Human, AI, and Organizational Performance (HAOP)

A Governance Framework for AI-Enabled Safety-Critical Work

Extending HOP for the age of AI performers, organizational accountability, and automated drift

by Jaina Ko, CSP

May 31, 2026

Contents

I. Executive Summary

AI is entering safety-critical work faster than most governance systems were designed to absorb. In EHS and operational settings, AI is not just summarizing documents or retrieving information. It is increasingly classifying risk, prioritizing signals, routing work, recommending action, generating controls, monitoring compliance, and shaping what human beings see, decide, and do.

Most organizations still govern these systems as a tool: a matter of data, accuracy, privacy, cybersecurity, and uptime. Those controls remain necessary, but they are insufficient in those circumstances where AI materially influences operational judgment or workflow execution. At that point, AI becomes a performer within the work system, introducing variability, shaping outcomes, and failing in ways that differ from human error.

Human and Organizational Performance corrected an earlier safety mistake: the belief that workers are the primary variable to control. HOP showed that human behavior is shaped by context, incentives, constraints, leadership response, and the gap between work-as-imagined and work-as-done. HAOP, *Human, AI, and Organizational Performance*, extends that logic for AI-enabled work systems.

HAOP recognizes three interacting performers. The human performer adapts under real operating conditions. The AI performer optimizes based on data, signals, permissions, constraints, and architecture. The organizational performer shapes both through governance, incentives, resources, metrics, authority, procurement, and tolerated tradeoffs.

The central claim of HAOP is simple: safety-critical AI cannot be governed by model performance alone. It must be governed as part of a socio-technical system where human adaptation, AI optimization, and organizational signaling interact.

This paper introduces the HAOP framework, defines the distinct failure signatures of each performer, and proposes a practical diagnostic - the **True Function Test** - for evaluating whether an AI-enabled workflow actually produces the safety outcome it claims to pursue, or merely produces a cleaner representation of safety.

HAOP is not a no-blame model. It is an accountability-by-control model. Responsibility should not collapse onto the most visible person at the point of failure. It should be mapped in advance to the people, teams, and functions with control over the relevant action, signal, constraint, permission, resource, metric, deployment decision, verification point, or escalation path.

For EHS professionals, HAOP provides a way to govern AI without abandoning the lessons of HOP. For AI technologists, it identifies operational failure modes that model metrics alone will not reveal. For AI governance leaders, it connects accountability, assurance, human oversight, and drift to real work. For academics, it offers a conceptual bridge between safety science, organizational theory, and AI-enabled socio-technical systems.

The central test is not whether an AI workflow looks modern, efficient, compliant, or data-driven. The test is whether it remains grounded in operational reality before its outputs become consequential.

Part I: The HAOP Framework

II. The Problem: AI Is Arriving and Is About to Amplify Safety's Oldest Mistake

Most workplaces still operate within control-based systems designed for a simpler era, when hierarchy, rules, and compliance were assumed to produce safety. Much of modern industrial management inherited assumptions from Taylor's scientific management, which turned workers into measurable units of labor and treated human error as deviation from an otherwise perfect plan.[1]

For nearly a century we have chased "zero harm" through more audits and stricter rules. The measurement of choice was progress against a lagging metric, Total Recordable Incident Rate. Organizations layered on valuable methods such as the hierarchy of controls, behavior-based safety, and versions of Plan-Do-Check-Act. These tools had real benefits, but they also incentivized corrosive practices that blamed the worker and managed the metric rather than the risk, and in too many cases, manipulating the data outright.

Meanwhile, the gap Hollnagel named as *work-as-imagined* and *work-as-done* kept widening.[2] Culture surveys tried to explain it and more new initiatives got launched. Yet recordable rates plateaued and serious injuries and fatalities persisted, challenging the precept that all accidents were preventable.

Human and Organizational Performance emerged as a correction to this broken frame. Grounded in reality, HOP recognized and accepted something fundamental that the old frameworks denied: *to err is human*. From this basic truth, the five principles were built out: people are fallible, context drives behavior, blame fixes nothing, learning is vital, and leader response matters.[3] By shifting from control to curiosity, organizations could move toward systems that produce what they claim to pursue and toward learning rather than enforcement.

But conditions are changing again. The workforce is shrinking, experienced people are retiring with their knowledge, and organizations are rushing to fill the gap with AI. In EHS work, AI is still governed as an IT and data risk, emphasizing privacy, accuracy, and uptime. The governance problem is not AI adoption itself. The problem appears when AI moves from information support into operational influence: classifying risk, routing work, prioritizing signals, recommending controls, approving actions, or shaping what a human reviewer sees first. At that point, the system is not just storing or displaying information. It is participating in the production of work.

Here is the collision. Deploy AI into a system still shaped by the old assumptions of control and compliance, and AI will amplify that philosophy. If an organization still believes, even quietly, that people are the problem, AI becomes a faster, colder way to monitor, blame, and punish. In a complex, fast-moving environment, automated judgment built on that premise is not merely ineffective. It is dangerous.

III. The Foundation: What HOP Established, and Where It Stops

Human and Organizational Performance established that human behavior is not random defect. It is locally rational performance shaped by context, constraints, incentives, tools, training, leadership response, and operational reality. It moved safety away from fixing people and toward understanding and improving the systems in which people work. Todd Conklin's five principles of HOP:[3]

1. Error is normal.

HOP rejects the assumption that perfect compliance is a realistic safety strategy. Humans forget, misread, adapt, get tired, improvise, and act on incomplete information. The goal is not to create perfect workers. The goal is to design systems where predictable human error does not become catastrophic. HOP joined Dekker's "Safety Differently" and Resilience Engineering movements in moving the field away from "fixing the human" and toward being able to fail safely.[4],[5]

2. Context drives behavior.

HOP made context central. People do things that make sense to them at the time given their immediate goals, their knowledge, their focus of attention, the norms around them, and the specific pressures of that exact moment. This is local rationality. As Dekker put it, no one goes to work planning on doing a bad job.[4]

3. Blame fixes nothing.

Blame may satisfy the need for consequence, but it hides the conditions that made the event possible, and it suppresses the reporting and weak signals that a system needs to stay informed that something is drifting.

4. Learning is vital.

HOP pushed safety away from only studying failures. The conditions that produce failure are the same conditions that produce success, which makes normal work a critical data source: studying when things go right yields far more data than waiting for them to go wrong.[6] The workarounds, adaptations, informal practices, and quiet compensations show how the system is actually functioning, and they reveal the gap between work-as-imagined and work-as-done.

5. How leaders respond matters.

What leaders do after error matters more than the error itself. A response of curiosity improves learning and lets the contributing factors get corrected. A response of punishment or defensiveness loses the information the system depends on, and the system grows more brittle.

HOP explains how humans perform inside systems. It corrected a deep misunderstanding of human performance and gave us a better way to understand workers, leaders, procedures, context, drift, learning, and accountability in those human-centered systems.

It also made the organization visible, because the organization supplies much of the context that shapes human behavior. And a system executes signals, not objectives. If an organization states an

objective of reducing risk but is only measuring TRIR, PPE violations, or throughput, it is performing through the signal architecture it created, which may be working directly against the objective it claims.

But HOP was not designed for a world with another performer in it.

IV. The Three Performers

HOP remains necessary, but it is no longer sufficient on its own for AI-enabled work systems. It explains how humans adapt inside organizational conditions. It does not, by itself, provide a vocabulary for optimization systems that classify, route, recommend, suppress, escalate, or act across workflows.

HAOP - Human, AI, and Organizational Performance - extends HOP when AI crosses a functional (even if somewhat blurry) threshold from tool to performer. AI remains a tool when it supports a bounded human task: summarizing a report, querying a database, drawing a procedure, or responding to a prompt. In those cases, a human initiates the task, a human reviews the output, and the AI's role is narrow and more bounded.

AI becomes performer-level when it no longer just supports a human task, but materially shapes the sequence, priority, visibility, recommendation, approval, routing, or execution of work. At that point, the AI becomes a part of the work system itself, introducing variability and shaping outcomes.

An autonomous AI agent makes this line easier to see. An agent may pursue objectives over time, make decisions, and act without human initiation at each step. When it fails, it does not fail through human-like intention, fatigue, fear, or local rationality. It fails through local optimization mechanisms such as mis-specified goals, weak constraints, degraded data, excessive permissions, tool misuse, goal hijacking, or cascading action. This is not theoretical. The tech industry itself recognizes the instability of these tools. When OWASP's Agentic Security Initiative published its *Top 10 for Agentic Applications* (2026 edition), it documented 10 failure categories specific to agents, including goal hijacking and cascading failures.[7]

AI remains a tool in many use cases. But when an AI system materially shapes what people see, decide, prioritize, approve, route, or do, it becomes a performer within the work system.

> ***Performer:*** *Any entity that can take action, influence outcomes, and introduce variability.*

HOP taught safety leaders to ask how the system looked to the worker. HAOP adds two additional questions: how did the system look to the AI, and what did the organization design the system to optimize?

HAOP uses "performer" functionally, not morally. A performer is any entity or organized system that can act, shape outcomes, and introduce variability. A human performer acts through judgment and adaptation. An AI performer acts through model output, tool use, ranking, routing, prediction, or workflow execution. An organizational performer acts through governance, incentives, metrics,

resources, authority, procedures, procurement, and tolerated tradeoffs. Accountability, however, still traces to people, roles, and governing bodies with control.

The organization as performer has grounding in organizational studies and safety science canon. In *A Behavioral Theory of the Firm*, Cyert and March[8] rejected the idea that organizations are neutral containers or that they behave like single rational profit-maximizing actors. In their seminal work, they described how organizations behave like coalitions of participants with multiple (potentially conflicting) goals: production, sales, safety, quality, cost, labor relations, market share, executive priorities, departmental interests, and local incentives.

Vaughan developed the concept of *normalization of deviance* in her study of the Challenger disaster, where O-ring erosion and blow-by warning signs were ignored. The core thesis was that organizations can gradually accept deviant conditions as normal when those conditions repeatedly occur without immediate catastrophe.[9]

Dekker's *Drift into Failure: From Hunting Broken Components to Understanding Complex Systems*[10] argued that complex system failures should not be explained by hunting for one broken component or one bad actor. His work pointed out that people and organizations make small adjustments that seem reasonable at the time. They borrow a little from safety margins to meet production, cost, schedule, staffing, or efficiency demands. Each adjustment appears tolerable because nothing bad happens immediately. Over time, *systems drift* closer to failure because the movement is incremental and does not feel like a major risk decision.

Therefore, naming the organization a performer is not a new claim; it operationalizes what safety science and organizational theory already assert.[11]

HAOP recognizes three interacting forms of performance: human performance, machine/AI performance, and organizational performance. The organization is the containing performer because it designs, constrains, authorizes, measures, and normalizes the conditions under which the human and AI performers operate.

HOP's corrective against worker blame held that humans are not the problem but part of the solution. HAOP keeps the human as the central operational performer, adds AI as a distinct performer that behaves and fails differently, and names the organization as the containing performer responsible for the interaction space between them. In AI-enabled work, there are no longer only human performers inside organizational systems. AI increasingly performs work, and the organization performs the conditions under which both human and AI action becomes possible, rewarded, constrained, or ignored. That makes the organization not background context, but a containing performer.

This is what gives the framework accountability precision. One of the critiques of HOP has been that it leaves no one accountable, because "the system" did it. But that critique emerges when HOP is adopted as a nice new idea, not as the rigorous culture transformation it was designed to be. Accountability follows control, and the ability to create change is an essential element of control. People always control some actions, with hierarchy carrying great influence. AI systems perform delegated functions. Organizations control the conditions, incentives, authority structure, and validation systems that shape both.

- **The human performer:** Already well understood. We perceive, adapt, hesitate, compensate, comply, deviate, speak up, stay silent, and make tradeoffs under real operating conditions, acting through cognition, interaction with physical reality, judgment, and adaptation.
- **The AI performer:** Still being understood. AI classifies, predicts, generates, routes, recommends, approves, prioritizes, escalates, suppresses, and optimizes based on data, signals, constraints, and architecture, acting through model output, tool use, workflow execution, ranking, routing, prediction, or automated decision support.
- **The organizational performer:** Authorizes, funds, measures, rewards, constrains, trains, validates, ignores, normalizes, and assigns accountability, acting through institutional design, incentives, norms, controls, resource allocation, governance, and sanctioned meaning.

The organization is the containing performer because it creates the conditions under which human and AI performance becomes possible, constrained, rewarded, normalized, hidden, or corrected. It contains the two operational performers - the human beings and the AI systems - but it also performs through them.

V. How the Performers Fail Differently

The first truth that HOP surfaced was simple, and so was the reaction to it: humans make mistakes - it is normal. Take a breath. Now let us understand why the mistake happened.

That posture - understanding before blame - extends to all three performers, though for different reasons. The human is owed the breath because blame wounds and silences. The AI and the organization are owed the same analytical patience because rushing to assign cause makes us blind to how the system actually failed. And the three do not fail the same way. You cannot govern what you cannot recognize, so before any framework can help, we have to see each performer's distinct signature of failure.

5.1 The Failure Signatures

> ***Failure signature:*** *The recognizable pattern that tells you which kind of performer failure you are seeing.*

- **Human failure signatures:** bounded attention, local rationality, fatigue, normalization, silence, automation overreliance, and adaptive overload.[12],[13] HAOP names one emergent failure pattern, ***cognitive overrun***: a condition where a worker remains formally accountable for verifying AI output while its rate, density, or ambiguity has exceeded their capacity to verify what matters.
- **AI failure signatures:** wrong-signal optimization, *confident incompetence* (fluent, authoritative, or decisive outputs that are wrong), context blindness, specification failure, data degradation/compression loss, and speed-scale amplification.[14]

- **Organizational failure signatures:** signal-objective mismatch, accountability gaps, procurement ahead of governance, symbolic oversight, normalized deviance, compliance theater, and under-resourced controls.

There are many more ways to fail than can be listed here, certain failure signatures are distinct to the performer involved. In broad terms, humans adapt, AI optimizes, and organizations signal. AI failures are often described with anthropomorphic language that attributes fatigue, intent, desire, judgment, or understanding to the system. AI failures may resemble human mistakes at the surface, but they do not arise from human-like motivation, experience, or error. Its mechanisms are AI-specific: optimization against the wrong signal, weak constraints, degraded data, flawed architecture, excessive permissions, or other failures of data, architecture, constraint, deployment, or governance. When human adaptation, AI optimization, and organizational signaling become misaligned, the system can move toward failure while still appearing functional for a period of time.

5.2 The Stakes Are Epistemic: System Information Can Rot

The current AI replacement narrative assumes that expertise can be extracted from humans, embedded into software, and then scaled while the human labor system is reduced or removed. In safety-critical work, that assumption is structurally dangerous.

AI systems do not create new operational knowledge from nothing. They compress, abstract, and reproduce patterns from the data and feedback they are given. Without continual grounding (refreshing) in real-world human expertise, these systems can narrow toward the average, lose sensitivity to rare cases, and erase the low-frequency signals that safety work exists to detect.[15]

Degrading information may not be an entirely new problem created by AI. Information passed from person to person and department to department would degrade, with tiny errors in transmission amplifying as it moved along. AI information, though, often rests on the assumption that it is hard-coded and unchanging while being transmitted. The assumption includes the belief that AI can absorb the intellect and experience of humans and keep it in a stable form - in perpetuity. These assumptions are a trap.

This matters because safety-critical work often depends on weak, contextual, and tacit signals: the mechanic who recognizes an unusual vibration, the operator who knows a machine has a non-OEM part, the supervisor who has seen a failure pattern before, or the experienced worker who senses that conditions “do not look right” despite normal instrumentation. These signals are difficult to formalize as data, but they are often the difference between early intervention and serious harm.[16]

The risk scales into organizational design. When companies replace entry-level and intermediate roles with AI, they do more than cut headcount - they weaken the pipeline that produces future senior judgment. Senior expertise is not instant; it develops through years of field exposure, minor mistakes, tacit learning, and contact with real conditions. An organization that extracts existing expertise into AI while eliminating the roles where future expertise forms creates a double collapse: degradation of the informational system AI depends on, and degradation of the human system that keeps the information grounded.[17]

AI also changes the time dimension of safety. Traditional drift develops slowly enough for humans to notice anomalies, huddle, improvise, and intervene. AI performers are built for speed and scale. When one acts on a distorted representation of reality, the latency buffer disappears, and the system can move from normal operation to significant consequence faster than human adaptation can respond.

The central governance failure is the *aesthetic illusion*: mistaking a clean, fluent, low-noise digital artifact for operational competence. The recent "tokenmaxxing" episode at Amazon shows the mechanism: once token consumption became the visible metric, developers maximized the measured signal, inflating usage with low-value AI calls rather than producing better work. This continued until Amazon removed the leaderboard and shifted to a measure of meaningful deployment.[18],[19] When the signal becomes the target, the system produces the signal.

For HAOP the implication is direct: AI cannot be governed as a tool once it becomes a performer in safety-relevant work. Human expertise must remain an active grounding mechanism, not a training input to be extracted and discarded. The governance question is not only whether AI output is accurate today, but whether the organization has preserved the human, technical, and organizational feedback loops that keep the system anchored to operational reality over time.

The stakes are epistemic, operational, and human: replace the people who ground the system in reality, and you risk corrupting what the system knows, accelerating how it fails, and destroying the expertise needed to recover.

5.3 Applied Case: The Warehouse That Drifted Toward a Fire

A fire captain described a scenario that shows how AI optimization can create risk without any obvious bad decision in the moment.

An AI-driven warehouse system is built to increase storage capacity and throughput. It continuously adjusts layouts, tightens storage patterns, and reallocates space to demand. Each change is small. Each snapshot looks efficient. From the dashboard, the system appears to be working exactly as intended.

But it was optimized for space and throughput, not fire protection. Because critical safety constraints were never built in, the optimization gradually creates conditions the dashboard cannot recognize as dangerous: sprinkler clearance drops below required thresholds, egress paths narrow during peak operations, fuel-load density concentrates. No alarm sounds. No one makes an obviously reckless decision. The system is doing exactly what it was designed to do: maximize the signal it was given.

Then a fire starts. Sprinklers underperform because clearance is obstructed. Egress is slower because pathways narrowed. Concentrated fuel load accelerates the event. What looked highly optimized becomes a major loss.

This is not ordinary human drift. It is unconstrained optimization. The system did not gradually ignore a rule it understood - it was never designed to treat fire protection as a non-negotiable constraint. Traditional HOP explains how human performers drift under pressure, adapt locally, and

normalize degraded conditions. AI performers fail differently: they do not rely on judgment, unease, or contextual hesitation. They optimize the specified signal unless explicit constraints prevent them.

The organizational performer is central here too. The organization chose the objective, approved the system, defined the success metric, and failed to specify the safety boundary. The failure was not only in the AI. It was in the organizational design that let efficiency become the dominant signal without embedding fire protection as a hard constraint.

VI. HAOP Operating Principles

HOP established the human-performance foundation by explaining how humans perform in systems. HAOP incorporates that foundation, but its operating principles govern a three-performer system: human adaptation, AI optimization, and organizational signaling. These principles come from the interaction dynamics through which these performers jointly produce outcomes in AI-enabled systems.

These HAOP Operating Principles transition the focus from the crisis (AI colliding with an old, control-based operating system) to the solution (a new socio-technical framework). They are an expansion of the core truths of HOP, updated for a world where algorithms are active team members.

1. **Performance is distributed.** Operational outcomes are produced by the interaction of human performers, AI performers, and organizational performers.
2. **Each performer has a distinct failure signature.** Humans tend to fail through adaptive overload - encountering obstacles and traps while trying to make the system work. AI becomes a performer when it materially shapes work and tends to fail through misaligned optimization and data degradation - acting on narrow directives with zero understanding of physical reality. AI does not possess situated operational judgment; it acts on representations of reality, not on physical contact with the worksite, equipment history, informal norms, or consequences. Organizations are the containing performer and tend to fail through operational drift, distorted metrics, and normalized deviance - slowly detaching the hierarchy from frontline reality.
3. **Systems execute signals, not intentions.** An organization may intend safety, but the system executes what is measured, rewarded, automated, and enforced. Stated objectives do not govern behavior unless they are translated into valid signals, constraints, feedback loops, and accountability structures.
4. **Grounding is a control.** AI-enabled systems must remain anchored to operational reality: real work, field expertise, equipment variability, regulatory constraints, weak signals, and consequences short of harm.[20] Repeatedly looping lossy data structures without grounding leads to system-wide operational failure risk.
5. **Human oversight is work, not a label.** Oversight requires time, competence, authority, attention, access to underlying data, and the protected ability to pause, verify, escalate, or intervene.[21] Without those conditions, “human in the loop” is symbolic.

6. **Constraints must be designed before optimization.** AI performers will optimize the signal they are given. Safety-critical boundaries must be specified as constraints before deployment, not discovered after failure.[22] *Human-in-the-design* validation points are necessary friction that prevents optimization from outrunning verification.
7. **Accountability follows control.** Accountability has three forms: *consequence, ownership, and design*. Consequence is reactive; it assigns responsibility after failure. Ownership is admirable but dependent on character, and therefore unreliable on its own, as it depends on individuals or senior leaders accepting responsibility for outcomes. Design is structural; it defines responsibility before failure occurs. In AI-enabled systems, accountability must be mapped to the person, team, or function with control over the action, signal, constraint, permission, resource, metric, deployment decision, or response.
8. **Learning must include all three performers.** Investigations must examine what the human adapted to, what the AI optimized, and what the organization signaled, funded, ignored, rewarded, or normalized.

VII. Accountability: Follow Control, Not Visibility

The oldest objection to systems thinking is that it dissolves accountability: if the system did it, no one is responsible. One of the founding principles of HOP is "blame fixes nothing." Some have read this as there being no accountability at all. By blaming "the system," individuals have sought to evade responsibility for their actions. Blame is the act of pinning accountability as consequence on one primary performer. It is narrow, focused, and (sometimes intentionally) creates blindness to other areas of accountability.

HAOP answers this problem directly by naming the organization a performer. This adds accountability. It does not remove it. The organization is not a scapegoat or a ghost. It acts through governance, incentives, resources, procedures, metrics, and constraints.

Accountability should follow control, not visibility. While HAOP recognizes three performers, accountability ultimately traces to humans, roles, teams, functions, officers, or governing bodies with control. Workers hold control over choices within their authority and capacity. Technical teams hold control over how the AI was selected, validated, constrained, and monitored. Supervisors hold control over how work is assigned, how exceptions are escalated, how controls are verified, and whether weak signals are acted on or normalized. Senior leaders hold control over resources, priorities, incentives, and risk tolerance. Executives hold control over the management systems: assurance, authority, oversight, and acceptable tradeoffs.

These are not mutually exclusive shares of one responsibility. They are distinct control obligations. Each is accountable within the boundary of the control it actually held: its authority, knowledge, resources, ability to intervene, and duty to verify. When a worker is held to account, that does not discharge the executive whose decisions shaped the conditions; both held control, and both answer for it.

7.1. Three Forms of Accountability

HAOP distinguishes three forms of accountability:

Consequence accountability is reactive. It assigns responsibility after failure - who is disciplined, cited, sued, removed, retrained, or otherwise held responsible.

Ownership accountability depends on the voluntary acceptance of responsibility by an individual or a leader. It is valuable, but unreliable on its own because it rests on character, courage, and culture.

Design accountability is structural. It operates in real time, before failure. It assigns responsibility in advance by defining who controls what, who verifies what, who has authority to stop the workflow, who monitors drift, and who approves deployment. Weak signals are recognized as attention points, and methods for surfacing and escalating them are established and assigned. The reality that error happens is built in, and methods for ensuring visibility are designed. HAOP is built on design accountability.

Consequence and ownership will always exist, but only design accountability can be engineered in advance. Only design accountability prevents responsibility from collapsing onto whoever happens to be standing closest when something goes wrong.

7.2. Accountability Must be Mapped to Control

If design accountability is the goal, control is the map. The question is not only who was closest to the failure, but who had authority over the conditions that made the failure possible: the action, signal, constraint, permission, metric, resource, verification point, escalation path, or deployment decision.

After an incident, the question is not simply, “Who made the error?” The first step is to identify the failure signature. Was it human adaptation under overload? Was it unconstrained AI optimization? Was it an organizational signal that contradicted the stated safety objective? The failure signature identifies the type of breakdown. The accountability trace then asks where control was supposed to exist, who held it, and where it failed.

That trace must move downward, laterally, and upward. It may include the worker’s decision. It may also include the supervisor’s work assignment, the engineer’s validation boundary, the product team’s optimization target, the executive’s risk tolerance, the procurement decision, the metric that rewarded speed over verification, or the governance process that allowed a workflow to become consequential without adequate review or pause capability.

This is the hard part. **Accountability-by-control** requires organizational courage because it does not stop at the most visible person. It follows authority, incentive, verification, and permission wherever they sit.

This is also where HAOP aligns with Just Culture. Just Culture protects learning by recognizing that people should not be punished for actions, omissions, or decisions that are reasonable given their training, experience, and operating conditions.[23] It also preserves accountability for reckless conduct,

willful violations, gross negligence, destructive acts, or knowing disregard of risk. The point is not to remove responsibility. The point is to locate responsibility accurately.

What accountability does not mean:

- It does not mean every person connected to a system is equally responsible.
- It does not mean front-line workers are excused from reckless or willful conduct.
- It does not mean executives are personally responsible for every local error.
- It does not mean AI failures are treated as mysterious technical events with no human or organizational owner.

> *Accountability means responsibility is mapped to control: authority, knowledge, resources, incentives, verification duties, and ability to intervene.*

What HAOP establishes is the discipline of designing accountability in advance: mapping responsibility to the person, team, or function with control over each action, signal, constraint, permission, metric, and deployment decision - before the system goes live, *while that mapping has the greatest amount of influence.*

Design accountability is not merely aspirational. It is enforceable in some jurisdictions. In *Gibson v Maritime New Zealand*,[24] the former chief executive of Ports of Auckland was convicted for failing to exercise due diligence as an officer after a worker was killed by a falling container. In March 2026 the High Court upheld his conviction, the NZ$130,000 fine, and a NZ$60,000 costs award.[25] Ports of Auckland was separately held accountable as the operating entity. The significance of the case is not that every executive is automatically liable for every incident. It is that officer accountability was traced to governance-level control:[26] whether critical risks were understood, whether controls were effective in practice, and whether assurance processes verified work as actually done.

That is the principle HAOP carries into AI-enabled safety work. The question is not, “Who was visible at the moment of failure?” The question is, “Who controlled the conditions, permissions, signals, constraints, and verification points that made the failure possible?”

This approach is not about blame. It is disciplined accountability across human, AI, and organizational control points. Its purpose is to make accountability more accurate, more distributed, and more useful before harm occurs.

Part II: Applying HAOP

The preceding sections define HAOP as a governance framework. The next section translates the framework into a practical diagnostic. The purpose is not to certify an AI system as safe. It is to test whether an AI-enabled workflow has True Function: whether it produces the safety outcome it claims to pursue, rather than merely producing a representation of safety.

VIII. The True Function Test: A HAOP Diagnostic

8.1 Purpose of the Test

This framework is not only a way of seeing. The concepts can be turned into a diagnostic any organization can run on its own AI-enabled safety work. Here is one that can be used, today, within your organization.

The purpose is simple: determine whether the workflow actually produces the safety outcome it claims to pursue, or whether it mainly produces a representation of safety - a dashboard, metric, report, checklist, approval, or compliance artifact.

The True Function Test begins with nine core diagnostic questions. The Governance Grid then expands those questions across the three performers and adds two framing rows: what work each performer is actually doing, and what control each one needs.

8.2 The Nine Diagnostic Questions

For any AI-enabled safety workflow, ask:

1. What outcome does this system claim to produce?
2. What signal is it actually optimizing?
3. What human judgment is it relying on?
4. What is the AI allowed to do without human initiation, review, or intervention?
5. What organizational incentive is shaping human behavior and AI optimization?
6. Where does verification occur before the output becomes consequential?
7. Who has authority to stop the workflow?
8. What would failure look like before harm occurs?
9. What weak signals would the system likely erase?

A workflow has **True Function** when it actually produces the outcome it claims to pursue - not just the documentation, dashboard, metric, or compliance artifact that represents the outcome.

Where the answers reveal a gap between the claimed outcome and the optimized signal, the workflow is performing a **fake function**: it looks like it is managing safety while it is actually managing something else.

8.3 Minimum Go-Live Requirement: The Ability to Pause

Stop Work Authority is recognized by many organizations and remains necessary where danger is imminent. But in practice, it is often framed as an emergency control: stop the job, halt production, trigger escalation. That framing can create reluctance to use it when the safer action is not a full stop, but a pause long enough to verify what is happening.

Much of operational safety depends on an earlier and less dramatic control: the protected ability to pause the flow of work. Workers, reviewers, supervisors, and technical owners need the protected ability to pause when something is unclear, when conditions have changed, when an input or output does not make sense, when a weak signal appears, or when the next step would turn uncertainty into consequence.

This is not a full stop. It is a pause long enough to ask, verify, adjust, escalate, or correct before the work continues. In AI-enabled workflows, this matters because an AI output may shape priority, approval, routing, escalation, or action before a person has had a realistic chance to challenge it.

The principle, though, is broader than AI. Safety-critical work depends on the ability to pause when the work or operating conditions no longer match the plan: when conditions change, a signal is unclear, an input or output does not make sense, or the next step would turn uncertainty into consequence. A workflow that cannot be paused long enough to ask, verify, adjust, or escalate is not under control.

No safety-critical workflow should go live unless people know when they may pause the work, how the concern is escalated, who must respond, and how the person raising the concern is protected from retaliation or penalty.

8.4 The Governance Grid

Run each question across all three performers:

	Diagnostic question	Human performer	AI performer	Organizational performer
1	**What work is being performed?**	judgment, adaptation, review, escalation	classification, generation, routing, optimization	work design, approval, resourcing, metrics
2	**What outcome is claimed?**	What is the person expected to accomplish?	What is the AI expected to produce or influence?	What business or safety outcome is the organization claiming?
3	**What signal is optimized?**	What behavior is rewarded, pressured, or normalized?	What data, metric, prompt, target, or proxy is the AI optimizing?	What metric, incentive, or executive priority is shaping the workflow?
4	**What judgment is relied on?**	What must the human notice, interpret, challenge, or verify?	What does the AI classify, recommend, generate, approve, or route?	What assumptions has the organization made about human capacity and AI reliability?
5	**What behavior is allowed?**	What shortcuts, adaptations, or overrides are tolerated?	What can the AI do without review or intervention?	What permissions, resources, and authority has the organization granted?

	Diagnostic question	Human performer	AI performer	Organizational performer
6	**Where does verification occur?**	Does the human have time, skill, context, and authority to verify?	Is the AI output checked against reality, constraints, or known limits?	Has the organization defined points of consequence and required validation there?
7	**Who can stop the workflow?**	Can the worker or reviewer pause or challenge the process?	Are there technical limits, hard stops, or escalation triggers?	Has stop-work authority been designed, protected, and resourced?
8	**What can early failure look like?**	Fatigue, silence, workarounds, attention or judgment miss, compliance miss, overtrust	Confident incompetence, false negatives, signal substitution, constraint failure	Dashboard confidence, compliance theater, accountability gaps, normalized deviance, weak oversight
9	**What is shaping behavior?**	context, fatigue, training, peer norms	data, prompts, model limits, permissions	incentives, policies, budgets, priorities
10	**What weak signals may be erased?**	Hesitation, informal warnings, tacit knowledge, near misses	Edge cases, anomalies, uncertainty, missing context	Bad-news resistance, bottlenecks, friction, dissent, underreported risk, resource constraints
11	**What control is needed?**	time, authority, training, stop-work ability	validation, constraints, monitoring, auditability	governance, ownership, escalation, consequence design

8.5 Red Flags

The workflow is likely performing a fake function if:

- The claimed objective is safety, but the optimized signal is speed, closure rate, adoption, cost reduction, or dashboard completion.
- The human is credited as a control but lacks time, authority, context, or competence to intervene.
- The AI output becomes consequential before verification occurs.
- The workflow produces cleaner documentation without improving field control.
- Weak signals are converted into low-resolution summaries before anyone with authority sees them.
- No one can clearly identify who owns the signal, constraint, permission, validation point, or stop-work decision.

8.6 What the Test Does Not Do

The True Function Test does not ask whether the workflow looks modern, efficient, compliant, or data-driven. It asks whether the workflow remains anchored to operational reality before the output becomes consequential.

IX. Beyond EHS: Why the Framework Extends

Although HAOP begins in safety, the structure it describes is not unique to safety. At the level of the containing performer, the governance problem repeats across HR, Quality, and organizational governance. Each is a domain where operational performers act within an organizational performer that can itself drift. Wherever humans and AI perform inside an organization that shapes both, the same three-performer logic applies, and the same questions of signal, control, and accountability follow. HAOP is a safety framework first, but it is a governance framework by structure.

X. Conclusion

AI did not create the metric-for-mission and blame-the-worker problem we find in organizations today. It will amplify and accelerate it. It is a powerful optimizer, and when deployed into a system that already mistakes the metric for the mission or assigns blame to the nearest person to the incident, it will pursue that signal faster, more consistently, and at greater scale than a human organization could manage manually.

Existing safety frameworks remain valuable. AI, though, introduces specific operational dynamics that those frameworks were not designed for. HAOP gives organizations a way to account for those dynamics by recognizing and treating the three performers now shaping safety-critical work: the human who adapts, the AI that optimizes, and the organization that contains, directs, and legitimizes both. Seeing them clearly is the precondition for governing them well and for keeping the human judgment that grounds the whole system present, capable, and accountable, rather than extracted and discarded.

HAOP offers two things: a conceptual framework for seeing human, AI, and organizational performance as interacting sources of safety-critical outcomes, and a practical diagnostic for testing whether AI-enabled workflows remain grounded in operational reality before their outputs become consequential.

> *HAOP Diagnostic Tools are under development to help organizations examine AI-enabled and consequential workflows BEFORE automation accelerates the wrong assumptions. They will not be certification tools or technical model audits. They will help surface the governance questions leaders and teams need to address. The HAOP framework described in this paper and the True Function Diagnostic are free to use either way.*

References

1. Frederick Winslow Taylor, The Principles of Scientific Management (New York: Harper & Brothers, 1911).
2. Erik Hollnagel, Safety-I and Safety-II: The Past and Future of Safety Management (Farnham, UK: Ashgate, 2014).
3. Todd Conklin, The 5 Principles of Human Performance: A Contemporary Update of the Building Blocks of Human Performance for the New View of Safety (Santa Fe, NM: PreAccident Investigation Media, 2019).
4. Sidney Dekker, The Field Guide to Understanding “Human Error”, 3rd ed. (Boca Raton, FL: CRC Press, 2014).
5. Sidney Dekker, Safety Differently: Human Factors for a New Era, 2nd ed. (Boca Raton, FL: CRC Press, 2014).
6. Erik Hollnagel, Robert L. Wears, and Jeffrey Braithwaite, From Safety-I to Safety-II: A White Paper (University of Southern Denmark, University of Florida, and Macquarie University, 2015).
7. OWASP GenAI Security Project, OWASP Top 10 for Agentic Applications - 2026 Edition (December 2025). OWASP describes the framework as addressing risks in autonomous and agentic AI systems that plan, act, and make decisions across workflows.
8. Richard M. Cyert and James G. March, A Behavioral Theory of the Firm, 2nd ed. (Malden, MA: Blackwell, 1992; originally published 1963).
9. Diane Vaughan, The Challenger Launch Decision: Risky Technology, Culture, and Deviance at NASA (Chicago: University of Chicago Press, 1996).
10. Sidney Dekker, Drift into Failure: From Hunting Broken Components to Understanding Complex Systems (Farnham, UK: Ashgate, 2011).
11. James G. March and Herbert A. Simon, Organizations, 2nd ed. (Cambridge, MA: Blackwell, 1993; originally published 1958).
12. Daniel Kahneman, Attention and Effort (Englewood Cliffs, NJ: Prentice-Hall, 1973).
13. Raja Parasuraman and Victor Riley, “Humans and Automation: Use, Misuse, Disuse, Abuse,” Human Factors 39, no. 2 (1997): 230–253.
14. Stuart Russell, Human Compatible: Artificial Intelligence and the Problem of Control (New York: Viking, 2019).
15. Ilia Shumailov et al., “AI Models Collapse When Trained on Recursively Generated Data,” Nature 631 (2024): 755–759. The article defines model collapse as a degenerative process in which generated data pollutes later training sets.
16. Michael Polanyi, The Tacit Dimension (Chicago: University of Chicago Press, 1966).

17. Gary Klein, Sources of Power: How People Make Decisions (Cambridge, MA: MIT Press, 1998).

18. Financial Times, “Amazon Scraps AI Leaderboard to Stop Workers Chasing Usage Scores,” May 2026.

19. Hugh Langley, “Amazon Says It Shut Down a Token Leaderboard: ‘Don’t Use AI Just to Use AI,’” Business Insider, May 2026.

20. International Organization for Standardization and International Electrotechnical Commission, ISO/IEC 42001:2023: Information Technology - Artificial Intelligence - Management System (Geneva: ISO, 2023). ISO describes ISO/IEC 42001 as an AI management system standard for managing AI risks and opportunities.

21. National Institute of Standards and Technology, Artificial Intelligence Risk Management Framework (AI RMF 1.0), NIST AI 100-1 (Washington, DC: U.S. Department of Commerce, January 2023). NIST frames AI risk management around governance, mapping, measuring, and managing AI risks.

22. Russell, Human Compatible; NIST, Artificial Intelligence Risk Management Framework (AI RMF 1.0).

23. EUROCONTROL, “Just Culture,” accessed May 31, 2026. EUROCONTROL defines Just Culture as protecting operators from punishment for actions commensurate with experience and training, while not tolerating gross negligence, willful violations, or destructive acts.

24. Gibson v Maritime New Zealand [2026] NZHC 813. The High Court dismissed the appeal and upheld the former Ports of Auckland CEO’s conviction, NZ$130,000 fine, and NZ$60,000 costs award under New Zealand’s Health and Safety at Work Act.

25. Maritime New Zealand, “High Court Dismisses Tony Gibson Appeal,” April 13, 2026. Maritime NZ stated that the Auckland High Court upheld the guilty finding and sentence after stevedore Pala’amo Kalati was killed by a falling container at the port.

26. New Zealand, Health and Safety at Work Act 2015, ss. 44 and 152.