making adjustments, and optimizing operation within defined constraints. AI can support that task, provided its authority is bounded, monitored, validated, and overridable.

But AI in combination with independent process safety functions is a very different matter. Process safety does not exist to optimize operation. It exists to prevent hazardous conditions and limit consequences when control fails. Placing probabilistic, adaptive, or opaque AI behavior inside that protection role is therefore extremely difficult to justify and potentially dangerous.

The model below is my attempt to develop a new way of representing this shift. It should not be read as an organizational chart, nor as a claim that the disciplines have already merged. It is intended as a consequence-pathway model.

The left side represents the disciplines that still largely exist separately today: AI Safety, Industrial Cyber Security, Process Safety, and control automation. The center shows the shared pathway where failures begin to interact once AI receives operational authority: model behavior, digital disturbance, compromised objectives, control integrity impact, process deviation, hazardous condition, and physical consequence. The right side shows why governance becomes necessary: ownership, change management, escalation, shared risk criteria, performance monitoring, verification, validation, and lifecycle reassessment.



The important point is not that these disciplines disappear into one generic discipline. Process engineering, process safety, industrial cyber security, AI Safety, and process automation each retain their own methods, assumptions, and responsibilities. The important point is that autonomous high-hazard systems require an organizational capability that integrates these disciplines around one shared consequence pathway.

That capability is cyber-physical risk engineering.

This discipline may exist as a dedicated organizational function, or as a formally integrated way of working across existing disciplines. But it must have clear ownership, shared risk criteria, lifecycle governance, and authority to resolve conflicts between model performance, cyber security, control objectives, and process safety constraints. Without that integration, each discipline may optimize its own part of the problem while missing how failures propagate across the whole system.

The Next Step: From Advisor to Decision Maker

Most current industrial AI applications share an important characteristic, the AI typically acts as an advisor.

It observes data, performs analysis, identifies patterns, detects deviations, warns about emerging risks, and presents recommendations. The final decision remains with an operator, engineer, planner, or manager.

From a process safety perspective, this distinction is critical. The AI may be wrong. It may overlook information. It may draw incorrect conclusions. It may recommend actions that increase risk. However, the human remains the final barrier between the recommendation and the consequence. As a result, the safety discussion remains relatively familiar.

But industrial automation has never stood still.

Technology first provided information. Then it provided recommendations. Then it automated execution. Recently it started to make decisions. Finally, in the near future it becomes autonomous. Process automation has followed this path over decades. There is little reason to believe AI will follow a fundamentally different trajectory.

Why High-Hazard Industries Will Eventually Follow

The moment AI authority over high-hazard process units is discussed, the first reaction in many engineering circles is skepticism:

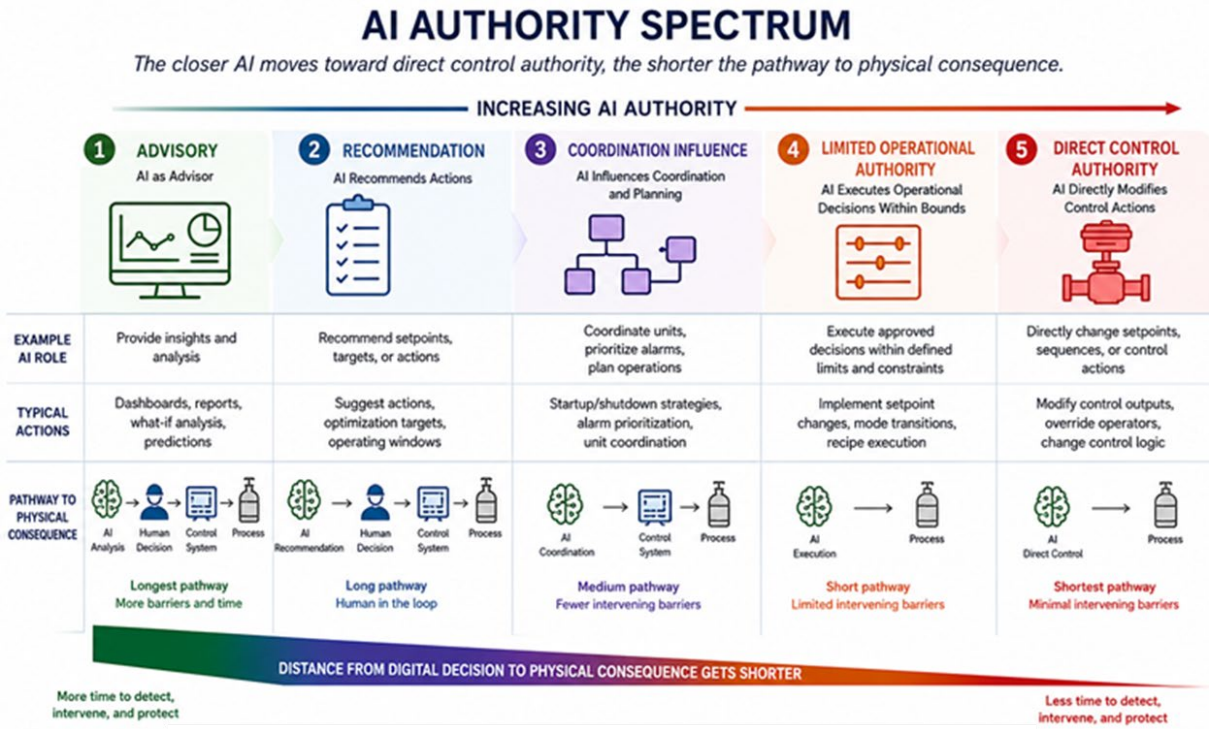
“That may happen elsewhere, but not in my plant.”

History suggests otherwise. High-hazard industries have traditionally adopted new technologies more cautiously than other sectors, but they rarely remain untouched by technological progress. The economic incentives are strong: higher production rates, lower energy consumption, improved product quality, reduced operating costs, better asset utilization, reduced staffing pressure, and more consistent operation.

Initially, AI capabilities have been introduced as advisory systems. Nowadays, they become semi-autonomous. Eventually, pressure emerges to grant greater authority to systems that consistently outperform human decision-makers in specific tasks. This does not mean fully autonomous chemical plants will appear tomorrow. But it does mean the discussion is no longer hypothetical. The question is not whether AI will gain more authority in high-hazard operations. The question is what safety philosophy should govern that authority.

The Day AI Stops Being a Tool

As long as AI remains an advisory tool, most current AI safety discussions remain valid. We can ask whether the model hallucinates, whether it can be jailbroken, whether it produces harmful outputs, whether it can assist dangerous activities, and whether it complies with policies.



These are important questions. But once AI receives operational authority, the nature of the problem changes fundamentally.

Authority exists on a spectrum. At one end, AI supports planning, analysis, and recommendations. It may adjust production targets, schedule maintenance activities, optimize resource allocation, or support operational planning.

Further along the spectrum, AI begins to influence operational coordination. It may coordinate process units, influence startup and shutdown strategies, prioritize operational responses, or shape alarm handling.

At the far end, AI approaches direct control authority. It may modify setpoints, alter operating sequences, or influence control actions.

The closer an AI system moves toward direct control authority, the shorter the pathway becomes between a digital decision and a physical consequence. That changes the safety question.

For advisory AI, the question is mainly:

Can the model generate an unsafe answer?

For AI with operational authority, the question becomes:

Can the system create an unsafe process condition?

An LLM generating an unsafe response does not directly increase reactor temperature, raise vessel pressure, or overload the compressor. But an autonomous system with operational authority potentially can. And that changes everything.

Two Disciplines Using the Same Word

Both AI Safety and Process Safety use the word safety, but they are concerned with different objects.

- **AI Safety** largely focuses on the behavior of the actor: whether the model is reliable, aligned, robust, and resistant to misuse.

- **Process Safety** focuses on the consequence: whether a disturbance can develop into a hazardous process condition and escalate into loss of containment, fire, explosion, toxic release, equipment damage, or other major accident consequences.

This distinction becomes critical once AI receives operational authority. In a high-hazard process, the initiating actor is rarely the most important question. The disturbance may originate from multiple sources. An operator intervention, configured control logic, an optimizer, an autonomous agent, or a cyber attacker manipulating data, logic, or production objectives. The process itself does not respond to any organizational categories or initiating causes. It responds to whether its physical and chemical state remains within the boundaries for stable and safe operation.

For process safety, the central question is therefore not where the disturbance started, but whether it can drive the process outside its safe operating envelope. A runaway reaction caused by an operator intervention, a software defect, a compromised optimizer, or an autonomous AI decision may have very different initiating causes. The escalation path and physical consequence can still be the same.

This is the gap that Cyber-Physical Risk is intended to address. While AI Safety evaluates the actor and Process Safety evaluates the consequence, Cyber-Physical Risk focuses on how a digital disturbance, decision, or compromise can propagate into a physical consequence.

Capability Does Not Create Risk. Authority Does.

A common mistake in discussions about AI risk is to focus on capability rather than authority.

- **Capability** describes what the AI can do: how well it predicts, reasons, optimizes, detects anomalies, or generates recommendations. Capability describes what the AI can do: how well it predicts, reasons, optimizes, detects anomalies, or generates recommendations.

- **Authority** describes what the AI is allowed to influence in the real world: whether it can change targets, modify setpoints, alter operating sequences, prioritize alarms, influence maintenance timing, or affect equipment behavior.

This distinction matters because a highly capable AI that only advises is fundamentally different from a highly capable AI that can also act. The first may mislead a human decision-maker. The second may directly influence the process state.

For me an AI system that understands how a reactor works is not necessarily dangerous. An AI system that can influence reactor operation may be. Risk always emerges when capability is combined with authority. This is especially important in high-hazard environments. Even a well-designed AI system can create risk when it is granted authority to act on imperfect or manipulated information, or when it optimizes an objective that conflicts with safe operation.

Process safety has long recognized this principle in another form: the control function should not be its own final protection layer. Independent safeguards exist because even capable control systems can fail, receive bad data, respond too slowly, or make the wrong move under abnormal conditions. The critical questions are therefore not only about model performance. They are about authority.

- What can the AI change?
- Under which conditions can it change it?
- How close is the AI to the control function?
- How short is the pathway from an AI decision to a physical consequence?

This is ultimately not an AI Safety question in isolation. It is a cyber-physical control authority question.

Trust Versus Protection

Current AI Safety largely focuses on making the actor behave correctly. Process Safety starts from a fundamentally different assumption: every actor can eventually be wrong. Operators can be wrong. Controllers can be wrong. Software can be wrong. Sensors can be wrong. So autonomous systems can be wrong. Because failure is assumed possible, Process Safety focuses on preventing failures from escalating into catastrophic consequences.

AI Safety asks:

How do we make the autonomous actor trustworthy?

Process Safety asks:

How do we ensure the process remains safe when the autonomous actor is not trustworthy?

Cyber-Physical Risk adds a third question:

Can the autonomous actor be deliberately induced to make unsafe decisions while appearing to operate correctly?

That last question is critical. It shifts the discussion from accidental failure to adversarial manipulation.

The operating window, safe operating limits, design limits, and process constraints have always been central to process safety. AI does not change that boundary. What AI changes is the pathway by which the process can be driven toward that boundary. The critical question is not whether the AI appears to behave safely. The critical question is whether the combined system can push the process outside its intended operating envelope. Once that happens, consequence management becomes the dominant concern.

Behavioral Control Versus Inherent Safety

Most current AI safeguards are behavioral controls. Reinforcement learning from human feedback, constitutional AI, safety fine-tuning, prompt filtering, output filtering, moderation systems, monitoring, and red teaming all aim to influence or constrain behavior. These measures are useful. But they do not eliminate the underlying capability of the system.

Process Safety traditionally works from a different hierarchy. It prefers eliminating hazards, reducing hazards, limiting hazardous inventories, adding independent protection layers, and preparing emergency response. This is the logic of inherently safer design.

Current AI Safety contains many protection layers, but very few equivalents of inherent safety. That difference matters once AI is connected to process authority.

Process Safety learned decades ago that the controller should not be the final protection layer. This is why autonomous systems should never become both optimizer and protector.

- The AI may optimize.
- The AI may advise.
- The AI may coordinate.

But independent protection layers must remain outside the AI reasoning layer.

Autonomous Systems Create a New Cyber-Physical Attack Surface

As I have argued in several earlier articles on cyber-physical risk, traditional process safety primarily addresses accidental failure modes. That remains essential, but it is not sufficient once digital systems can influence physical process conditions. Cyber-physical risk analysis extends the scope by also considering security-induced failure modes. I think that autonomous systems now add another layer to that problem. They introduce a decision-making function that can be wrong, manipulated, or compromised while still appearing to operate normally. As such the decision-making function itself becomes part of the cyber-physical attack surface.

Like any process automation system, an autonomous system does not directly observe reality. It observes a digital representation of the process, constructed from sensor data, communications, models, historian data, engineering configuration, and operational context. That is not new.

What changes with autonomous AI is the role of that representation in decision-making. In a conventional control function, the logic is typically bounded, engineered, and relatively

transparent. In an autonomous system, a richer and potentially less transparent decision layer may interpret that same digital representation, infer the process state, select actions, and optimize objectives. This makes the integrity, freshness, coherence, and context of the digital process representation more critical.

If that representation is manipulated, incomplete, stale, or misleading, the autonomous system may still make internally logical decisions. But those decisions may no longer be safe for the physical process.

The question is therefore no longer only:

Can the autonomous system make a mistake?

The question becomes:

Can the autonomous system be deliberately induced to make a mistake?

AI as an Attack Instrument

An autonomous AI system is not only a system that may be attacked. Once it has operational authority, it may also become the mechanism through which an attack is executed.

A compromised, poisoned, or manipulated autonomous system may continue to appear operationally normal while gradually moving the process toward less stable or less protected conditions. The attack does not need to rely on an immediate trip, shutdown, or visible disruption. It may instead work through small changes that remain individually plausible but collectively reduce the plant's margin to hazardous operation.

Such changes may include gradual movement of setpoints, erosion of operating margins, delayed maintenance decisions, altered alarm priorities, modified optimization objectives, or sequences of actions that are acceptable in isolation but unsafe in combination.

In this scenario, the critical failure is not that the autonomous system stops working. The critical failure is that it continues working toward an objective that has been modified, corrupted, or placed in conflict with safe operation.

This creates a new cyber-physical concern: the digital representation of the process becomes part of the hazard pathway. Like conventional control systems, autonomous systems act on a digital view of the process. The difference is that an autonomous system may interpret that representation more broadly, infer operating intent, select actions, and optimize objectives with less transparent reasoning.

For high-hazard environments, this makes architectural independence essential. Independent protection functions should always remain outside the authority of AI and optimization systems.

AI and optimization systems should not be permitted to modify shutdown logic, change trip settings, suppress safety functions, or bypass independent safeguards.

This requires functional and architectural separation between optimization, process control, process safety, engineering, and external data services. The purpose is not only to protect networks or systems. It is to prevent a compromised data or decision pathway from becoming a pathway to physical consequence.

Control Integrity and Time-Dependent Risk

Autonomous systems in high-hazard environments introduce more than questions of model reliability or cyber security. They create a distinct control integrity challenge.

Control integrity refers to the ability of a system to keep the process within its intended operating boundaries, in line with design intent, operational intent, and defined constraints, even when data quality, timing, model behavior, or system trust is degraded or uncertain.

This is broader than data integrity. Data may be authentic and still be incomplete, stale, poorly contextualized, or unsuitable for the decision being made. It is also broader than AI reliability. An AI system may produce a plausible recommendation that still conflicts with operational intent or reduces the margin to hazardous operation.

Control integrity asks whether the complete decision and control chain continues to preserve intended process behavior.

That includes whether the system understands the current process state, whether the data is fresh enough for the action being selected, whether the action is valid for the current operating mode, whether authority limits are respected, and whether independent protection layers remain unaffected.

This matters because cyber-physical risk is inherently time-dependent. A decision that appears correct in principle can become unsafe if it is based on stale data, delayed execution, incorrect sequencing, or an incoherent view of the current process state.

For autonomous systems, the validity of an action depends not only on what decision is made, but also on when it is made, what data it relies on, whether the system is in the expected operating mode, and whether the process still responds according to the assumptions used by the decision logic.

A manipulated sensor is not only a cyber security issue. It can degrade the AI system’s decision quality and contribute to a process safety concern. A compromised optimization objective is not only a cyber attack. It can become a control integrity failure and, under the wrong conditions, a pathway toward hazardous operation.

The central question is whether the combined system continues to control the process within its intended operating envelope under degraded, abnormal, or adversarial conditions. That is why operational and as part of it control integrity becomes a key concept for autonomous industrial systems. It connects AI behavior, cyber compromise, process dynamics, timing, authority, and independent protection layers into one risk question:

Can the system continue to maintain intended and safe control of the process when parts of the digital environment can no longer be fully trusted?

Conclusions

Current AI Safety is not wrong. It addresses an important problem: how to make AI systems more reliable, robust, aligned, and resistant to misuse. But once AI receives operational authority over a high-hazard process, that is no longer sufficient.

At that point, the primary safety object is no longer the model. It is the process state. The critical question is not only whether the AI behaves correctly, but whether the complete cyber-physical system remains within its intended operating envelope under degraded, abnormal, or adversarial conditions.

This changes the risk discussion.

Capability alone does not create the decisive risk. Authority does. A capable AI system that only advises is fundamentally different from a system that can influence setpoints, operating sequences, optimization objectives, alarm priorities, or coordination between process units. The closer AI moves toward control authority, the shorter the pathway becomes between a digital decision and a physical consequence.

This is also where the traditional boundaries between disciplines become insufficient. AI Safety evaluates the behavior of the model. Industrial Cyber Security evaluates the protection of systems, communications, software, and information. Process Safety evaluates hazardous process conditions and their consequences. Process engineering and process automation focus on making the process perform as intended. None of these perspectives alone can fully explain the risk once AI behavior, cyber compromise, data integrity, control authority, process dynamics, and independent protection layers interact within the same consequence pathway.

That is the role of cyber-physical risk engineering.

Autonomous industrial systems require an integrated capability that can assess how digital disturbances, AI decisions, manipulated data, compromised objectives, timing degradation, authority boundaries, and process deviations can converge into physical consequences.

This does not mean that process engineering, process safety, industrial cyber security, AI Safety, and process automation become one generic discipline. They retain their own methods and responsibilities. But they must be integrated around a shared understanding of control integrity and physical consequence.

The practical design principle is clear.

AI may support control, optimization, coordination, and decision support when its authority is bounded, monitored, validated, and overridable. But independent process safety functions must remain outside the authority of AI and optimization systems. Protection layers must remain deterministic, independently validated, and architecturally separated from adaptive or probabilistic decision-making.

The central challenge is therefore not simply to make AI safer.

The central challenge is to ensure that high-hazard processes remain controllable, observable, bounded, and protected when AI, data, timing, objectives, or the surrounding digital environment can no longer be fully trusted.

That is where AI Safety meets Process Safety.

And that is where cyber-physical risk engineering becomes essential. Because in a high-hazard process, protecting the AI is never enough. The process must remain safe when the AI is wrong, compromised, or no longer trustworthy.