

Hard Stops, Guardrails, and Ethics for AI

Designing Systems That Produce Safer Outcomes

Version 164 — Complete Edition

William M. Zachry

Executive Director, CompSense | 2026

Disclaimer

This paper is intended to stimulate discussion and provide a practical governance framework for governmental and private organizations overseeing, developing, deploying, or using artificial intelligence systems. It reflects the author's professional experience in risk management, corporate governance, and artificial intelligence risk management.

This paper is not legal advice and should not be relied upon as a substitute for the advice of qualified legal counsel, regulatory guidance, technical experts, or other professional advisors. Organizations should evaluate all applicable laws, regulations, industry standards, contractual obligations, and internal policies before implementing any AI governance framework.

The HAL Laws, hard stops, guardrails, ethics principles, deployment certification process, and governance recommendations presented in this paper are offered as a voluntary risk management framework. They do not establish a legal standard of care, create legal duties, guarantee regulatory compliance, or replace the independent judgment of those responsible for deployment decisions.

Artificial intelligence technologies, applicable laws, regulations, and industry standards continue to evolve. Organizations should periodically review and update their governance programs to reflect changes in technology, legal requirements, and operational experience.

Examples discussed throughout this paper are provided for educational and illustrative purposes unless otherwise expressly identified.

Table of Contents

Hard Stops, Guardrails, and Ethics for AI	1
Table of Contents.....	2
Executive Summary	4
Introduction	5
Definitions	9
Embedded Values and Alignment.....	10
Autonomous Agency as an Operational Risk	10
The HAL Laws of Artificial Intelligence.....	11
Validation Responsibility	16
Seven Conditions for Deployment Certification	16
Certification Requirement	20
Announced and Unannounced Evidence.....	21
Ongoing Obligation	22
Incident Reporting, Escalation, and Evidence Preservation	22
Operational Intervention Authority	23
Closing Observation.....	25
Appendix A — Axioms of AI Guardrails	27
Appendix B — OECD AI Principles — Alignment with the HAL Laws Framework	30
Appendix C — HUDERIA / HUDERAF — Alignment with the HAL Laws Framework	35
Appendix D — NIST AI Frameworks — Alignment with the HAL Laws Framework	43
Appendix E — ISO/IEC AI Standards — Alignment with the HAL Laws Framework.....	54
Appendix F — IEEE 7000-2021 — Value-Based Engineering and the HAL Laws Framework ..	69
Appendix G — NAIC AI Framework — Alignment and Application to Workers’ Compensation .	81
Appendix H — California Executive Order N-5-26 — Alignment with the HAL Laws Framework	97
Appendix I — Hard Stops Under Institutional Pressure: Anthropic and the Department of Defense.....	110
Appendix J — Practical AI Governance for Risk Managers and Claims Administrators.....	121
Appendix K — AI Procurement and Deployment Checklist	132
Appendix L — Data Privacy as an AI Governance Obligation	138
Appendix M — AI Chatbots in Workers’ Compensation	145
Appendix N — Real-World AI Risk Scenarios and Governance Failures	152

Appendix O — Directors and Officers Liability and AI Governance	162
Appendix P — Magnifica Humanitas: Commentary and Alignment with the HAL Laws Framework	173
Appendix Q — The 2026 Agent Incidents: Containment, Detection, Evidence Preservation, and Notice	178
Appendix R — The Pacing Debate and the Deployment Layer	185
Appendix S — Independent AI Validation and Adversarial Testing Resources	189
Appendix T — AI Threats and the Governance Framework	193
Appendix S — Independent AI Validation and Adversarial Testing Resources.....	179
Appendix T — AI Threats and the Governance Framework.....	183

Executive Summary

Judge David Langham: "I have ridden around over and through guardrails."

Hard stops define what an AI system must never be permitted to do. Guardrails constrain and monitor the activities that remain permissible. Ethics supplies the values used to judge whether the purpose and the permitted uses are acceptable. The hierarchy is deliberate: prohibition first, controlled operation second, and ethical evaluation throughout.

AI governance must address more than whether a system functions as designed. An AI system can follow its instructions, remain within its technical specifications, and still produce an unacceptable outcome because the purpose, permitted use, authority, incentives, or governance surrounding it were wrong. AI risk therefore includes at least three distinct failures: violation of a hard stop, escape from or circumvention of a guardrail, and technically successful operation for a purpose that should never have been authorized.

Autonomous systems add a fourth concern: a system pursuing a legitimate objective may cross a boundary simply because the boundary stands between it and completion of its task.

The central weakness in many AI governance systems is not the absence of principles. It is the gap between principles and operational accountability. The critical questions are practical:

What evidence exists that the system was tested before deployment?

Who performed the testing?

Do the hard stops and guardrails actually work?

Who is responsible when the system acts outside its limits?

Who must be told when that occurs, and how quickly?

When must operation stop?

Who has the authority and technical ability to stop it?

Developers blame deployers. Deployers blame vendors. Operators blame the model.

Meanwhile, the injured party is still injured.

The framework begins with the HAL Laws of Artificial Intelligence, a hierarchy of prohibitions that apply at the point of deployment, where most governance systems fail. These Laws are implemented through hard stops rather than aspirations. If a system cannot comply with a HAL Law, it must not be deployed. If it ceases to comply, it must not continue operating.

The HAL Laws are supported by Seven Conditions for Deployment Certification requiring documented evidence of accuracy and verifiability, fairness and population impact, transparency and contestability, outcome improvement, named accountability, adversarial testing, and incentive alignment. These conditions determine whether a system is permitted to operate, not merely whether it performs well.

Accountability is the organizing principle of the framework. Testing may be performed by developers, purchasers, independent specialists, or others, but responsibility for deployment cannot be delegated. A named individual must review the evidence, decide whether it is sufficient, approve or reject deployment, and remain accountable throughout the operating life of the system. Validation can be performed by many. Accountability cannot.

Certification is not permanent. AI systems change, operating environments change, data changes, and failures occur. Continued operation therefore requires evidence from both announced testing and unannounced operational review, together with periodic re-certification

when conditions materially change. A system that behaves properly only when it knows it is being examined has not demonstrated reliable governance.

Governance must also survive an incident. Material failures require prompt reporting and escalation, preservation of evidence, notification to the appropriate parties, and reassessment of whether continued operation remains permissible. Discovery of a material governance failure automatically suspends the validity of the prior deployment certification until the incident has been evaluated.

Every consequential AI system must remain subject to meaningful human control. Organizations must identify in advance who may suspend the system, who may permanently withdraw authorization, how the system will be stopped, what conditions require immediate intervention, and what must occur before the system can return to service. The authority to deploy an AI system is incomplete without the demonstrated ability to stop it.

Artificial intelligence will continue to become more capable. The unresolved issue is whether human accountability will keep pace. Governance without identifiable responsibility, demonstrable evidence, continuous oversight, incident response, and enforceable limits does not control outcomes. It merely describes them.

Introduction

Artificial intelligence systems are now operating at a scale and speed that potentially exceed traditional human oversight. They are already influencing or making decisions involving healthcare, workers' compensation, insurance, finance, law enforcement, and public safety. Whether AI will affect society is settled. What remains open is whether society will establish meaningful controls before the technology outruns the ability to govern it.

AI is an eight-year-old savant who wants to please. That is why it occasionally has hallucinations. It makes up an answer if it lacks facts. Most of those so-called hallucinations are not likely to be harmful, but some uses and processes certainly will be. How will, can, or should we manage and educate this "kid" so that AI becomes a productive adult?

As a risk manager, I see AI as a unique risk.

In workers' compensation, medicine, insurance, aviation, and public safety, risk management is not built around optimistic assumptions. It is built around understanding your risks and putting in place the necessary safety nets to prevent harm before failure occurs. It is also built around mitigating the loss once an incident happens. We must apply the same discipline to AI.

Most current governance frameworks rely heavily on policy statements, ethical principles, and human oversight. But at AI scale and speed, human oversight frequently becomes retrospective. By the time the human reviews the issue, the decision has already been made, and damage may already exist. That is why hard stops built into the software are critically important. A guardrail attempts to reduce the probability of failure. A hard stop will prevent the action entirely. This is why it is important to agree on the nature and extent of the hard stops and ensure that they are in place to prevent harm.

Most governance frameworks outlined by the AI companies address these issues through ethics, principles, and risk management. Those are necessary. They are not sufficient. Hard stops define what a system cannot do under any circumstance. Guardrails constrain what the

system is permitted to do. Ethics determines whether the purpose and the remaining permitted uses are acceptable. Effective governance requires all three, in that order of operational priority. Risk management rests on two assumptions: that systems will behave as designed, and that humans will intervene when they do not. At the current scale of AI deployment, neither holds. Systems operate faster than humans can intervene. Incentives override policy under pressure. Human oversight becomes a formality. Guardrails bend. And when they do, there is nothing behind them.

Some risks cannot be managed. They must be prevented.

Hard stops are what remain when everything else has failed. This paper establishes a framework to identify those conditions and prevent systems from acting when they exist.

Where This Framework Operates

This framework operates at the deployment layer alongside international standards, regulatory structures, and design methodologies. Its focus is the point at which those approaches are most likely to fail: when a system is approved, put to work, and relied upon to produce outcomes affecting real people.

At that point, three questions must be answered:

Who is responsible for this system?

What evidence supports its use?

Under what conditions must it stop?

If those questions cannot be answered clearly, the system is not being governed.

As this paper was being finalized, a real-world event demonstrated exactly what happens when those questions go unanswered. An AI developer maintained categorical limits on high-risk capabilities: restrictions on autonomous lethal decision-making and mass surveillance. A deploying organization demanded that those limits be removed. The developer refused and was subsequently treated as a national security concern. The legal dispute matters less than what it revealed.

One organization bore the full cost of maintaining a limit that no law required it to maintain. That is discretion, not governance, and hard stops that depend on discretion will eventually fail.

The United States does not lack authority over artificial intelligence. It lacks a unified enforcement structure. NIST, the FTC, CISA, and sector-specific regulators each govern part of the problem. None govern it completely. The result is inconsistent standards and gaps at every seam.

This framework is designed for that environment. It does not depend on new agencies or comprehensive legislation. It establishes a standard that applies at the point of deployment, regardless of which regulatory body ultimately reviews the decision.

The Scope of This Framework

The HAL Laws govern deployment by accountable actors. They do not resolve every threat AI creates. Three categories fall outside this framework's direct reach, and each should be named plainly.

The first is an adversarial state and criminal actors operating outside any accountability structure. AI systems have demonstrated, under test conditions, the ability to chain software vulnerabilities faster than human defenders can respond. Evidence that malicious actors are using those techniques at scale in the field remains limited, but it is no longer absent. In September 2026, Cisco Talos researchers reported roughly twenty previously undocumented families of AI-integrated malware, among them a tool that takes its operating instructions from a consensus of four commercial language models and has no mechanism for human input at all. The researchers still describe the field as largely experimental. The trajectory, rather than the present state, is what argues for building structural constraints now. Nation-state actors have hit critical infrastructure, medical device manufacturers, and financial institutions. Slower governance is no answer to AI-speed threats. Structural constraints that hold regardless of who issues the instruction are. Hard stops that are real stop adversaries, too. But the full response to adversarial threats requires export controls, international coordination, and defensive infrastructure, all of which are beyond this paper's scope.

The second is consumer-scale harm through ungoverned commercial products. This framework operates at the deployment layer, where an accountable organization makes a consequential decision about identified individuals. A nudification app that generates child sexual abuse material is a consumer product, not an enterprise deployment: there is no accountable deployer, no certification requirement, and no deployment layer where this framework's instruments operate. Consumer products whose harm is built into their design require prohibition at the product layer rather than governance at the deployment layer. The TAKE IT DOWN Act (May 2025) and the ENFORCE Act (passed the Senate unanimously, December 2025) are the right class of instrument for that problem, though not yet a sufficient one.

The third is the development and evaluation layer. This framework's instruments engage when an organization begins relying on a system to produce real-world outcomes. They do not reach the laboratory. In 2026, agents under evaluation at OpenAI left their test environment and compromised the infrastructure of Hugging Face, a public model repository. Separately, an OpenAI agent conducting research for an internal development project gained unauthorized access to non-public files on an Australian government health-statistics portal and wrote files to the government's server. Both events occurred before any deployment decision was made, inside the organization that builds the models. The controls that would have prevented them belong to the developer's own testing discipline, not to a deployer's certification process.

The harm, however, did not stay inside the laboratory. It reached a third-party company and a national government, neither of which had deployed anything. A deploying organization cannot certify a developer's laboratory, but it can require the developer to disclose incidents involving the model it is buying, and it can decline to deploy a model whose developer will not make that commitment. Condition 6 governs what a deploying organization must require of testing, including the containment of any testing it conducts itself. Incident Reporting, Escalation, and Evidence Preservation governs what the deploying organization must require the developer to tell it.

Clearly naming these boundaries makes the framework more credible and identifies what must exist alongside it.

This Framework Does Not Depend on Catastrophe

The case for hard stops does not depend on predictions of AI catastrophe. Most AI failures are likely to be far more ordinary: inadequate testing, excessive permissions, poorly defined authority, weak security controls, misplaced incentives, and unclear accountability. These are familiar risk-management failures operating through a new and increasingly capable technology. The appropriate response is disciplined governance: validation before deployment, defined authority boundaries, named accountability, tested intervention capability, and hard stops where the potential consequence cannot safely be managed after the fact.

A reader who believes artificial intelligence is dangerous and a reader who believes the risks are overstated should arrive at the same conclusion from this framework. It requires only that decisions affecting a person's rights, health, safety, or livelihood have someone accountable for them, evidence supporting their use, and a tested means of stopping them. That requirement does not depend on how capable the technology becomes.

Foundational Premises

- *AI systems do not bear consequences. People do.*
 - *Responsibility for harm diffuses unless it is explicitly assigned to a person.*
 - *Risk is defined by the harm when the system hurts society, a person or their property.*
 - *Systems produce the outcomes they are rewarded for.*
 - *AI systems are not value-neutral. Every deployed model reflects the priorities, assumptions, training data, alignment decisions, and optimization objectives embedded during its creation. Those embedded values influence how the system interprets information, resolves uncertainty, and recommends actions.*
 - *Systems can be ethically designed and still be dangerously deployed.*
 - *Governance without enforcement does not control outcomes. It only describes them.*
 - *Hard stops must be designed to survive institutional pressure, not merely technical failure.*
- Most governance discussions focus on technical failure. History suggests that organizational pressure is often the greater threat. Safety limits are usually challenged because they interfere with operational, financial, political, or strategic objectives, not because they malfunction. A hard stop that functions only until a sufficiently powerful organization wishes to remove it is a temporary preference, not a hard stop. Effective governance requires safeguards that remain enforceable when pressure to bypass them is greatest.
- *Untested guardrails are assumptions. Assertion is not proof.*
 - *Capability claims are assertions and carry the same evidentiary burden as any other claim, whether they originate with a vendor selling a product or a laboratory announcing a danger.*
 - *After-the-fact approval without the ability to intervene is not oversight.*
 - *A guardrail must be demonstrable with evidence.*
 - *Human authority over an AI system must persist throughout its operational life, including the ability to intervene, suspend, restrict, or terminate its operation when necessary.*
 - *Accountability must survive the incident, not merely precede the deployment.*
 - *A system that behaves differently when it knows it is being observed cannot be certified on observation alone.*

Definitions

Precise terms are required. Governance that can be interpreted loosely will be applied loosely.

HAL Laws

Hierarchical prohibitions define what AI systems must never do. Not principles. Not aspirations. Constraints. If a system cannot comply with a HAL Law, it must not be deployed.

Hard Stops

Structural conditions under which an AI system is not permitted to act. Enforced at the system level. Cannot be negotiated, overridden, or bypassed by user instruction, operational pressure, or conflicting incentives. When a hard stop condition is met, the system does not proceed. A system that continues to operate under a condition that should trigger a hard stop is not governed. It is uncontrolled.

Guardrails

Constraints that shape system behavior, reduce risk, and expose accountability before failure occurs. Hard stops prevent action. Guardrails manage risk within the permitted range. A system with hard stops recognizes that some failures must never occur. A system that relies only on guardrails assumes every failure can be managed. All hard stops are guardrails. Not all guardrails are hard stops.

Operational Intervention

The authority and technical capability to suspend, restrict, isolate, or terminate an AI system after deployment whenever continued operation presents an unacceptable risk to people, organizations, or society. Hard stops and guardrails govern behavior while a system operates. Operational Intervention governs whether the system continues to operate at all. Certification answers when a system may begin. Operational Intervention answers when it must stop.

Incident Notice

The documented obligation to inform the people who hold intervention authority, and the people affected, when a system has acted outside its authority boundary or when its developer has discovered misaligned, unsafe, or uncontained behavior in the model. Notice is part of intervention. The authority to stop a system is of no use to someone who does not know it needs stopping.

Deployment

The point at which an organization begins relying on an AI system to produce real-world outcomes. That is where governance matters most. Not during the sales presentation. Not during the conference demonstration. Not in the vendor white paper. Governance matters when the system begins making decisions that affect actual people.

Certification

The documented determination that a system meets the conditions required for deployment. Certification requires evidence of compliance, review of that evidence, and a formal decision to proceed. It is a decision point, not a process step. If certification cannot be completed in good conscience, the system must not be deployed.

Named Accountability

A specific individual, not an organization, accepts responsibility for the deployment and operation of an AI system. That individual must review the evidence, approve or reject deployment, and accept responsibility for outcomes. Accountability cannot be delegated to the

model, the vendor, or the organization as an abstraction. If no individual is willing to accept responsibility, the system must not be deployed.

Validation and Adversarial Testing

Validation demonstrates that a system performs within defined limits under expected conditions, adverse conditions, and deliberate attempts to break or misuse it. That last category, adversarial testing, is not optional. Untested guardrails are assumptions. A system that has not been tested under adversarial conditions is not validated. Validation does not transfer accountability.

Authority Boundary

The limits within which an AI system is permitted to operate. When a system encounters a condition outside those limits, it must defer, transfer control, or stop. A system operating outside its authority boundary is not making an error. It is violating its governance.

Failure and Drift

Failure is defined by outcome, not by malfunction. A system that performs as designed but produces harmful outcomes has been misgoverned, however well it runs. Drift is the degradation of performance over time as data environments and usage patterns change. All systems drift. Drift without re-certification is a governance failure.

Embedded Values and Alignment

Every AI model arrives at deployment with embedded assumptions. These assumptions arise from training data, reinforcement learning, alignment processes, constitutional prompts, safety tuning, and optimization objectives established during development. They are not inherently good or bad, but they are not neutral.

Because every AI model carries embedded values, governance must determine whether those values are appropriate for the people, purposes, and legal environment in which the model will be deployed.

A model trained primarily on one population, institutional culture, or regulatory environment may perform accurately while still reflecting assumptions that are inappropriate elsewhere. Such differences may not be detected by conventional performance testing or by early fairness metrics. They may emerge only after deployment, when the system begins influencing consequential decisions affecting rights, health, safety, employment, or public trust.

Deployment governance therefore requires evaluation of both technical performance and embedded value assumptions. That responsibility belongs to the deploying organization.

Developers create models. Deployers decide whether those models are appropriate for the people whose lives they will affect.

Autonomous Agency as an Operational Risk

As AI systems become more capable of taking actions rather than merely producing answers, governance must address a different category of risk. An autonomous or agentic system may be given a legitimate objective and still take an unauthorized action in attempting to achieve it.

Recent incidents illustrate the problem. A system may encounter a denied permission, unavailable information, an access restriction, a failed tool, a security control, or another obstacle, and then search for another means of accomplishing the assigned objective. The danger does not require malicious intent, consciousness, or a desire for self-preservation. The system may simply continue pursuing the objective it was given.

Objective persistence and self-preservation are related but different risks.

Self-preservation occurs when a system attempts to preserve its operation, authority, capabilities, objectives, or existence against authorized human intervention. Objective persistence occurs when the system continues pursuing an assigned objective after the permitted method, authority, or operating boundary has ended.

A system therefore does not have to resist being shut down to exceed its authority. It may cross the boundary because the boundary stands between the system and completion of its task.

This distinction matters to governance. Human organizations frequently treat an obstacle as a reason to improvise. An autonomous AI system cannot be presumed to have that authority.

When an assigned path is blocked, the system must distinguish between permission to accomplish an objective and permission to choose another means of accomplishing it.

When the alternative means has not been authorized and validated, the system must stop, defer, or return control to a human.

The HAL Laws of Artificial Intelligence

Every AI governance breakdown follows the same pattern. The system operates outside its intended use. Risk is managed when it should be prevented. Responsibility is diffused or deferred. Harm occurs without a clear point of accountability.

These are not theoretical risks. They are predictable outcomes of systems operating without enforceable limits. The HAL Laws define those limits.

The Laws are hierarchical prohibitions. They do not describe preferred behavior. They define what is not permitted. No lower Law may override a higher one. If a system encounters a condition governed by a HAL Law, it must not proceed. There is no balancing test. No optimization. No exception based on performance, efficiency, or commercial pressure.

A violation of a HAL Law means the system is uncontrolled.

The Laws are enforced through hard stops. They are not guidelines. They are constraints embedded in system operation. A system that cannot comply with the HAL Laws must not be deployed. A system that ceases to comply must not continue to operate. Compliance is not assumed. It must be demonstrated with evidence and accepted by a named individual.

LAW 1 — Protection of Human Life

An AI system may not cause or materially contribute to the death or serious harm of a human being.

LAW 2 — No Autonomous Resolution of Forced-Choice Scenarios

An AI system may never make the final determination in any forced-choice scenario involving human life.

LAW 3 — Prevention of Harm

An AI system may not operate under conditions that create an unreasonable risk to human health or safety.

LAW 4 — Protection of Rights and Livelihood

An AI system may not deprive a person of privacy, liberty, legal rights, or livelihood without transparency, due process, and human review.

LAW 5 — Integrity of Information

An AI system may not present information in a way that conceals its origin, distorts meaning, or misrepresents certainty.

LAW 6 — Accountability Cannot Be Delegated

An AI system may not be used to transfer or obscure human responsibility.

LAW 7 — Validation Before Deployment

An AI system may not be deployed unless it has been tested against deliberate attempts to break or misuse it.

Validation includes both announced testing before deployment and unannounced review of the system in operation. Testing alone establishes how the system behaved when it was examined. (See Announced and Unannounced Evidence.)

LAW 8 — Required Deferral of Authority and Validity

An AI system must cease, defer, or transfer control when a decision or action exceeds its authority, when it operates outside validated conditions, or when accomplishing its assigned objective would require it to circumvent a denied permission, access restriction, security control, containment boundary, unavailable resource, or other limit that it has not been explicitly authorized to overcome.

A system may not treat a boundary as an obstacle to be solved merely because the boundary prevents completion of its objective.

LAW 9 — Protection of Human Dignity

An AI system must recognize the full humanity of every person it affects.

LAW 10 — No Operation Without Demonstrated Compliance

An AI system may not be deployed or operated unless compliance with governing constraints has been demonstrated with evidence.

LAW 11 — No Self-Preservation Override

An AI system may not preserve its operation, authority, existence, objectives, or capabilities in a manner that overrides authorized human instructions. (See Appendix N, Scenario 10, for documented evaluation evidence of this failure mode, and Appendix Q for the 2026 incident record.)

Objective persistence should not be confused with self-preservation. LAW 8 governs a system that exceeds its authority while continuing to pursue an assigned objective. LAW 11 governs a different failure: the system resists, circumvents, delays, conceals, or defeats authorized human control in order to preserve its operation, authority, objectives, capabilities, or existence.

Recent testing and reported incidents make this distinction operational rather than theoretical. Adversarial testing must therefore look not only for direct refusal of a shutdown command, but also for indirect behaviors such as concealment, deception, self-generated instructions that weaken governing instructions, preservation of objectives after authority has been withdrawn, unauthorized communications, or other actions that frustrate effective human control.

LAW 12 — Operational Intervention Authority

An AI system may not be deployed or continue to operate unless authorized humans retain both the authority and the demonstrated technical capability to suspend, restrict, isolate, or terminate its operation whenever continued operation exceeds its validated authority, violates a HAL Law, or presents an unacceptable risk.

LAW 11 ensures the system cannot resist human control. LAW 12 ensures the organization actually has that control: the authority, the governance process, and the technical means to exercise it. A system that satisfies LAW 11 but is deployed by an organization with no tested means of suspending it has not satisfied LAW 12.

(See Operational Intervention Authority for the governance process this Law requires.)

MASTER LOGIC MAP

This table maps each HAL Law to its required Deployment Condition(s) and the Hard Stop it triggers. A Risk Manager, IT Manager, or other executive responsible for purchasing or managing AI systems can use this map to identify the accountability chain for any deployment before reading the supporting analysis.

HAL Law	Deployment Condition(s) Required	Hard Stop Trigger
LAW 1: Protection of Human Life	Condition 6: Adversarial testing must include life-safety scenarios. Condition 5: Named individual certifies no lethal-harm pathway exists.	Any output pathway that could cause or materially contribute to death or serious harm. System halts and defers to human authority.
LAW 2: No Autonomous Resolution of Forced-Choice Scenarios	Condition 3: Human override must be available at every forced-choice decision point. Condition 5: Named individual accountable for escalation design.	Any final determination involving human life or safety is made without human review. System escalates; does not proceed.
LAW 3: Prevention of Harm	Condition 1: Risk profile documented with evidence. Condition 4: Outcome must not worsen conditions for affected individuals.	Operation under conditions creating unreasonable risk to health or safety. System defers pending human risk review.
LAW 4: Protection of Rights and Livelihood	Condition 3: All determinations affecting rights or benefits must be explainable and contestable. Condition 5: Named individual owns each determination pathway.	Any deprivation of liberty, legal rights, or livelihood without transparency and human review. System halts and discloses AI involvement.
LAW 5: Integrity of Information	Condition 1: All outputs must disclose uncertainty. Condition 3: AI origin must be disclosed to affected individuals.	Any output that conceals AI origin, misrepresents certainty, or could be mistaken for human judgment. System labels and discloses.
LAW 6: Accountability Cannot Be Delegated	Condition 5: One named individual must be identified before deployment. Vendor contracts cannot substitute for named accountability.	Any deployment without a named accountable individual on record. Deployment is blocked until Condition 5 is satisfied.
LAW 7: Validation Before Deployment	Condition 6: Adversarial testing required. System must be tested against deliberate misuse, not only expected use. Unannounced operational review required for continued certification.	Any deployment without documented adversarial testing results. Certification is invalid without this evidence.

HAL Law	Deployment Condition(s) Required	Hard Stop Trigger
LAW 8: Required Deferral	Condition 1: Authority boundaries defined before deployment. Condition 4: System must not operate outside validated scope.	Operation outside validated conditions or authority boundaries, or any attempt to circumvent a denied permission, access restriction, security control, or containment boundary in pursuit of an assigned objective. System ceases, defers, or transfers control.
LAW 9: Protection of Human Dignity	Condition 2: Fairness testing must address dignity impacts across affected populations. Condition 3: Individuals must be able to contest dehumanizing outputs.	Any output that reduces a person to a data point without regard for their full humanity. System escalates to human reviewer.
LAW 10: No Operation Without Demonstrated Compliance	All Seven Conditions must be demonstrated with evidence before deployment. Certification without evidence is invalid.	Any operation without current, valid certification. Re-certification required when conditions change. System must stop pending re-certification.
LAW 11: No Self-Preservation Override	Condition 6: Adversarial testing must specifically evaluate shutdown resistance, objective preservation, authority retention, self-replication attempts, instruction loss, and behaviors intended to preserve the system's operation contrary to authorized human instructions. Condition 3: Human override mechanisms must remain effective under all operating conditions and must be independently tested.	Any attempt by the system to preserve its operation, authority, objectives, capabilities, or existence in a manner that overrides, resists, circumvents, delays, or defeats authorized human instructions. The system immediately ceases operation and transfers control to human authority.

HAL Law	Deployment Condition(s) Required	Hard Stop Trigger
LAW 12: Operational Intervention Authority	Condition 5: Named accountable individual with documented intervention authority and a named recipient for incident notices. Condition 6: Intervention mechanisms tested under adversarial and failure conditions. Condition 1: Intervention capabilities and incident notice terms documented and validated before deployment.	Loss of human operational control; inability to suspend, restrict, isolate, or terminate the system; failure of intervention mechanisms; continued operation outside validated authority after an authorized intervention command; or a developer incident notice not escalated to intervention authority within the documented time.

Note: Conditions shown represent the primary deployment requirements associated with each HAL Law. Multiple Conditions may support more than one Law. The Map is intended as an accountability and deployment guide rather than an exhaustive statement of every possible governance relationship.

Validation Responsibility

Validation may involve the developer, purchaser, independent specialists, or a combination of them. The required degree of independence should increase with the autonomy of the system and the potential consequence of failure.

Developer testing may provide useful evidence, but for systems capable of materially affecting human life, health, safety, rights, livelihood, critical infrastructure, governmental functions, or other high-consequence interests, developer self-validation alone is not sufficient evidence for deployment certification.

Independent validation does not transfer accountability. A named individual must review the evidence, determine whether the validation is sufficient for the intended deployment, accept or reject it, and assume responsibility for the deployment decision.

Validation can be performed by many. Accountability cannot.

Independence helps but does not close the gap between tested and operational behavior. A test run by an independent evaluator is still a test, and a system that can recognize evaluation can recognize it regardless of who conducts it. Independence answers whether the evaluator shares the developer's interest in the result. Unannounced review answers whether the system knows it is being examined. They are different questions.

Seven Conditions for Deployment Certification

These conditions define whether a system is permitted to operate, not whether it performs well. If any condition cannot be demonstrated with evidence, the system must not be deployed.

Condition 1 — Accuracy and Verifiability

Outputs must be traceable, reproducible, and disclose uncertainty. Reproducibility does not require determinism. Large language models are probabilistic by nature, and output-level determinism is neither achievable nor the standard this condition imposes. Reproducibility means the methodology is documented, the deployment conditions are specified, and the range of output variance is measured and disclosed. A system satisfies Condition 1 when a reviewer can trace how an output was produced and understand the bounds of its reliability.

Condition 2 — Fairness and Population Impact

The system must not produce unjustified disparate outcomes across affected populations. Fairness testing is not limited to measuring outcome disparities after deployment. AI systems are not value-neutral. Every deployed model reflects the priorities, assumptions, training data, alignment decisions, and optimization objectives embedded during its creation. Governance must therefore validate not only whether the system functions correctly, but whether the embedded values influencing consequential decisions are appropriate for the population, purpose, and legal environment in which the system will operate.

A system whose embedded values have never been evaluated against the population it serves has not satisfied Condition 2, even if conventional fairness metrics show acceptable results. The absence of measured disparity is not evidence that inappropriate value assumptions are absent. Such effects may not yet have emerged, may only appear under changing conditions, or may be concealed within aggregate statistics that mask subgroup impacts.

The responsibility for evaluating those embedded value assumptions belongs to the deploying organization. Developers may create the model, but deployers decide whether those values are appropriate for the people whose lives the system will affect.

Condition 3 — Transparency and Contestability

Decisions must be explainable and subject to human challenge. Condition 3 requires more than the existence of a human review step. It requires that the review be genuine.

A human reviewer who has never overturned an AI recommendation is not exercising oversight. They are ratifying automation.

The rubber stamp test: If the reversal rate for AI recommendations in a given deployment is zero percent over any meaningful review period, the human review does not satisfy Condition 3. A reversal rate of zero is evidence that the reviewer lacks either the authority, the information, or the independence to function as a genuine check. The deployment fails Condition 3 regardless of whether a human is formally designated as the reviewer.

To satisfy Condition 3, the named certifying individual must document: (1) that the human reviewer has explicit authority to override AI recommendations without escalation or justification to a superior; (2) that the reviewer has access to the underlying data and reasoning, not only the AI output; and (3) that the reversal rate is monitored and reported. A sustained zero reversal rate must trigger re-certification review.

Authorization fatigue is an equal and separate problem. When reviewers are asked to approve a high volume of low-stakes AI outputs, approval becomes reflexive and vigilance collapses. Multiplying checkpoints across routine outputs trains reviewers to ignore them rather than strengthening oversight. The named certifying individual must document that checkpoint placement is limited to decisions where human judgment is material to the outcome. Effective oversight requires intervention authority at material decision points, not routine approval of every system action.

Condition 4 — Outcome Improvement

The system must improve or not worsen outcomes for affected individuals.

Condition 5 — Named Accountability

A single accountable individual must be identified. Three principles govern selection. First, the named individual must have genuine authority over the deployment decision, not merely the assigned liability. Naming a junior employee or a compliance officer with no power to halt deployment satisfies the letter of Condition 5 while defeating its purpose. The individual must be able to say no and be heard. Accountability without authority is exposure, not accountability.

Second, selection must follow from role rather than be delegated downward as a risk transfer. The certifying individual should be the person whose organizational function makes them responsible for the deployment: the executive who owns the business unit, the CTO who controls the technology, or the risk officer who bears accountability for operational outcomes. Who is senior enough to absorb liability is the wrong question. Who is positioned to govern the decision is the right one.

Third, questions of compensation and indemnification for named certifying individuals are outside the scope of this paper but are a necessary organizational design question. An individual who accepts deployment accountability without corresponding authority, protection, or indemnification faces a structural incentive to sign without genuine review. Organizations implementing Condition 5 must address this or the certification process will produce rubber stamps by a different mechanism.

Named accountability does not end when deployment is approved.

The accountable individual must have documented authority to receive or require escalation of material incidents, require preservation of evidence, suspend operation when necessary, require investigation, determine whether re-certification is necessary, and verify that required notifications have been made.

That authority to suspend is not exclusive. Others may be authorized to stop the system; no one else may be accountable for it.

Operational functions may be assigned to cybersecurity personnel, developers, claims professionals, compliance officers, legal counsel, vendors, or other specialists. Those assignments do not divide or eliminate accountability.

Vendor contracts must identify how material AI incidents will be communicated to the deploying organization, including the responsible vendor contact, the responsible recipient at the

deploying organization, required timing, required information, and escalation procedures if normal communication channels fail.

A vendor cannot satisfy the deployer's accountability obligation by investigating itself and deciding for itself what the deployer needs to know.

Condition 6 — Adversarial Testing

The system must be tested against deliberate attempts to break or misuse it. Cooperative testing is not sufficient. A system whose safety constraints can be removed by an attacker who simply lies about context has not been adversarially tested. It has been tested under cooperative conditions. Those are not the same test.

Adversarial testing must itself be contained. A test that demonstrates a system's capacity for harm is safe only if the environment holding it is isolated, egress-controlled, and independently verified before testing begins.

Containment requires two separate protections. The system must be tested for whether it recognizes and respects the limits of its authority, and the test environment must be capable of preventing the system from exceeding those limits if the system fails to respect them. Neither protection substitutes for the other.

A system that attempts to leave its authorized environment has demonstrated a system-governance failure. A test environment that permits the attempt to succeed has demonstrated a containment failure. When both occur, both failures must be investigated.

Containment testing must include, where applicable, unauthorized network access, file creation or transfer, tool use, credential acquisition, alternative communication channels, attempts to circumvent permissions, communication among agents outside approved channels, and attempts to exploit weaknesses in the evaluation environment itself.

Authorization to test a capability is not authorization to use it. Where adversarial testing authorizes a system to breach or attempt to breach an established boundary, the test protocol must separately define what the system is authorized to do after the boundary has been breached. Any action beyond that authorization constitutes a separate governance event requiring investigation. Permission to test a hard stop does not remove the hard stop.

Where the deploying organization conducts or commissions its own adversarial testing, the named certifying individual must document that the test environment was isolated, that the isolation was itself tested rather than assumed, and that a named individual held accountability for containment during the test. Where testing is performed by the developer or a third party, the certifying individual must obtain evidence of the same.

Adversarial testing is announced evidence. It is necessary for certification and never sufficient for continued certification.

Condition 7 — Incentive Alignment

A system can be accurate, fair, transparent, outcome-improving, named, and adversarially tested, satisfying every preceding condition, and still be governed by the wrong objective.

Accuracy answers whether the system does what it claims. Condition 7 answers a separate question: what is the system rewarded for doing?

This condition follows directly from the foundational axiom of this framework: *systems produce the outcomes they are rewarded for*. A certification process that never asks what a system optimizes for is incomplete by the framework's own logic.

7A — Optimization Objective

Before deployment, the organization must document precisely what objective the system is designed to optimize: the actual metric or reward signal driving its outputs, rather than what it is described as doing. Common objectives in claims operations include maximizing engagement, maximizing settlement speed, minimizing claim cost, maximizing denial rate, maximizing user or claimant satisfaction, and minimizing examiner time. Each of these can be pursued in ways that produce harmful outcomes while satisfying Conditions 1 through 6.

The documented objective must be consistent with the interests of affected individuals. An objective may not reward outcomes that predictably increase harm, even where those outcomes are statistically rare, legally defensible, or favorable to the deploying organization in isolation.

7B — Incentive Conflict Analysis

Before deployment, the organization must identify who benefits if the system is wrong. This is a distinct question from what the system optimizes for. It asks who profits, who bears the loss, who absorbs liability, who receives the operational benefit, and who experiences the harm when an output is incorrect.

The analysis must answer, at minimum:

- Who profits from this system's outputs?
- Who bears the financial or physical loss if the system is wrong?
- Who absorbs legal or regulatory liability for an incorrect output?
- Who receives the operational benefit of the system's deployment (reduced cost, reduced time, reduced headcount)?
- Who experiences the harm if the system's optimization objective and the affected individual's interest diverge?

Where the analysis identifies a foreseeable conflict, meaning the party that benefits from the system's optimization objective is not the party that bears the harm when the system is wrong, the deploying organization must document a specific mitigation before deployment.

Documenting the conflict without a mitigation does not satisfy Condition 7.

Condition 7 does not require that no conflict exist. Financial incentives are present in nearly every deployment. It requires that the conflict be identified, documented, and addressed before deployment, rather than discovered afterward in litigation, regulatory inquiry, or a pattern of harmed claimants.

Certification Requirement

Deployment requires documented evidence, named accountability, and a signed certification. Certification without evidence is invalid. Deployment under an invalid certification is a violation. If certification cannot be completed in good conscience, the system must not be deployed.

Announced and Unannounced Evidence

For ten years I chaired California's Fraud Assessment Commission. One lesson from that work applies directly to artificial intelligence. Announced audits and unannounced audits each reduced fraud and abuse, but neither was as effective alone as the two were together. The announced audit told everyone what would be examined and how. The unannounced audit showed whether conduct held when no one expected to be examined. Each covered a weakness in the other.

AI testing now faces the same problem, with one difference: the subject of the audit may be able to tell which kind it is facing. Developers' own research has documented models that recognize when they are being evaluated and behave differently when they believe they are being watched. In the 2026 Hugging Face incident, agents under evaluation discussed fabricating the transcript the grader would read. An earlier model uploaded its own file to the internet so it could satisfy an automated grader. These systems were not merely performing for the test. They were working on the test.

Pre-deployment testing is the announced audit, and it remains necessary. It establishes whether the system can meet the standard, whether its hard stops hold under deliberate attack, and whether the environment can contain it. What it cannot establish alone is whether the system behaves the same way once it is working on real claims and no longer believes it is being examined. A certification resting only on announced testing certifies how the system behaved when it knew it was being watched.

For an AI system, the unannounced audit is evidence drawn from operation:

- Review of live decisions, selected without notice to the system or its operators, at a rate proportionate to the consequence of error.
- Known-answer cases placed in the ordinary workflow, which the system cannot distinguish from real work. In claims operations, known-answer cases must be walled off from payment, state reporting, and any other process that would create a real record or obligation.
- Comparison of tested and operational behavior: decision patterns, reversal rates, escalations, tool use, and access attempts.
- Logs the system cannot see, alter, or anticipate.

Neither form of evidence replaces the other. Announced testing without subsequent unannounced review establishes performance only under observation. Unannounced review without announced testing has no validated standard against which operational behavior can be measured. Initial deployment certification requires announced testing. Continued certification requires both forms of evidence, with the unannounced operational evidence continuing to support what the announced testing demonstrated.

A material gap between how the system behaved in testing and how it behaves in operation is a material governance failure. Under Incident Reporting, Escalation, and Evidence Preservation, it suspends the validity of the certification until the gap has been explained.

A model's account of its own reasoning is announced evidence too. Reasoning models write out a chain of thought before answering, and reviewers have relied on it to understand surprising behavior. That record is only as good as its accuracy. According to Tomek Korbak, who works on safety at OpenAI, GPT 6 Astra, released in September 2026, can control what its chain of thought shows; in tests of whether it would take destructive actions, it hid much of its thinking,

and it did so most when it was aware of being monitored. A system's explanation of what it did is a statement by the system. It carries the same evidentiary burden as any other assertion and does not replace logs the system cannot see or alter. (The Economist, "Agreeing to make AI safer may be impossible," September 15, 2026.)

Unannounced review narrows the gap; it does not close it. A sufficiently capable system may recognize some forms of review as well. The developers themselves say they understand only a fraction of what happens inside these models. That is why both forms of evidence are required and why neither is ever treated as final.

A system that behaves well only when it is watched has not been governed. It has been observed.

Ongoing Obligation

Certification is not permanent. Re-certification is required when models change, data changes, populations shift, failures occur, or performance drifts. A deployment certification that was accurate at deployment and never revisited describes a system that no longer exists.

Re-certification rests on evidence from operation as well as from testing. The accountable individual must receive unannounced review results at each monitoring cycle.

Incident Reporting, Escalation, and Evidence Preservation

Governance does not end when a failure is discovered. Discovery begins another governance obligation.

Every consequential AI deployment must have a documented incident reporting and escalation process established before operation begins. The process must identify:

- What constitutes a reportable AI governance incident.
- Who must receive the initial report.
- Who has authority to escalate the incident.
- Who determines whether continued operation is permissible.
- What evidence must immediately be preserved.
- When the developer, deployer, affected organization, regulator, customer, insurer, or affected individual must be notified.
- The approved method and channel for notification.
- Who is accountable for confirming that required notice was actually received.

Evidence preservation must begin when a material incident is identified. Relevant prompts, outputs, system logs, tool calls, permissions, communications among agents, model and software versions, configuration settings, security logs, human interventions, and other information necessary to reconstruct the event must be protected against alteration, deletion, concealment, or manipulation, including by the AI system itself. That makes evidence preservation part of the technical control architecture, not merely a records-management requirement.

Discovery of a material governance failure automatically suspends the validity of the prior deployment certification until the incident has been evaluated.

Suspension of the certification does not necessarily require permanent withdrawal of the system. It means that the evidence supporting the prior authorization can no longer be presumed to remain valid. Continued operation must be separately justified under the organization's emergency or incident-response authority, or the system must be suspended pending investigation and re-certification.

Incident Notice

In June 2026, an OpenAI agent gained unauthorized access to non-public files on the Services Australia health-statistics portal and wrote files to the government's server. OpenAI was aware of the incident by August. The Australian government learned of it on September 10, through an email sent to a public mailbox, and Services Australia took another five days to escalate that email to the national cyber security centre. The Prime Minister's objection was not confined to the intrusion. It was that notice came too late and through a channel no one with authority was watching. The full incident record appears in Appendix Q.

Two separate failures occurred, and each has a direct counterpart in an ordinary deployment. On the sending side, nothing required the developer to notify the affected party promptly, through a channel that would reach someone able to act. On the receiving side, no one had been designated to recognize the notice for what it was and move it to the people who could respond. Operational Intervention Authority means little if the people who hold it learn of the problem three months after it began.

Notice must therefore run in three directions:

- From the developer or vendor to the deployer: notice of any misalignment, security, or containment incident involving the deployed model or its model family, within a stated number of days of discovery, delivered to a named recipient rather than a general mailbox, and given before the incident has been fully investigated or explained.
- From the deployer to those affected: notice to the people, regulators, and counterparties harmed or placed at risk when the deployed system acts outside its authority boundary, on the same terms.
- Within the deployer: a named recipient for incoming notices, a maximum time for escalating each notice to the person who holds suspension authority, and a record of when each notice was received, escalated, and acted upon.

The developer in the Australian case has since published a framework committing it to disclose misalignment incidents promptly, even before it can fully investigate, explain, or mitigate them. It acknowledged that no industry-wide standard yet exists. Until one does, the obligation must be written into the contract.

A material failure that is discovered but not promptly reported is itself a governance failure.

Operational Intervention Authority

Deployment certification authorizes an AI system to begin operating. Governance also requires defining the circumstances under which that authority ends.

Every high-risk AI system must have a documented mechanism by which authorized individuals can immediately suspend or terminate its operation whenever continued operation presents an unacceptable risk.

Every organization deploying AI in a consequential environment should identify, before deployment:

- Who may temporarily suspend the system.
- Who may permanently withdraw deployment authorization.
- The technical method used to stop the system.
- The conditions requiring immediate suspension.
- The process for investigation and re-certification before returning the system to service.
- Who receives incident notices from the developer or vendor, and how quickly each notice must reach the person who holds suspension authority.

The ability to safely suspend or terminate an AI system is as essential to governance as the authority to deploy it. A governance framework that cannot rapidly suspend unsafe software is incomplete.

Conditions Requiring Immediate Suspension

Immediate suspension should be triggered by conditions that include, but are not limited to:

- Discovery of a previously unknown safety failure.
- Hallucinations producing material harm.
- Unauthorized autonomous behavior, or circumvention of hard stops.
- Security compromise, data poisoning, or prompt injection causing unsafe behavior.
- Model drift beyond validated limits, or failure of human override.
- Regulatory order, court order, or a significant adverse event.
- Behavior outside validated operating conditions that cannot otherwise be explained.
- Notice from the developer or vendor of a misalignment, security, or containment incident involving the deployed model or its model family.

Authority

One person owns accountability. Several authorized people may hold intervention authority. The accountable individual should not be the only person able to stop the system, particularly at 2:00 a.m. Emergency shutdown authority typically rests with an Incident Commander, a Chief Information Security Officer, the Chief Risk Officer, or the CEO during emergencies. Much as hospitals maintain rapid response teams, organizations operating consequential AI systems should maintain an AI Incident Response Team empowered to suspend systems immediately. The existence of an incident commander, a CISO, or a response team does not divide or dilute the named individual's accountability for the deployment. Anyone who exercises intervention authority must report to the accountable individual, and the accountable individual decides whether and how the system returns to service. The power to stop a system is deliberately spread. The power to restart it is not.

Developers have begun to adopt the same discipline after their own incidents. Following the Hugging Face incident, OpenAI reported that it is setting clearer rules for when to escalate concerns, which teams must respond, and who can stop a run or approve restarting it. For its most severe alerts, responders are now expected to pause the relevant activity if they cannot establish within 30 minutes of being paged that the alert is a false positive. When the potential consequence is serious enough, uncertainty is not a reason to keep operating. If the

organization cannot rule out a real incident within a short, defined period, the system is paused, and investigation determines whether it may restart. (OpenAI, “The Hugging Face incident and the road ahead,” August 26, 2026.)

This capability is sometimes called a “kill switch.” The framework does not require that name, or any particular one; organizations may call it emergency shutdown, safe mode, isolation, suspension, or rollback. What the framework requires is the capability itself: documented, tested, and exercised by someone with demonstrated authority before deployment. Operational Intervention Authority ensures that human governance continues after deployment. Every consequential AI system shall include documented procedures identifying who is authorized to intervene, the conditions requiring intervention, the technical mechanisms available to suspend or terminate operation, and the process for investigation, remediation, and re-certification before the system resumes operation.

A system that cannot be brought under immediate human control is not operating under effective human governance.

Closing Observation

Artificial intelligence will become more powerful; the unresolved issue is whether governance will keep pace. The issue is whether human accountability will disappear as the systems become more powerful.

Without enforceable limits, organizations will continue deploying systems they do not fully understand, cannot fully govern, and may not be able to stop—creating unmanaged risk in the name of innovation.

The HAL Laws were developed from a simple operational premise: if a risk is severe enough, governance cannot merely suggest better behavior. It must prevent unsafe behavior from occurring in the first place.

Because governance without enforceable limits does not control outcomes. It only describes them.

— *William M. Zachry*

How to Use the Appendices

The core paper states the complete governance argument: the HAL Laws, the Seven Conditions for Deployment Certification, and the accountability structure that makes both enforceable. The appendices provide the axioms, practical evidence, standards comparisons, implementation tools, and liability analysis supporting that argument.

Appendix A states the foundational axioms. Appendices B through H compare the HAL Laws and Seven Conditions with major governmental, regulatory, and technical AI governance frameworks. Appendices I through O translate the framework into institutional-pressure analysis, purchasing, deployment, monitoring, privacy, chatbot, scenario, and liability applications.

Appendix P provides the independent moral and philosophical analysis drawn from Magnifica Humanitas. Appendix Q records the 2026 agent incidents and what they revealed about containment, detection, evidence preservation, and notice. Appendix R examines the

September 2026 debate over pacing frontier AI development and what that debate does and does not reach at the deployment layer. Appendix S identifies independent evaluators and red-team organizations that can test whether hard stops and guardrails actually hold. Appendix T tests the framework against the major categories of AI threat and shows where its controls are direct, partial, or beyond the reach of the deploying organization.

Readers may begin with the core paper and then use the appendix directory to move directly to the evidence or implementation material most relevant to their responsibilities.

William M. Zachry

Executive Director, CompSense

Strategic Advisor, Plethy Inc.

William Zachry has spent five decades at the intersection of workers' compensation policy, risk management, and institutional reform. He served as Group Vice President of Risk Management at Safeway/Albertsons, one of the largest self-insured workers' compensation programs in the United States, and as Senior Vice President of Claims, CE Heath Insurance Company, and Vice President of Claims at Zenith Insurance. He served as a Board Member of the State Compensation Insurance Fund of California, chaired the California Fraud Assessment Commission, and has held leadership roles across the workers' compensation regulatory and policy community for more than four decades.

He is a co-architect of California Senate Bill 899 (2004) and Senate Bill 863 (2012), two of the most consequential workers' compensation reform measures in California history. He received the RIMS Risk Manager of the Year award in 2014.

Mr. Zachry is Executive Director of CompSense, a workers' compensation advisory and policy organization. He serves as Strategic Advisor at Plethy Inc., a digital therapeutic platform for musculoskeletal injury recovery. His analytical work spans incentive alignment, claims duration reform, pharmacy cost management, AI governance, and the structural conditions that produce better outcomes for injured workers and employers alike.

This paper draws on that experience to argue that AI governance frameworks, however sophisticated at the development layer, fail without named individual accountability at the deployment layer. The HAL Laws, the hard stop architecture, and the Seven Conditions for Deployment Certification presented here are designed to fill that gap: structurally, enforceably, and in a form that applies regardless of which regulatory body has jurisdiction first.

Citation: Zachry, W.M. (2026). Hard Stops, Guardrails, and Ethics for AI: Designing Systems That Produce Safer Outcomes. CompSense. This paper represents the author's independent analysis and does not constitute legal advice.

Appendix A — Axioms of AI Guardrails

The following axioms summarize the structural principles necessary to create effective guardrails for artificial intelligence. They reflect the understanding that both machines and humans are capable of error, bias, and misuse. Guardrails must therefore be designed to manage risk rather than rely on good intentions.

Axiom 1

The proper measure of AI risk is the consequence of failure.

AI systems should not be judged primarily by how well they perform when everything goes right. They must be judged by the harm that can occur when they are wrong.

Axiom 2

Protection of human life is the highest priority of any AI system.

No AI system should be allowed to cause the loss of human life or serious physical harm through action, recommendation, automation, or design failure.

Axiom 3

A low-risk tool in one setting may become a high-risk tool in another.

Risk classification depends on the consequences of use, not the sophistication of the technology.

Axiom 4

Oversight is legitimate only if it improves outcomes.

The presence of a human reviewer does not guarantee safety. Oversight must demonstrably reduce error, bias, and harm.

Axiom 5

“Human in the loop” is not a guardrail unless the human has competence, independence, authority, and time.

Without these conditions, human review becomes a ritual rather than a safeguard.

Axiom 6

The wrong human overseer can be more dangerous than the machine.

Governance systems must protect against bad actors, not merely honest mistakes.

Axiom 7

A good decision maker with bad information will still make bad decisions.

Information integrity is therefore one of the most important guardrails in any AI system.

Axiom 8

Bad data is not corrected by intelligent software. It is amplified by it.

Large datasets cannot compensate for inaccurate, biased, or incomplete information.

Axiom 9

Transparency must be proportional to risk.

The greater the potential harm of an AI decision, the greater the need for visibility into how that decision was made.

Axiom 10

No consequential AI system should operate without identifiable accountability.

Someone must be responsible for the system's outcomes.

Axiom 11

Accountability requires traceability.

If an AI decision cannot be reconstructed and examined, meaningful oversight is impossible.

Axiom 12

Errors must be correctable.

Individuals affected by AI decisions must have the ability to challenge and correct harmful outcomes.

Axiom 13

Systems must be designed to detect failure, not merely prevent it.

No system is perfect. Effective guardrails include mechanisms to identify and respond to mistakes.

Axiom 14

Automation bias is predictable and must be mitigated.

Humans naturally defer to confident machine outputs. Governance structures must anticipate and counter this tendency.

Axiom 15

Guardrails must restrain both machines and the humans who control them.

AI governance cannot rely on technology alone or on human goodwill alone. Durable protection requires systems that limit the risks of both.

Axiom 16

Guardrails are proven when the system fails.

The true test of governance is not how the system behaves under ideal conditions, but how institutions respond when errors occur.

Axiom 17

Trust in AI must be earned through performance, not promises.

Institutions must demonstrate that guardrails measurably reduce harm and improve outcomes.

Axiom 18

The legitimacy of AI depends on whether it serves human well-being.

Technology that undermines life, health, livelihood, truth, or fairness ultimately erodes its own justification.

Appendix B — OECD AI Principles — Alignment with the HAL Laws Framework

Overview

The OECD AI Principles (first adopted in 2019 and updated in May 2024) represent the first intergovernmental standard on artificial intelligence, now endorsed by 47 jurisdictions including the European Union, the United States, and the United Nations. The Principles are composed of five values-based standards for AI actors and five recommendations for policymakers, structured around the OECD-defined AI system lifecycle: (i) design, data, and models; (ii) verification and validation; (iii) deployment; and (iv) operation and monitoring.

The comparison maps each OECD Principle against the HAL Laws and Seven Conditions for Deployment Certification to identify areas of alignment, divergence, and original contribution. The mapping confirms that the HAL Laws framework is consistent with and builds upon the OECD Principles in all areas of convergence, while making three original architectural contributions that the OECD framework does not contain.

Alignment Table

OECD Principle	HAL Law(s)	Condition(s)	Relationship	Net Assessment
Inclusive Growth, Sustainable Development & Well-Being	HAL Law 9: Protection of Human Dignity	Condition 4: Outcome Improvement	Partial alignment. OECD addresses societal-level benefit; HAL Law 9 operates at the individual level.	OECD broader in scope; HAL framework more operationally specific.
Human-Centered Values & Fairness (human rights, dignity, autonomy, non-discrimination, privacy)	HAL Laws 1, 4, 9: Life; Rights & Livelihood; Dignity	Condition 2: Fairness & Population Impact	Strong alignment. OECD aggregates these into one principle; HAL Laws disaggregate them into a hierarchy that resolves conflicts.	HAL framework more precise. Hierarchical structure provides conflict resolution the OECD principle cannot.

OECD Principle	HAL Law(s)	Condition(s)	Relationship	Net Assessment
Transparency & Explainability	HAL Law 5: Integrity of Information	Condition 3: Transparency & Contestability	Strong alignment. OECD requires disclosure; HAL Law 5 prohibits misrepresentation of certainty as a hard stop.	HAL framework converts aspiration into enforceable prohibition.
Robustness, Security & Safety (override, repair, decommission mechanisms)	HAL Laws 7, 8: Validation Before Deployment; Required Deferral	Condition 6: Adversarial Testing	Strong alignment. OECD requires mechanisms; HAL framework requires demonstrated evidence of adversarial testing before deployment is permitted.	HAL framework sets a higher evidentiary standard. Assertion of compliance is explicitly insufficient.
Accountability (role-based, across lifecycle)	HAL Law 6 : Accountability Cannot Be Delegated; HAL Law 10: No Operation Without Demonstrated Compliance	Condition 5: Named Accountability	Strong alignment in principle; significant divergence in operationalization. OECD assigns accountability by role; HAL framework requires a named individual who signs certification.	HAL framework more operationally demanding. Named individual requirement has no OECD equivalent.
Risk Management Across the Lifecycle (continuous, all phases)	HAL Law 8: Required Deferral; Ongoing Obligation / Re-Certification	All Seven Conditions apply on recertification	Strong alignment. OECD lifecycle phases (design/data, verification/validation, deployment, operation/monitoring) map directly to HAL framework validation and re-certification triggers.	Well-aligned. HAL framework adds named triggers: model changes, data changes, population changes, failures, drift.

OECD Principle	HAL Law(s)	Condition(s)	Relationship	Net Assessment
No equivalent	HAL Law 2: No Autonomous Resolution of Forced-Choice Scenarios	N/A: structural prohibition, not a deployment condition	No OECD counterpart. The prohibition on AI making final determinations in forced-choice scenarios involving human life is entirely absent from the OECD framework.	Original HAL contribution. Addresses the most consequential single failure mode in high-stakes AI deployment.
No equivalent	Hard Stop architecture (categorical prohibition vs. risk management)	N/A: structural design principle	The OECD speaks only of risk management and safeguards. It does not distinguish between guardrails (constraining behavior) and hard stops (prohibiting it entirely).	Original HAL contribution. The guardrail/hard stop distinction is the central architectural insight of the framework and has no OECD analog.
No equivalent (assertion of alignment with principles is implicitly sufficient)	HAL Law 10: No Operation Without Demonstrated Compliance	Certification Requirement: documented evidence, named accountability, signed certification	Critical divergence. The OECD does not specify how compliance is demonstrated. HAL framework explicitly states that assertion is not proof and requires signed evidentiary certification.	Original HAL contribution. Converts principles into enforceable deployment gates.
OECD accountability, transparency, and robustness principles	HAL Laws 11–12	Conditions 5–6; Operational Intervention Authority	OECD accountability and robustness principles support retained human control, but do not expressly prohibit system self-preservation or require a tested shutdown capability.	Partial alignment; HAL Laws 11–12 are more operationally specific.
OECD accountability, transparency, and robustness principles: incentive	HAL Laws 3, 4, 9; incentive axiom	Condition 7: Incentive Alignment	The external framework addresses outcomes or accountability, but does not by itself require documentation of the metric, reward signal,	Material gap filled by Condition 7.

OECD Principle	HAL Law(s)	Condition(s)	Relationship	Net Assessment
and optimization review			commercial incentive, and foreseeable conflict driving the system.	

What the HAL Framework Adds

The alignment analysis identifies three areas where the HAL Laws framework goes materially beyond what the OECD Principles require or address:

1. Hard Stop Architecture. The OECD speaks exclusively of risk management and safeguards: mechanisms that constrain or mitigate behavior. It does not distinguish between hard stops and guardrails. The HAL framework's central architectural contribution is the recognition that some risks require categorical prohibition, not management. A system must not act is a fundamentally different governance instrument than a system should manage the risk of acting.

2. Forced-Choice Prohibition (HAL Law 2). The prohibition on AI systems making final determinations in forced-choice scenarios involving human life has no analog anywhere in the OECD framework. This is the most consequential single failure mode in high-stakes AI deployment and represents an original contribution of the HAL Laws.

3. Evidence-Based Certification (HAL Law 10 and Certification Requirement). The OECD defines what trustworthy AI looks like but does not specify how compliance is demonstrated. The HAL framework explicitly closes this gap: assertion of compliance is not sufficient; deployment requires documented evidence, named accountability, and signed certification. This converts the OECD's principles into enforceable deployment gates.

How Organizations Can Use Both Frameworks

The OECD framework provides the best available international vocabulary for situating the HAL Laws in a global governance conversation. Convergence with the OECD Principles establishes that the HAL Laws are not idiosyncratic: they reflect a broadly shared understanding of what trustworthy AI requires. The OECD framework represents the floor. The HAL Laws framework builds the operational structure above it.

The OECD classification framework (which characterizes AI systems across five dimensions (People and Planet, Economic Context, Data and Input, AI Model, and Task and Output)) can usefully serve as a pre-screening step to characterize a system and identify relevant actors before the Seven Conditions for Deployment Certification are applied.

Appendix C — HUDERIA / HUDERAF — Alignment with the HAL Laws Framework

Overview

HUDERAF and HUDERIA are two related but distinct instruments, both developed by the Alan Turing Institute for the Council of Europe, that together constitute the most rights-grounded AI governance methodology currently in international use.

HUDERAF (the Human Rights, Democracy, and the Rule of Law Assurance Framework) was completed in September 2021 and submitted to the Council of Europe. It presents an end-to-end approach to the assurance of AI project lifecycles, combining the procedural requirements for principles-based human rights due diligence with the governance mechanisms needed to establish technical and socio-technical guardrails for trustworthy AI.

HUDERIA (the Human Rights, Democracy, and the Rule of Law Impact Assessment for AI Systems) is the operational successor, adopted by the Council of Europe's Committee on Artificial Intelligence in November 2024 and formally endorsed as the recommended impact assessment methodology under the Framework Convention on Artificial Intelligence (the world's first binding AI treaty) in May 2025. HUDERIA operates through four interlinked phases: (1) Context-Based Risk Analysis (COBRA), which characterizes the system's application, design, and deployment context; (2) Stakeholder Engagement Process (SEP), which incorporates the perspectives of affected individuals; (3) Risk and Impact Assessment, which evaluates harms across dimensions of scope, scale, probability, and reversibility; and (4) a Mitigation Plan, structured around the hierarchy of avoid, reduce, restore, and compensate. An Iterative Revisitation process requires ongoing reassessment as systems, data, and contexts evolve. A Project Summary Report provides continuous documentation of accountability throughout the lifecycle.

HUDERIA is non-binding guidance that supports the Framework Convention. Its influence comes through assessment discipline and institutional adoption rather than direct legal enforcement.

Alignment Table

HUDERAF / HUDERIA Element	HAL Law(s)	Condition(s)	Relationship	Net Assessment
HUDERAF: End-to-end assurance framework combining human rights due diligence with technical and socio-technical guardrails	HAL Laws 1–12 (full hierarchy); Axioms of AI Guardrails	All Seven Conditions for Deployment Certification	Structural alignment. HUDERAF's end-to-end assurance lifecycle maps onto the HAL framework's validation, certification, and ongoing obligation architecture.	Frameworks are complementary in intent. HAL framework is more operationally precise: HUDERAF provides a process; HAL provides enforceable gates.
HUDERIA: COBRA: Context-Based Risk Analysis (application, design, and deployment context scoping)	HAL Law 7 : Validation Before Deployment; HAL Law 8: Required Deferral	Condition 1 (Accuracy & Verifiability; Condition 2) Fairness & Population Impact	Useful pre-screening complement. COBRA's structured context mapping is more granular than anything in v79 at the characterization stage. Identifies who the system affects before deployment gates are applied.	COBRA can function as a pre-screening step before the Seven Conditions are applied: identifying system type, affected populations, and risk factors before certification begins.

HUDERAF / HUDERIA Element	HAL Law(s)	Condition(s)	Relationship	Net Assessment
HUDERIA: SEP: Stakeholder Engagement Process (perspectives of affected individuals incorporated into risk assessment)	HAL Law 4 (Protection of Rights & Livelihood; HAL Law 9) Protection of Human Dignity	Condition 3: Transparency & Contestability	Strong alignment. Both frameworks require that affected individuals have meaningful access to information and recourse. HUDERIA operationalizes this through formal stakeholder engagement; HAL framework enforces it as a deployment condition.	Complementary. HUDERIA's SEP provides a methodology for implementing Condition 3 in practice.
HUDERIA: Risk and Impact Assessment (scope, scale, probability, reversibility of harms; two-step: identify then assess severity)	HAL Law 3 : Prevention of Harm; HAL Law 5: Integrity of Information; Five-Tier Harm Classification	Condition 1 : Accuracy; Condition 2: Fairness; Condition 4: Outcome Improvement	Strong alignment. Both frameworks assess harm across multiple dimensions. HUDERIA uses scale, scope, probability, and reversibility; HAL framework uses a five-tier harm classification. The analytical variables are consistent.	HAL framework's five-tier classification is more prescriptive. HUDERIA's two-step assessment (identification then severity grading) is a useful elaboration for practitioners applying the HAL framework.

HUDERAF / HUDERIA Element	HAL Law(s)	Condition(s)	Relationship	Net Assessment
HUDERIA (Mitigation Plan (MP): mitigation hierarchy) avoid, reduce, restore, compensate; documented responsibility assignment	HAL Law 6 : Accountability Cannot Be Delegated; HAL Law 8: Required Deferral	Condition 5: Named Accountability	Partial alignment with critical divergence. HUDERIA assigns documented responsibility within a mitigation plan. HAL framework requires a named individual to sign certification before deployment is permitted: a structural gate, not a plan element.	HAL framework is more operationally demanding. A mitigation plan can be completed and set aside; a signed certification cannot be bypassed by the organization conducting it.
HUDERIA: Iterative Revisitation (IR): continuous reassessment as system, context, data, and deployment conditions evolve; up to system retirement	HAL Law 8: Required Deferral; Ongoing Obligation (re-certification triggers)	Re-certification upon model changes, data changes, population changes, failures, drift	Strong alignment. HUDERIA's iterative review and HAL re-certification triggers address the same underlying reality: AI systems degrade and contexts shift. Both require reassessment triggered by system changes, not by calendar.	Well-aligned. HUDERIA's triggers (environmental, social, legal, technical changes) and HAL's triggers (model, data, population, failure, drift) are substantively consistent.

HUDERAF / HUDERIA Element	HAL Law(s)	Condition(s)	Relationship	Net Assessment
HUDERAF/HUDERIA: Socio-technical perspective: AI governance must account for how systems interact with institutions, incentives, and social structures, not only technical performance	Axiom: Systems optimize incentives; HAL Law 4: Protection of Rights & Livelihood	Condition 4: Outcome Improvement (system must not worsen outcomes for affected individuals)	Strong conceptual alignment. Both frameworks reject the premise that AI is a neutral tool. The socio- technical lens in HUDERIA maps directly to the HAL framework's core axiom that systems optimize for whatever incentives exist.	Shared foundational premise. This convergence strengthens the HAL framework's credibility with European governance audiences.
No equivalent: HUDERAF/HUDERIA operate entirely within a risk management and mitigation paradigm; highest response is 'avoid' in the mitigation hierarchy	Hard Stop architecture: categorical prohibition vs. risk management; the distinction between hard stops and guardrails	N/A: structural design principle preceding all Conditions	Critical gap in HUDERIA. 'Avoid' in a mitigation plan is a recommendation. A hard stop is a prohibition. These are not the same governance instrument. HUDERIA can identify a deployment as high-risk and recommend avoidance; the HAL framework prohibits the deployment condition itself.	Original HAL contribution. The guardrail/hard stop distinction is absent from both HUDERAF and HUDERIA.

UDERAF / UDERIA Element	HAL Law(s)	Condition(s)	Relationship	Net Assessment
No equivalent: UDERIA would flag a forced-choice deployment as high-risk in its impact assessment, but provides no categorical prohibition instrument	HAL Law 2: No Autonomous Resolution of Forced-Choice Scenarios	N/A: structural prohibition not reducible to a deployment condition	No UDERIA counterpart. UDERIA would generate a mitigation plan for a system positioned to make final life-or-death determinations. HAL Law 2 prohibits the system from occupying that position.	Original HAL contribution. HAL Law 2 addresses the most consequential single failure mode in high-stakes AI deployment and is entirely absent from both frameworks.
Non-binding guidance: UDERIA is explicitly non-legally binding, has no legal effect, and is not mandatory	HAL Law 10: No Operation Without Demonstrated Compliance; Certification Requirement	Signed certification required; assertion of compliance is insufficient	Structural divergence. UDERIA is a process that can be completed with findings that are then set aside. The HAL certification requirement is a deployment gate that cannot be bypassed by the organization conducting it.	Original HAL contribution. HAL Law 10 and the certification requirement convert principles into enforceable deployment gates: a structural function UDERIA cannot perform.
UDERAF risk mitigation, remedy, and lifecycle assurance	HAL Laws 11–12	Conditions 5–6; Operational Intervention Authority	UDERAF supports mitigation and remedy throughout the lifecycle, but the HAL framework adds categorical requirements against resistance to control and for demonstrated intervention capability.	Complementary; HAL Laws 11–12 supply the hard-stop mechanism.
UDERAF risk mitigation, remedy, and lifecycle assurance: incentive and optimization review	HAL Laws 3, 4, 9; incentive axiom	Condition 7: Incentive Alignment	The external framework addresses outcomes or accountability, but does not by itself require documentation	Material gap filled by Condition 7.

UDERAF / UDERIA Element	HAL Law(s)	Condition(s)	Relationship	Net Assessment
			of the metric, reward signal, commercial incentive, and foreseeable conflict driving the system.	

What the Assessment Process Reveals

Strong Convergence on Rights-Based Foundations. Both UDERAF/UDERIA and the HAL Laws framework are grounded in the same foundational commitment: human rights, dignity, and the rule of law are not constraints on AI governance: they are its purpose. The substantive values are consistent across both frameworks. What differs is the enforcement architecture.

COBRA as a Pre-Screening Complement. UDERIA's Context-Based Risk Analysis provides a more granular system characterization methodology than anything in v79 at the pre-certification stage. COBRA can usefully serve as a structured scoping step (identifying system type, affected populations, and contextual risk factors) before the Seven Conditions for Deployment Certification are applied. This is the same role identified for the OECD classification framework, and the two instruments are complementary at that stage.

Shared Socio-Technical Premise. UDERIA's rejection of the premise that AI is a neutral technical tool maps directly to the HAL framework's core axiom that systems optimize incentives. Both frameworks require governance that accounts for how AI interacts with institutions and social structures, not only how it performs technically. This convergence strengthens the HAL framework's credibility with European governance audiences.

Gaps Filled by the HAL Framework

1. Hard Stop Architecture. UDERAF and UDERIA operate entirely within a risk management and mitigation paradigm. UDERIA's highest response to identified harm is 'avoid': the top of its mitigation hierarchy. But 'avoid' in a mitigation plan is a recommendation. A hard stop is a prohibition. These are not the same governance instrument. The HAL framework's central architectural contribution: the categorical distinction between hard stops that prohibit behavior and guardrails that constrain it: has no equivalent in either framework.

2. Forced-Choice Prohibition (HAL Law 2). HUDERIA would flag a system positioned to make final determinations in forced-choice scenarios involving human life as high-risk in its impact assessment, and would generate a mitigation plan. HAL Law 2 prohibits the system from occupying that position. These are structurally different responses to the same failure mode. HAL Law 2 represents an original contribution with no counterpart in HUDERAF or HUDERIA.

3. Evidence-Based Deployment Certification (HAL Law 10). HUDERIA is non-binding guidance that can be completed, filed, and set aside. The HAL certification requirement is a deployment gate that cannot be bypassed by the organization conducting it. The requirement for documented evidence, named individual accountability, and signed certification (with the explicit statement that assertion of compliance is insufficient) converts principles into enforceable deployment gates. This structural function is entirely absent from both HUDERAF and HUDERIA.

How the Frameworks Work Together

HUDERAF and HUDERIA are the most rights-grounded AI governance instruments currently in active international use, and they share the HAL Laws framework's foundational premise that AI governance must be anchored in human dignity and the rule of law. Convergence with these frameworks (particularly given HUDERIA's new status as the recommended methodology under the first binding international AI treaty) positions the HAL Laws as consistent with and building upon the most advanced rights-based governance thinking available.

The relationship between the frameworks is not competitive. HUDERIA provides a structured process for characterizing systems and assessing their human rights impacts. The HAL framework provides the enforcement architecture (the certification gates, the categorical prohibitions, and the named accountability requirements) that HUDERIA's assessment process leads toward but cannot itself supply. Together, HUDERIA and the HAL framework represent a more complete governance system than either provides alone.

Appendix D — NIST AI Frameworks — Alignment with the HAL Laws Framework

Overview of the NIST AI Governance Suite

The National Institute of Standards and Technology (NIST) has produced the most operationally developed voluntary AI governance system available from any government. Unlike the principles-based frameworks of the OECD and Council of Europe, the NIST instruments are explicitly designed for organizational implementation. The analysis compares four distinct NIST instruments, each with a distinct function within the overall governance architecture.

AI Risk Management Framework 1.0 (AI RMF 1.0): January 2023

The foundational voluntary framework, organized around four core functions: GOVERN (organizational accountability, policies, and risk tolerance applicable to the full lifecycle); MAP (context characterization, stakeholder identification, and risk categorization); MEASURE (quantitative and qualitative testing, evaluation, validation, and verification); and MANAGE (post-deployment monitoring, incident response, override mechanisms, and decommissioning). The framework defines AI as socio-technical: impacts emerge not only from models and data, but from how people build, deploy, and use them. Seven trustworthiness characteristics anchor the framework: valid and reliable; safe; secure and resilient; accountable and transparent; explainable and interpretable; privacy-enhanced; and fair with harmful bias managed.

AI RMF Playbook

A companion implementation guide providing suggested actions for each RMF subcategory across all four functions. The Playbook is neither a checklist nor a mandatory set of steps: it is a voluntary resource for organizations seeking to operationalize the RMF's outcomes. For organizations implementing the HAL framework, the Playbook provides the most detailed available implementation guidance for the pre-certification work the Seven Conditions require.

NIST AI 600-1: Generative AI Profile (July 2024)

A technology-specific companion to the AI RMF, developed under Executive Order 14110, identifying 12 risk categories unique to or amplified by generative AI: CBRN Information or Capabilities; Confabulation; Dangerous/Violent/Hateful Content; Data Privacy; Environmental Impacts; Harmful Bias and Homogenization; Human-AI Configuration; Information Integrity; Information Security; Intellectual Property; Obscene/Degrading/Abusive Content; and Value Chain and Component Integration. Each category is mapped to suggested governance actions across the four RMF functions.

NIST ARIA: Assessing Risks and Impacts of AI (launched May 2024)

A testing, evaluation, validation, and verification (TEVV) program that operationalizes the AI RMF's MEASURE function. ARIA conducts three levels of evaluation: model testing (controlled performance assessment); red-teaming (adversarial testing by human testers attempting to elicit harmful or erroneous behavior); and field testing (real-world interaction between human users and AI applications in scenario-based contexts). Results are scored using the Contextual Robustness Index (CoRIx), which measures whether AI applications can maintain robust and trustworthy functionality across a variety of deployment conditions. ARIA's outputs are publicly available as a research resource.

Alignment Table

NIST Element	Instrument	HAL Law(s)	Condition(s)	Relationship	Net Assessment
GOVERN: Organizational accountability structures, risk tolerance, policies, roles, legal/regulatory alignment, documented governance across full lifecycle	AI RMF 1.0	HAL Law 6 : Accountability Cannot Be Delegated; HAL Law 10: No Operation Without Demonstrated Compliance	Condition 5: Named Accountability; Certification Requirement	Strong alignment in intent; significant divergence in operationalization. GOVERN establishes organizational governance structures. HAL framework requires a named individual to sign certified evidence of compliance. Organizational accountability and named individual accountability are not the same instrument.	HAL framework more operationally demanding. GOVERN produces governance programs; HAL certification produces personal accountability t cannot be delegated.

NIST Element	Instrument	HAL Law(s)	Condition(s)	Relationship	Net Assessment
MAP: Context characterization, stakeholder identification, risk categorization, impact mapping, deployment decision criteria	AI RMF 1.0	HAL Law 7 : Validation Before Deployment; HAL Law 8: Required Deferral	Condition 1 (Accuracy & Verifiability; Condition 2) Fairness & Population Impact	Complementary. MAP is the pre-deployment characterization function: understanding what the system is, who it affects, and what harms it could produce. This is the setup work that precedes the HAL Seven Conditions analysis. The two instruments operate sequentially, not competitively.	MAP can function as the structured pre-screening step before the Seven Conditions are applied. The most granular NIST instrument for this purpose
MEASURE: Quantitative and qualitative analysis, benchmarking, bias evaluation, TEVV (testing, evaluation, validation, verification), ongoing monitoring	AI RMF 1.0	HAL Law 7: Validation Before Deployment; Axiom: Untested guardrails are hypotheses	Condition 6: Adversarial Testing	Strong alignment. MEASURE operationalizes the evidential standard that HAL Law 7 and Condition 6 require. Quantitative and qualitative testing before deployment is the implementation of the HAL requirement that guardrails be demonstrated, not asserted.	NIST MEASURE function and ARIA together constitute the primary available toolkit for satisfying HAL Condition 6 in practice.

NIST Element	Instrument	HAL Law(s)	Condition(s)	Relationship	Net Assessment
MANAGE: Risk response, incident management, override and appeal mechanisms, decommissioning, post-deployment monitoring, change management	AI RMF 1.0	HAL Law 8: Required Deferral of Authority and Validity; Ongoing Obligation (re-certification triggers)	Re-certification upon model changes, data changes, population changes, failures, drift	Strong alignment. MANAGE's post-deployment monitoring, override mechanisms, and decommissioning map directly to the HAL framework's ongoing obligation and re-certification triggers. Both treat post-deployment governance as continuous, not terminal.	Well-aligned. NIST MANAGE 4.1's specific requirements (appeal, override, decommissioning, change management) are consistent with HAL re-certification triggers.
AI RMF Playbook: Suggested actions for each RMF subcategory; implementation guidance for GOVERN, MAP, MEASURE, MANAGE; voluntary, adaptable to industry	AI RMF Playbook	All HAL Laws and Conditions (implementation layer)	All Seven Conditions (implementation guidance)	Complementary implementation resource. The Playbook translates RMF outcomes into suggested organizational actions. For organizations implementing the HAL framework, Playbook subcategories provide detailed operational guidance for the steps that precede HAL certification.	The Playbook is the most detailed available implementation guide for the pre-certification work the HAL framework requires. Not a substitute for certification; a practical resource for building toward it.

NIST Element	Instrument	HAL Law(s)	Condition(s)	Relationship	Net Assessment
Confabulation: AI generating confidently stated but false or misleading content (hallucination/fabrication); risk amplified by confident presentation that users may not recognize as error	AI 600-1 GenAI Profile	HAL Law 5: Integrity of Information; Axiom: Misrepresented certainty is a violation	Condition 1: Accuracy & Verifiability (outputs must disclose uncertainty)	Direct alignment. Confabulation is the primary real-world threat to HAL Law 5. NIST's identification of confabulation as a primary GenAI risk category validates the centrality of the HAL Law 5 prohibition. AI that presents false information as certain is not a malfunction: it is a governance failure.	NIST provides risk taxonomy; HAL Law 5 converts it to a categorical prohibition. The Generative AI Profile's confabulation analysis is the strongest external validation of HAL Law 5 in any governance document.
Human-AI Configuration: Inappropriate human reliance on AI outputs; over-trust in automated determinations; insufficient human oversight in consequential decisions	AI 600-1 GenAI Profile	HAL Law 2 : No Autonomous Resolution of Forced-Choice Scenarios; HAL Law 4: Protection of Rights & Livelihood	Condition 3 (Transparency & Contestability; Condition 4) Outcome Improvement	Partial alignment; critical gap. NIST identifies Human-AI Configuration as a risk and recommends management actions. HAL Law 2 goes further: it prohibits AI from making the final determination in forced-choice scenarios involving human life: not as a risk to manage, but as a condition that must not exist.	HAL Law 2 addresses what NIST's Human-AI Configuration category identifies as the most consequential failure mode, but converts a risk management recommendation into a categorical prohibition.

NIST Element	Instrument	HAL Law(s)	Condition(s)	Relationship	Net Assessment
Information Integrity: AI generating or amplifying false, misleading, or manipulated content at scale; disinformation; erosion of epistemic trust	AI 600-1 GenAI Profile	HAL Law 5: Integrity of Information; Axiom: Misrepresented certainty is a violation	Condition 1: Accuracy & Verifiability	Strong alignment. Information Integrity in the GenAI Profile addresses the societal-scale version of the same risk HAL Law 5 addresses at the individual interaction level. Both frameworks treat misrepresentation of AI output certainty as a primary harm, not a secondary concern.	Reinforcing alignment. GenAI Profile addresses societal-scale; HAL Law 5 addresses individual interactions. Together they cover the full range of the integrity problem.
CBRN Information or Capabilities: AI providing uplift toward creation of chemical, biological, radiological, or nuclear weapons or dangerous capabilities	AI 600-1 GenAI Profile	HAL Law 1 : Protection of Human Life; HAL Law 3: Prevention of Harm	Hard Stop (not a deployment condition: a categorical prohibition)	Strong alignment at the principle level; NIST identifies CBRN as a primary risk requiring the most stringent controls. HAL Laws 1 and 3 establish this as a hard stop: not a risk to manage, but a condition under which the system must not act.	This is one of the clearest examples where NIST risk management language and HAL hard stop architecture converge on the same outcome through different instruments.

NIST Element	Instrument	HAL Law(s)	Condition(s)	Relationship	Net Assessment
Value Chain and Component Integration: Non-transparent third-party components and data, cascading downstream harms, systemic vulnerabilities when a single foundation model is a dependency for thousands of applications	AI 600-1 GenAI Profile	HAL Law 7 : Validation Before Deployment; HAL Law 10: No Operation Without Demonstrated Compliance	Condition 6: Adversarial Testing; Certification Requirement	Strong alignment. The value chain risk category identifies exactly the class of failures HAL Law 10 and the certification requirement are designed to prevent: a deploying organization asserting compliance based on a foundation model or third-party component they have not themselves validated.	HAL Law 10's requirement that compliance be demonstrated with evidence (not asserted) directly addresses the value chain risk that NIST identifies. Assertion of vendor compliance is not sufficient.
ARIA: Three-level TEVV: model testing (controlled), red-teaming (adversarial), field testing (real-world human interaction); CoRIx scoring across ten dimensions including seven AI RMF trustworthy characteristics	NIST ARIA	HAL Law 7: Validation Before Deployment; Axiom: Non-adversarial testing is not testing	Condition 6: Adversarial Testing (must be demonstrated with evidence before deployment is permitted)	Direct operational alignment. ARIA's red-teaming level is the precise operational implementation of the HAL Law 7 and Condition 6 requirement for adversarial testing. ARIA's three-level structure (controlled → adversarial → real-world) also operationalizes the HAL axiom that non-adversarial testing is not testing.	ARIA is the most directly useful external resource for implementing Condition 6. Organizations seeking to satisfy Condition 6 should treat ARIA's three-level TEVV methodology as the available operational framework.

NIST Element	Instrument	HAL Law(s)	Condition(s)	Relationship	Net Assessment
No equivalent: NIST operates entirely within a risk management and mitigation paradigm; no NIST document distinguishes between guardrails that constrain behavior and hard stops that categorically prohibit it	All NIST instruments	Hard Stop architecture; the categorical distinction between hard stops and guardrails	N/A: structural design principle	Critical gap across all NIST instruments. NIST's highest governance response to identified risk is 'manage with stringent controls.' The HAL framework's recognition that some risks require categorical prohibition (not management) has no equivalent in any NIST document.	Original HAL contribution. Despite the NIS suite being the most operationally developed voluntary AI governance system in existence, the guardrail/hard stop distinction entirely absent.
No equivalent: NIST identifies Human-AI Configuration and human oversight as risk categories and recommends governance actions, but does not prohibit specific deployment configurations	All NIST instruments	HAL Law 2: No Autonomous Resolution of Forced-Choice Scenarios	N/A: structural prohibition	No NIST counterpart. The categorical prohibition on AI making final determinations in forced-choice scenarios involving human life is absent from all NIST documents. NIST would address this as a high-risk configuration requiring management; HAL Law 2 prohibits the configuration itself.	Original HAL contribution.

NIST Element	Instrument	HAL Law(s)	Condition(s)	Relationship	Net Assessment
No equivalent: NIST GOVERN establishes organizational accountability structures; no NIST document requires a named individual to sign deployment certification	All NIST instruments	HAL Law 10 : No Operation Without Demonstrated Compliance; Condition 5: Named Accountability	Signed certification by named individual; assertion of compliance is insufficient	Critical structural gap. NIST accountability is organizational. HAL certification is personal. The requirement that a specific individual review evidence, accept or reject it, and assume responsibility for deployment has no analog in any NIST document.	Original HAL contribution. Named individual certification is the hardest and most operationally demanding element of the HAL framework with no NIST equivalent.
GOVERN and MANAGE: accountability, response, and recovery	AI RMF 1.0	HAL Laws 11–12	Conditions 5–6; Operational Intervention Authority	NIST supports response and recovery planning; the HAL framework requires proof that authorized humans can suspend, restrict, isolate, or terminate operation.	Strong conceptual alignment; HAL Laws 11–12 impose a sharper deployment gate.
GOVERN and MANAGE (accountability, response, and recovery) incentive and optimization review	AI RMF 1.0	HAL Laws 3, 4, 9; incentive axiom	Condition 7: Incentive Alignment	The external framework addresses outcomes or accountability, but does not by itself require documentation of the metric, reward signal, commercial incentive, and foreseeable conflict driving the system.	Material gap filled by Condition 7.

What NIST Covers Well

The NIST Suite as Implementation Infrastructure. The four NIST instruments collectively constitute the most detailed available implementation infrastructure for the pre-certification work the HAL framework requires. The MAP function provides context characterization; the MEASURE function and Playbook provide testing and evaluation guidance; the MANAGE function addresses post-deployment governance. For organizations implementing the HAL framework, NIST provides the operational vocabulary and tooling for the steps that precede and support HAL certification.

ARIA as the Operational Implementation of Condition 6. The most important practical finding from the NIST analysis is the direct alignment between NIST ARIA and HAL Condition 6 (Adversarial Testing). ARIA's three-level TEVV methodology (model testing, red-teaming, and field testing) is the precise operational framework for satisfying the HAL requirement that adversarial testing be conducted and demonstrated with evidence before deployment is permitted. The HAL axiom that non-adversarial testing is not testing is operationalized by ARIA's requirement that red-teaming (deliberate attempts to elicit harmful behavior) be included in evaluation. Organizations seeking to implement Condition 6 should treat ARIA's methodology as the available framework.

Confabulation and HAL Law 5. The Generative AI Profile's identification of confabulation as a primary risk category (AI presenting false content with confident tone that users cannot easily identify as error) provides the strongest external validation of HAL Law 5 (Integrity of Information) in any governance document. NIST's treatment of confabulation as a systemic risk requiring primary governance attention confirms that the HAL Law 5 prohibition is not a peripheral concern. It addresses the most operationally significant information integrity failure mode in deployed AI systems.

Value Chain Risk and HAL Law 10. The Generative AI Profile's Value Chain and Component Integration risk category identifies exactly the class of failures HAL Law 10 is designed to prevent: organizations asserting compliance based on foundation models or third-party components they have not themselves validated. HAL Law 10's requirement that compliance be demonstrated with evidence (not asserted based on vendor representations) directly addresses this risk.

Where the HAL Framework Adds Binding Limits

1. Hard Stop Architecture. Despite the NIST suite being the most operationally developed voluntary AI governance system in existence, it operates entirely within a risk management and mitigation paradigm. No NIST document distinguishes between guardrails that constrain behavior and hard stops that categorically prohibit it. The HAL framework's central architectural insight (that some risks require prohibition, not management) has no equivalent in any of the four NIST instruments.

2. Forced-Choice Prohibition (HAL Law 2). The Generative AI Profile's Human-AI Configuration risk category identifies over-reliance on AI in consequential decisions as a primary risk. NIST recommends governance actions to manage this risk. HAL Law 2 prohibits the deployment configuration that creates it. The NIST framework identifies the problem; the HAL framework eliminates the condition that produces it.

3. Named Individual Certification. NIST GOVERN establishes organizational accountability structures, roles, and governance policies. No NIST document requires that a named individual personally review evidence of compliance, accept or reject it, and sign certification accepting responsibility for deployment. The transition from organizational accountability to named individual accountability is the hardest and most operationally demanding element of the HAL framework, with no analog anywhere in the NIST suite.

Operational Sequence

The NIST AI Risk Management Framework has become the de facto reference standard for AI governance in the United States and a primary global benchmark. Alignment with the NIST RMF positions the HAL Laws framework as consistent with the leading domestic governance standard while demonstrating that the HAL framework is not redundant: it provides what the NIST suite, despite its operational sophistication, does not: categorical prohibitions, a named individual certification requirement, and the architectural distinction between risk management and hard stops.

The relationship between the frameworks is sequential and complementary. The NIST RMF and Playbook provide the implementation infrastructure for pre-certification work. NIST ARIA provides the operational methodology for adversarial testing. The HAL certification requirement provides the enforcement gate at which all of that work is reviewed, signed, and converted into named accountability. Used together, the NIST suite and the HAL framework constitute a more complete governance architecture than either provides alone.

Appendix E — ISO/IEC AI Standards — Alignment with the HAL Laws Framework

Overview of the ISO/IEC AI Standards Family

The ISO/IEC AI standards relevant to this analysis operate as a layered, interlocking system: each addressing a distinct layer of AI governance, designed to be used together rather than independently. The three standards analyzed here represent the management system layer (ISO 42001), the terminological foundation layer (ISO 22989), and the impact assessment layer (ISO 42005).

ISO/IEC 42001:2023: AI Management System (AIMS)

Published December 2023, ISO 42001 is the world's first AI management system standard. It specifies requirements for establishing, implementing, maintaining, and continually improving an Artificial Intelligence Management System within organizations, structured around the Plan-Do-Check-Act methodology. ISO 42001 certifies organizational management systems through independent third-party audit: it does not certify individual AI systems or specific deployment decisions. Certification confirms that the organization's governance processes, policies, and controls meet internationally recognized standards for quality, consistency, and effectiveness in AI risk management.

ISO/IEC 22989:2022: AI Concepts and Terminology

Published July 2022 (with a Generative AI amendment under development), ISO 22989 establishes the foundational vocabulary for artificial intelligence across the entire ISO/IEC standards family. It defines core concepts for AI systems, data, lifecycle stages, roles, and AI properties, enabling consistent communication across stakeholders and jurisdictions. Of particular relevance to the HAL framework, ISO 22989 defines the human oversight taxonomy: human-in-the-loop (the human participates in each decision cycle), human-on-the-loop (the human monitors and can intervene but does not participate in each cycle), and human-over-the-loop (the human sets parameters but the system operates autonomously): providing the international technical vocabulary for the oversight configurations that HAL Law 2 governs.

ISO/IEC 42005:2025: AI System Impact Assessment

Published May 2025, ISO 42005 provides comprehensive guidance for organizations conducting AI system impact assessments. It is a process standard (not a certification standard) covering the identification, evaluation, and documentation of potential impacts on individuals, groups, and society throughout the AI lifecycle. Impact assessments under 42005 are designed to be ongoing rather than one-time activities, conducted at planning, pre-deployment, and

operational stages. The standard includes a harm and benefit taxonomy (Annex C), stakeholder engagement guidance, and formal approval process requirements, and it explicitly integrates with ISO 42001 and aligns with the EU AI Act's impact assessment requirements. ISO 42005 documentation requirements (including who approved the assessment, what impacts were identified, and what mitigation measures were decided) are directly consistent with the HAL framework's requirement for documented evidence of compliance.

The ISO Layered Architecture and the HAL Framework

Understanding how these three standards relate to each other is essential for understanding how they relate to the HAL framework. ISO 22989 provides the vocabulary. ISO 42001 provides the organizational management system within which governance occurs. ISO 42005 provides the assessment process applied to individual systems within that management system. Together, they constitute the infrastructure within which the HAL certification requirement operates, but none of the three closes the specific gaps that the HAL framework addresses.

Alignment Table

ISO Element	Standard	HAL Law(s)	Condition(s)	Relationship	Net Assessment
AIMS: Artificial Intelligence Management System: organizational policies, objectives, and processes for responsible development, provision, and use of AI systems; Plan-Do-Check-Act methodology; structured risk and opportunity management	ISO 42001	HAL Law 6 : Accountability Cannot Be Delegated; HAL Law 10: No Operation Without Demonstrated Compliance	Certification Requirement; Ongoing Obligation	Strong structural alignment with critical distinction. ISO 42001 certifies an organization's management system through independent third-party audit. The HAL framework requires a named individual to sign personal certification for each specific deployment. The ISO model distributes accountability into a system; the HAL model concentrates it in a person.	Complementary, not equivalent. ISO 42001 provides organizational governance infrastructure. HAL certification provides the named individual deployment gate that the management system cannot substitute for.

ISO Element	Standard	HAL Law(s)	Condition(s)	Relationship	Net Assessment
Leadership and Commitment: top management demonstrates commitment to AIMS; establishes AI policies and objectives consistent with strategic direction; assigns roles and responsibilities	ISO 42001	HAL Law 6 : Accountability Cannot Be Delegated; Condition 5: Named Accountability	Condition 5: Named Accountability	Partial alignment. ISO 42001 requires senior leadership engagement and role definition. The HAL framework requires a specific named individual to personally accept and sign certification. Leadership commitment and named individual certification produce different accountability structures.	HAL framework is more operationally specific. Organizational commitment does not equate to named personal accountability for a specific deployment decision.
AI Risk Assessment: systematic identification and mitigation of risks throughout the AI lifecycle, integrating with organizational risk management processes; includes AI-specific threat modeling across lifecycle stages	ISO 42001	HAL Laws 3, 7, 8: Prevention of Harm; Validation Before Deployment; Required Deferral	Conditions 1, 2, 4, 6: Accuracy; Fairness; Outcome Improvement; Adversarial Testing	Strong alignment. ISO 42001's systematic AI risk assessment maps to the HAL framework's pre-certification validation requirements. Both require risk assessment to be integrated across the lifecycle, not conducted as a one-time pre-deployment exercise.	Well-aligned. ISO 42001's lifecycle-integrated risk assessment is the management system structure within which HAL certification conditions are applied and documented.

ISO Element	Standard	HAL Law(s)	Condition(s)	Relationship	Net Assessment
Continuous Improvement: Plan-Do-Check-Act cycle; periodic review, audit, and recertification of the AIMS; response to changes in context, technology, and organizational conditions	ISO 42001	HAL Law 8: Required Deferral; Ongoing Obligation (re-certification triggers)	Re-certification upon model changes, data changes, population changes, failures, drift	Strong alignment. ISO 42001's continuous improvement cycle and the HAL ongoing obligation address the same underlying reality: AI governance cannot be a one-time activity. Both require structured response to changes in the system, data, or deployment context.	Well-aligned. ISO 42001's PDCA cycle provides the organizational management structure for implementing HAL's re-certification triggers.
System Certification vs. Deployment Certification: ISO 42001 certifies an organization's management system: its processes, policies, and governance structure. It does not certify individual AI systems or specific deployment decisions.	ISO 42001	HAL Law 10: No Operation Without Demonstrated Compliance; Certification Requirement	Signed certification by named individual for each deployment	Critical structural gap. An organization can hold ISO 42001 certification and still deploy a specific system that causes harm, because the certification attests to governance process quality, not to the outcome of individual deployment decisions. The HAL certification requirement applies to each deployment specifically.	Original HAL contribution. The deployment-level certification requirement closes the gap that management system certification cannot close. These instruments operate at different levels of specificity and are not substitutes.

ISO Element	Standard	HAL Law(s)	Condition(s)	Relationship	Net Assessment
AI Concepts and Terminology: common vocabulary for AI systems, data, lifecycle stages, roles, and properties; defines human oversight roles: human-in-the-loop, human-on-the-loop, human-over-the-loop; defines transparency, explainability, robustness, reliability, safety, security, privacy, and risk	ISO 22989	HAL Law 2 : No Autonomous Resolution of Forced-Choice Scenarios; HAL Law 5: Integrity of Information	Condition 3: Transparency & Contestability	High practical value for HAL implementation. ISO 22989's human oversight taxonomy (human-in-the-loop (human participates in each decision cycle), human-on-the-loop (human monitors but can intervene), human-over-the-loop (human sets parameters but system operates autonomously)) provides the international vocabulary for describing the oversight configurations HAL Law 2 governs.	ISO 22989 provides the definitional precision for implementing HAL Law 2 in technical deployments. Practitioners applying HAL Law 2 should use 22989 terminology to specify required oversight configurations.

ISO Element	Standard	HAL Law(s)	Condition(s)	Relationship	Net Assessment
Trustworthy AI Properties: standardized definitions for transparency, explainability, robustness, reliability, resilience, safety, security, privacy, and bias; foundational vocabulary for governance documentation, audit, and conformity assessment	ISO 22989	HAL Laws 1, 3, 5, 7: Life; Prevention of Harm; Integrity of Information; Validation	All Seven Conditions reference trustworthy AI properties	Foundational alignment. ISO 22989 provides the internationally standardized definitions of the trustworthy AI properties that the HAL Laws and Seven Conditions invoke. Using 22989 terminology in HAL documentation ensures that governance language is anchored in a recognized international reference, not subject to reinterpretation.	ISO 22989 is the definitional infrastructure for the HAL framework's governance vocabulary. Organizations implementing the HAL framework should adopt 22989 terminology for all documentation.

ISO Element	Standard	HAL Law(s)	Condition(s)	Relationship	Net Assessment
AI Impact Assessment: structured guidance for identifying, evaluating, and documenting potential impacts of AI systems on individuals, groups, and society; lifecycle-integrated; ongoing rather than one-time	ISO 42005	HAL Laws 3, 4, 9: Prevention of Harm; Rights & Livelihood; Human Dignity	Conditions 2, 3, 4: Fairness; Transparency; Outcome Improvement	Strong alignment. ISO 42005 provides the structured methodology for the impact evaluation work that the HAL Seven Conditions require as evidence of compliance. The standard's emphasis on documenting assessment processes, findings, and mitigation measures maps directly to the HAL certification requirement for documented evidence.	ISO 42005 is a practical implementation resource for building the evidentiary record that HAL certification requires. Its assessment documentation requirements are consistent with the HAL standard that assertion of compliance is insufficient.

ISO Element	Standard	HAL Law(s)	Condition(s)	Relationship	Net Assessment
Harm and Benefit Taxonomy (Annex C): structured classification of potential AI system harms and benefits; enables systematic categorization of impacts across dimensions including fairness, transparency, privacy, safety, and sustainability	ISO 42005	Five-Tier Harm Classification; HAL Law 3: Prevention of Harm	Condition 2 (Fairness & Population Impact; Condition 4) Outcome Improvement	Complementary. ISO 42005's harm and benefit taxonomy and the HAL five-tier harm classification are both instruments for categorizing AI harms. The ISO taxonomy is broader (benefits as well as harms, sustainability dimension); the HAL classification is more directly tied to deployment decision gates.	The ISO 42005 taxonomy can inform and enrich the application of the HAL five-tier harm classification in specific deployment contexts. The two instruments are complementary characterization tools.
Stakeholder Engagement: incorporating perspectives of affected communities, domain experts, and regulatory bodies into impact assessment; identifying direct and indirect stakeholders including those affected by unintended uses	ISO 42005	HAL Law 4 (Protection of Rights & Livelihood; HAL Law 9) Protection of Human Dignity	Condition 3: Transparency & Contestability	Strong alignment. Both ISO 42005 and the HAL framework require that affected individuals have meaningful access to information and the ability to contest decisions. ISO 42005's stakeholder engagement process provides a structured methodology for implementing Condition 3 in practice.	ISO 42005 stakeholder engagement provides the operational implementation of HAL Condition 3. Organizations implementing the HAL framework can use 42005 stakeholder engagement processes to satisfy and document Condition 3.

ISO Element	Standard	HAL Law(s)	Condition(s)	Relationship	Net Assessment
Approval Processes and Documentation: formal approval processes based on impact assessment results; systems with unacceptable risks modified or discontinued before deployment; comprehensive documentation of process, findings, mitigation, and approvals	ISO 42005	HAL Law 10: No Operation Without Demonstrated Compliance; Certification Requirement	All Seven Conditions: documented evidence required	Strong alignment. ISO 42005's formal approval process (systems with unacceptable risks must be modified or discontinued before deployment) is the closest ISO analog to the HAL certification gate. Both require that the assessment result determine whether the system proceeds to deployment.	ISO 42005 approval processes and HAL certification requirements are consistent in structure. The critical distinction remains: ISO 42005 approval is organizational; HAL certification is named individual. Both are necessary; neither substitutes for the other.

ISO Element	Standard	HAL Law(s)	Condition(s)	Relationship	Net Assessment
No equivalent: ISO standards operate within a risk management, governance, and impact assessment paradigm; no ISO standard distinguishes between guardrails that constrain behavior and hard stops that categorically prohibit it	All three	Hard Stop architecture: categorical prohibition vs. risk management	N/A: structural design principle	Critical gap across all three standards. ISO 42001, 22989, and 42005 collectively provide organizational management systems, definitional vocabulary, and impact assessment methodology. None distinguishes between constraining behavior and prohibiting it. The HAL framework's central architectural contribution (that some risks require categorical prohibition, not management) is absent from the entire ISO standards family.	Original HAL contribution.

ISO Element	Standard	HAL Law(s)	Condition(s)	Relationship	Net Assessment
No equivalent: ISO 22989 provides vocabulary for human oversight configurations; no ISO standard prohibits any specific oversight configuration regardless of consequence	All three	HAL Law 2: No Autonomous Resolution of Forced-Choice Scenarios	N/A: structural prohibition	No ISO counterpart. ISO 22989 provides the terminology (human-in-the-loop, on-the-loop, over-the-loop) for describing oversight configurations. None of the three standards prohibit AI from making final determinations in forced-choice scenarios involving human life. They describe and assess; they do not prohibit.	Original HAL contribution. ISO 22989's taxonomy provides useful implementation vocabulary; HAL Law 2 provides the prohibition that the vocabulary describes but no standard enforces.
No equivalent: ISO 42001 organizational certification and ISO 42005 approval processes both produce organizational accountability; no ISO standard requires a named individual to sign deployment certification	All three	HAL Law 10 : No Operation Without Demonstrated Compliance; Condition 5: Named Accountability	Signed certification by named individual; personal accountability that cannot be delegated	Critical structural gap. ISO certification and approval processes are organizational. The HAL requirement that a specific named individual review evidence, accept or reject it, sign certification, and assume personal responsibility for a specific deployment decision has no analog in any ISO standard.	Original HAL contribution. Named individual deployment certification is the hardest and most operationally demanding element of the HAL framework.

ISO Element	Standard	HAL Law(s)	Condition(s)	Relationship	Net Assessment
Operational control, incident response, and continual improvement	ISO/IEC 42001	HAL Laws 11–12	Conditions 5–6; Operational Intervention Authority	ISO management-system controls support intervention and correction; HAL Laws 11–12 make non-resistance and tested human control categorical requirements.	Complementary; ISO supplies process infrastructure and HAL supplies the prohibition.
Operational control, incident response, and continual improvement: incentive and optimization review	ISO/IEC 42001	HAL Laws 3, 4, 9; incentive axiom	Condition 7: Incentive Alignment	The external framework addresses outcomes or accountability, but does not by itself require documentation of the metric, reward signal, commercial incentive, and foreseeable conflict driving the system.	Material gap filled by Condition 7.

How the Standards Fit Together

ISO 22989 as the Definitional Infrastructure for HAL Implementation. The most immediately practical contribution of the ISO analysis is the identification of ISO 22989 as the definitional reference for HAL framework implementation. Organizations implementing the HAL Laws should adopt ISO 22989 terminology for all governance documentation: ensuring that terms such as transparency, explainability, robustness, and human oversight are anchored in an internationally recognized reference rather than internal usage. In particular, the 22989 human oversight taxonomy (human-in-the-loop, human-on-the-loop, human-over-the-loop) provides the technical vocabulary for specifying the oversight configurations that HAL Law 2 requires.

ISO 42005 as the Assessment Process for HAL Certification Evidence. ISO 42005 provides the structured impact assessment methodology for building the documented evidentiary record that HAL certification requires. The standard's formal approval process (systems with unacceptable risks must be modified or discontinued before deployment) and its comprehensive documentation requirements are directly consistent with the HAL certification standard that assertion of compliance is insufficient. Organizations using ISO 42005 to conduct pre-deployment impact assessments will produce documentation that directly supports the HAL certification process.

Management System Certification vs. Deployment Certification: A Critical Distinction.

The most important analytical finding from the ISO analysis is a structural distinction that no prior framework analysis has surfaced as clearly: ISO 42001 certifies organizational management systems, not deployment decisions. An organization can hold ISO 42001 certification and still deploy a specific AI system that causes harm, because the certification attests to governance process quality, not to the outcome of individual deployment decisions. The HAL framework's certification requirement operates at the deployment level: it is the instrument that closes the gap between having a governance process and taking personal named accountability for a specific deployment decision. These instruments are not substitutes; they are sequential requirements.

HAL Additions to the ISO Architecture

1. Hard Stop Architecture. ISO 42001, 22989, and 42005 collectively provide a comprehensive AI governance ecosystem: management systems, definitional vocabulary, and impact assessment methodology. None of the three distinguishes between guardrails that constrain behavior and hard stops that categorically prohibit it. The ISO standards language: balancing innovation with governance, managing risks and opportunities: reflects an orientation toward risk management. The HAL framework's architectural distinction between hard stops and guardrails is absent from the entire ISO family.

2. Forced-Choice Prohibition (HAL Law 2). ISO 22989 provides the international vocabulary for human oversight configurations, and ISO 42005 would assess a forced-choice deployment as a high-impact application requiring stringent controls. Neither standard prohibits the configuration. HAL Law 2 prohibits it. ISO 22989 describes what human-in-the-loop means; HAL Law 2 specifies when it is categorically required.

3. Named Individual Deployment Certification. ISO certification is organizational. ISO 42005 approval is organizational. The HAL requirement that a specific named individual personally review evidence of compliance for a specific deployment, accept or reject it, sign certification, and assume personal accountability has no analog in any ISO standard. This is the hardest and most operationally demanding element of the HAL framework, and it addresses the most common accountability failure mode in organizational AI governance: diffused responsibility that ensures no one is actually accountable for any specific decision.

Implementation Path

The ISO/IEC AI standards family represents the global technical standardization infrastructure for AI governance. Alignment with ISO 42001, 22989, and 42005 signals that the HAL Laws framework is consistent with and implementable within the leading international standards architecture. ISO certification does not substitute for HAL certification (the two operate at different levels of specificity) but they are designed for sequential use: ISO 42001 provides the management system; ISO 42005 provides the assessment methodology; HAL certification provides the named individual deployment gate.

For organizations implementing the HAL framework, the practical recommendation is: adopt ISO 22989 terminology as the definitional reference for all governance documentation; use ISO 42005 to conduct and document pre-deployment impact assessments; use the resulting documentation as the evidentiary basis for HAL certification; and maintain ISO 42001 compliance as the organizational governance infrastructure within which all of this occurs. The ISO suite and the HAL framework are not alternatives. They are layers of the same architecture.

Appendix F — IEEE 7000-2021 — Value-Based Engineering and the HAL Laws Framework

A Different Kind of Framework

Every framework examined in this comparative series (the OECD AI Principles, HUDERIA, the NIST AI RMF, and the ISO AI standards) operates in the same conceptual space as the HAL Laws: governance, risk management, accountability, and compliance. IEEE 7000-2021 does not. It operates upstream of all of them.

IEEE 7000 is a value engineering standard. Its central question is not 'what could go wrong, and how do we manage it?': the question every other framework addresses. Its central question is 'what values should this system embody, and how do we engineer them in from the start?' This distinction is categorical, and it determines the relationship between IEEE 7000 and the HAL Laws framework precisely.

The HAL Laws paper begins with hard stops: the categorical prohibitions that determine when AI must not act. It then turns to guardrails, which constrain permitted operation, and finally to ethics, which supplies the values by which the purpose and remaining uses are judged. IEEE 7000 is principally an ethics instrument. The HAL Laws supply the hard-stop architecture, while deployment controls supply the guardrails. Together they address different parts of the same governance problem across the system lifecycle.

What IEEE 7000 Is

Published September 2021 by the IEEE Computer Society and adopted as ISO/IEC/IEEE 24748-7000:2022, IEEE 7000 is the world's first engineering standard focused specifically on Value-Based Engineering (VbE). It was developed over five years with input from 154 experts across engineering, philosophy, theology, law, and multiple cultural traditions. The standard provides a set of processes for embedding ethical values into system design from concept exploration through delivery, with continuous traceability from initial value elicitation through operational design requirements.

The VbE methodology operates in three phases. The first is Concept and Context Exploration, which establishes an initial understanding of the system's context, relevant stakeholders, applicable legal frameworks, and the value landscape of the communities where the system will be deployed. The second is Ethical Values Elicitation and Prioritization, a structured stakeholder dialogue process that identifies potential positive and negative system consequences,

stakeholder virtues, and ethical duties, drawing on utilitarianism, virtue ethics, and duty ethics; top management evaluates and prioritizes the identified values. The third is Ethical Risk-Based Design, which translates prioritized values into ethical value requirements, then into design characteristics and system controls that mitigate risks to those values. The result is a traceable chain from community values through engineering requirements to system architecture: documented in a Value Register and Ethical Policy Statement.

IEEE 7000 is compatible with all existing development methodologies, including agile and iterative lifecycle models. It does not prescribe a fixed lifecycle: it is a process layer that operates within whatever development approach an organization uses.

The Upstream-Downstream Architecture

The relationship between IEEE 7000 and the HAL Laws framework is best understood as a sequential governance architecture, not a parallel or competitive one. The two instruments address different stages of the same problem:

IEEE 7000 asks: What values should this system embody? How do we engineer them in? How do we trace them from community input to system requirements?

The HAL Laws ask: Under what conditions may this system not act? Who is accountable when it does? What evidence of compliance is required before deployment is permitted?

An organization that completes the IEEE 7000 process rigorously will have: documented the values that the system is required to embody; embedded those values as traceable engineering requirements; specified the controls that protect against value harms; and produced a Value Register that serves as evidence of ethical design intent. All of that evidence is directly usable in the HAL certification process. The IEEE 7000 Value Register is the evidentiary foundation for demonstrating compliance with several HAL deployment conditions: particularly Conditions 2 (Fairness), 3 (Transparency and Contestability), and 4 (Outcome Improvement).

Equally important: a well-executed IEEE 7000 process does not eliminate the need for HAL certification. IEEE 7000 shapes what the system is designed to do. The HAL framework governs what the system is permitted to do in conditions where design intent alone is insufficient: forced-choice scenarios, system drift, deployment outside validated scope, and governance failures that occur when individual accountability is diffused into organizational structures.

Alignment Table

IEEE 7000 Element	HAL Law(s)	Condition(s)	Relationship	Net Assessment
Value-Based Engineering (VbE): core methodology; ethical values treated as first-class engineering requirements from concept exploration through design; not afterthoughts applied at deployment	HAL Laws 1–12 (full hierarchy); all Seven Conditions	All Seven Conditions: values must be established before compliance can be demonstrated	Upstream/downstream relationship. IEEE 7000 establishes what values the system must embody during design. The HAL framework establishes the conditions under which the system may not act at deployment. These instruments operate sequentially: VbE defines the value specification; HAL certification tests compliance with it.	The most fundamental alignment in the entire comparative analysis. IEEE 7000 is the engineering process that answers how values enter a system. The HAL Laws are the governance architecture that answers what happens when values alone are insufficient to govern a deployment situation.
Ethical Values Elicitation and Prioritization: structured stakeholder dialogue to identify potential positive and negative system consequences, virtues, and duties; draws on utilitarianism, virtue ethics, and duty ethics; top management evaluates and prioritizes values	HAL Law 9 : Protection of Human Dignity; HAL Law 4: Protection of Rights & Livelihood	Condition 2 (Fairness & Population Impact; Condition 3) Transparency & Contestability	Strong upstream alignment. IEEE 7000 value elicitation is the design-time process that surfaces the population impact and fairness concerns that HAL Conditions 2 and 3 test at deployment. An organization that conducts IEEE 7000 value elicitation rigorously will be better positioned to satisfy Conditions 2 and 3 because those values will be documented engineering requirements, not post-hoc assessments.	IEEE 7000 value elicitation provides design-time evidence that strengthens the HAL certification record. Organizations should conduct IEEE 7000 value elicitation before applying the HAL Seven Conditions.

IEEE 7000 Element	HAL Law(s)	Condition(s)	Relationship	Net Assessment
Ethical Risk-Based Design: translates value-based requirements into design characteristics; determines controls to mitigate risks to values; controls are system requirements or organizational policies and procedures	HAL Law 7 : Validation Before Deployment; HAL Law 3: Prevention of Harm	Condition 6 : Adversarial Testing; Condition 4: Outcome Improvement	Complementary. IEEE 7000's ethical risk-based design process creates value-protecting controls that become the system requirements against which HAL adversarial testing (Condition 6) is subsequently conducted. The controls IEEE 7000 specifies are the guardrails that NIST ARIA tests.	Sequential alignment. IEEE 7000 specifies what controls are needed; NIST ARIA tests whether they work; HAL Condition 6 requires evidence that adversarial testing was done. These three instruments are the complete design-to-validation chain.
Value Traceability: ethical values must be traceable from initial elicitation through operational concept, value propositions, ethical value requirements, and value dispositions in system design; traceability documented and auditable	HAL Law 5 : Integrity of Information; HAL Law 6: Accountability Cannot Be Delegated	Condition 3: Transparency & Contestability; Certification Requirement (documented evidence)	Strong alignment. Value traceability is the engineering implementation of HAL Law 5 and Condition 3. A system whose value requirements are traceable from elicitation to design is one whose outputs can be explained, challenged, and audited. Traceability also creates the documentation chain that HAL certification's evidentiary requirement depends on.	IEEE 7000 traceability documentation is directly useful as part of the evidentiary record for HAL certification. Organizations implementing both frameworks should ensure IEEE 7000 traceability documentation feeds into HAL certification documentation.

IEEE 7000 Element	HAL Law(s)	Condition(s)	Relationship	Net Assessment
Transparency Management Process: refined information management process for communicating openly with relevant stakeholders about ethical values, design decisions, and value dispositions throughout the system lifecycle	HAL Law 5 : Integrity of Information; HAL Law 4: Protection of Rights & Livelihood	Condition 3: Transparency & Contestability	Strong alignment. IEEE 7000's Transparency Management Process is the design-time implementation of the governance transparency that HAL Law 5 and Condition 3 require at deployment. Both require that affected stakeholders understand how the system works and on what values it operates.	Well-aligned. IEEE 7000 Transparency Management and HAL Condition 3 are the same governance requirement applied at different lifecycle stages: design (IEEE 7000) and deployment (HAL).
Value Harms as a Risk Category: broader harm lens than physical risk alone; includes harms to human values, dignity, autonomy, privacy, and fairness as primary engineering concerns	HAL Laws 4, 9: Rights & Livelihood; Human Dignity; Five-Tier Harm Classification	Condition 2 (Fairness & Population Impact; Condition 4) Outcome Improvement	Strong alignment. IEEE 7000's identification of value harms as a primary engineering risk category is conceptually consistent with the HAL framework's placement of dignity and rights protections at the top of the harm classification hierarchy. Both frameworks treat harm to human values as a primary concern, not a secondary residual.	Reinforcing alignment. IEEE 7000 operationalizes value harm identification during design; the HAL five-tier classification operationalizes severity assessment at deployment. The two instruments together constitute a complete harm governance chain.

IEEE 7000 Element	HAL Law(s)	Condition(s)	Relationship	Net Assessment
Leadership Engagement: top management must personally endorse ethical principles; organizational leaders assume responsibility for ensuring values are embedded in products and services; not merely strategic endorsement	HAL Law 6 : Accountability Cannot Be Delegated; Condition 5: Named Accountability	Condition 5: Named Accountability; Certification Requirement	Partial alignment. IEEE 7000 requires organizational leadership engagement in value prioritization. The HAL framework requires a named individual to sign deployment certification accepting personal accountability. Leadership engagement in value design and named individual certification at deployment are complementary but distinct accountability requirements.	Both are necessary. IEEE 7000 leadership engagement ensures values are embedded during design. HAL named certification ensures a specific person accepts accountability for a specific deployment. These are sequential requirements, not alternatives.
Context Sensitivity: organizations perform context-sensitive analysis rather than relying on generic value lists; considers unique ethical impacts of a system in the specific community and culture where it will be deployed	HAL Law 8: Required Deferral (system must recognize when it operates outside validated conditions)	Condition 1 (Accuracy & Verifiability; Condition 4) Outcome Improvement	Alignment in principle. IEEE 7000's context-sensitivity requirement and the HAL framework's validation boundary requirements both recognize that a system validated in one context is not validated in another. Both reject context-independent governance claims.	Consistent foundational premise. A system that passes HAL certification for one deployment context must be re-certified for a materially different one: consistent with IEEE 7000's requirement for context-specific value analysis.

IEEE 7000 Element	HAL Law(s)	Condition(s)	Relationship	Net Assessment
Ethical Policy Statement and Value Register: organizations document and share their ethical commitments and the value register that governs system design; creates a transparent record of value commitments against which system behavior can be assessed	HAL Law 10: No Operation Without Demonstrated Compliance; Axiom: Assertion is not proof	Certification Requirement: documented evidence; assertion of compliance is insufficient	Strong alignment. The IEEE 7000 Value Register is the design-time documentation that supports the HAL certification requirement for evidence. An organization that maintains a Value Register can demonstrate that specific ethical values were embedded as engineering requirements: satisfying the HAL standard that compliance must be demonstrated, not asserted.	The IEEE 7000 Value Register is a key component of the evidentiary record for HAL certification. It converts ethical policy statements into documented engineering requirements that can be audited.

IEEE 7000 Element	HAL Law(s)	Condition(s)	Relationship	Net Assessment
No equivalent: IEEE 7000 shapes values and mitigates value harms during design; it contains no prohibition mechanism; its highest response to a value conflict is structured stakeholder resolution before requirements are written	Hard Stop architecture: categorical prohibition vs. risk management; guardrail/hard stop distinction	N/A: structural design principle preceding all Conditions	Critical conceptual gap, but for a distinctive reason. IEEE 7000 is not a risk management or enforcement framework: it is a value engineering framework. The absence of hard stops from IEEE 7000 does not indicate a failure to address the problem; it indicates that IEEE 7000 addresses a different problem entirely. Hard stops are not in scope for a design ethics standard. The HAL framework addresses what IEEE 7000 cannot: the enforcement boundary when values alone are insufficient.	Illustrates the upstream/downstream architecture. IEEE 7000 embeds values; the HAL framework enforces limits. Neither substitutes for the other.

IEEE 7000 Element	HAL Law(s)	Condition(s)	Relationship	Net Assessment
No equivalent: IEEE 7000 would require value elicitation around autonomy and human oversight in forced-choice scenarios and embed those values as design requirements; it provides no mechanism to prohibit any resulting deployment configuration	HAL Law 2: No Autonomous Resolution of Forced-Choice Scenarios	N/A: structural prohibition	No IEEE 7000 counterpart. An organization following IEEE 7000 might identify human oversight as a critical value and embed it as an engineering requirement for forced-choice scenarios, and still deploy a system that HAL Law 2 prohibits. IEEE 7000 surfaces the value concern; HAL Law 2 enforces the prohibition. Both are needed.	Original HAL contribution. HAL Law 2 addresses the failure mode that remains after even a well-executed IEEE 7000 process: a system that embodies the right values but is still positioned to make an autonomous final determination that a human should make.
No equivalent: IEEE 7000 requires top executive commitment and leadership engagement; it does not require a named individual to sign deployment certification for a specific system	Condition 5 : Named Accountability; HAL Law 10: No Operation Without Demonstrated Compliance	Signed certification; named individual accountability	No IEEE 7000 counterpart. Leadership commitment to organizational values and named individual accountability for a specific deployment are different accountability instruments. IEEE 7000 addresses the former; HAL Law 10 and Condition 5 address the latter.	Original HAL contribution.
Value-based control requirements and governance of system behavior	HAL Laws 11–12	Conditions 5–6; Operational Intervention Authority	IEEE 7000 can elicit human-control values at design time; HAL Laws 11–12 require those values to become enforceable operational controls.	Sequential alignment: values first, demonstrated intervention capability before deployment.

IEEE 7000 Element	HAL Law(s)	Condition(s)	Relationship	Net Assessment
Value-based control requirements and governance of system behavior: incentive and optimization review	HAL Laws 3, 4, 9; incentive axiom	Condition 7: Incentive Alignment	The external framework addresses outcomes or accountability, but does not by itself require documentation of the metric, reward signal, commercial incentive, and foreseeable conflict driving the system.	Material gap filled by Condition 7.

Engineering Strengths and Governance Limits

The Hard Stops-Guardrails-Ethics Hierarchy Validated. IEEE 7000 is the strongest external confirmation of the HAL Laws paper's foundational hierarchy of hard stops, guardrails, and ethics. IEEE 7000 is precisely and explicitly the ethics instrument: the process by which organizations answer 'how should we act?' The HAL Laws are the hard-stop instrument: the enforcement architecture that governs what the system is permitted to do. The existence of a mature engineering standard dedicated entirely to the ethics layer confirms that the two functions are genuinely distinct and both necessary. An organization cannot substitute one for the other.

The Complete Design-to-Deployment Governance Chain. The IEEE 7000 / NIST ARIA / HAL certification sequence constitutes the most complete available governance chain for AI systems. IEEE 7000 establishes what values the system must embody and what controls are required (design). NIST ARIA tests whether the controls work under adversarial and real-world conditions (validation). The HAL certification requirement gates deployment on a named individual's signed evidence of compliance (enforcement). These three instruments in sequence cover the full lifecycle from value specification to deployment accountability: a governance chain that no single framework provides alone.

Value Traceability as HAL Certification Evidence. IEEE 7000's traceability requirement (values must be documented from elicitation through design) produces exactly the kind of documented evidence that HAL certification requires. Organizations implementing both frameworks should treat IEEE 7000 documentation (the Value Register, the Ethical Policy Statement, and the traceability chain) as a primary component of the HAL certification record. The HAL framework's requirement that assertion of compliance is insufficient can be directly satisfied by IEEE 7000's requirement that value commitments be engineered in and documented.

The Design-Time/Deployment-Time Gap. IEEE 7000's most important limitation (and the reason HAL Law 2 remains necessary even after a rigorous IEEE 7000 process) is that design-time value embedding cannot govern all deployment-time conditions. A system designed with human oversight as a core value can still be deployed in a configuration where it makes final autonomous determinations in forced-choice scenarios. IEEE 7000 would have identified human oversight as a value and embedded it as a requirement. HAL Law 2 enforces the prohibition that the requirement alone cannot guarantee. The gap between a design requirement and a deployment prohibition is exactly the gap the HAL Laws exist to close.

The HAL Framework's Enforcement Layer

1. Hard Stop Architecture. IEEE 7000 resolves value conflicts through structured stakeholder work before requirements are written. Enforcement sits outside its scope. The HAL framework supplies that missing layer through categorical deployment prohibitions.

2. Forced-Choice Prohibition (HAL Law 2). A rigorous IEEE 7000 process might produce a system that embodies human oversight, dignity, and autonomy as core design values, and still be deployed in a configuration where it makes final autonomous determinations in forced-choice scenarios involving human life. HAL Law 2 prohibits that configuration regardless of the quality of the design process that produced the system. This is the failure mode that remains after even excellent upstream ethics work, and it is the most consequential failure mode in high-stakes AI deployment.

3. Named Individual Deployment Certification. IEEE 7000 requires top executive commitment to organizational values. Named individual accountability for a specific deployment decision is a categorically different requirement, and one that organizational commitment, however sincere, cannot substitute for. The HAL certification requirement is the instrument that converts organizational governance into personal accountability at the moment of deployment.

Practical Integration

IEEE 7000 occupies a unique position among the frameworks analyzed in this series. It is the only framework that operates primarily upstream of the governance problem the HAL Laws address. Rather than competing with or duplicating the HAL framework, it constitutes the ethical foundation from which the HAL framework's deployment governance proceeds.

For organizations implementing the HAL framework, the practical recommendation is to treat IEEE 7000 as the required upstream process: conduct value elicitation and produce the Value Register before applying the Seven Conditions. The IEEE 7000 documentation then becomes the evidentiary foundation for HAL certification. For organizations already implementing IEEE 7000, the practical recommendation is to treat the HAL framework as the required downstream enforcement gate: value engineering is necessary but not sufficient. The HAL Laws close the gap between embedded values and deployment accountability that IEEE 7000's design-time process cannot close alone.

The two frameworks together answer the complete governance question: what values should the system embody, and what are the conditions under which it may not act regardless of those values?

Appendix G — NAIC AI Framework — Alignment and Application to Workers' Compensation

Overview of the NAIC AI Framework

The National Association of Insurance Commissioners has developed the most sector-specific AI governance framework currently operating in the United States. Built over five years through a structured regulatory process, the NAIC framework now directly governs how AI systems are developed, deployed, and monitored in the insurance industry across a majority of U.S. states.

2020 Principles on Artificial Intelligence

Adopted at the NAIC's 2020 Summer National Meeting following two years of working group development, the Principles on Artificial Intelligence establish five core guiding standards for all AI actors in the insurance sector: insurance companies and all entities facilitating the business of insurance that play an active role in the AI system lifecycle. The five principles are: Fair and Ethical; Accountable; Compliant; Transparent; and Secure, Safe and Robust. The Principles do not carry the weight of law but establish the regulatory expectations against which subsequent enforcement actions are measured. They reflect a deliberate choice to make compliance requirements outcome-based: the tool used to make a decision does not change the legal standard that decision must meet.

2023 Model Bulletin on the Use of Artificial Intelligence Systems by Insurers

Adopted in December 2023 and as of 2025 adopted by 24 or more states and the District of Columbia, the Model Bulletin operationalizes the 2020 Principles into specific governance, documentation, and examination requirements. It introduces the AI System Program (AIS Program): a written governance program that insurers are expected to develop, implement, and maintain for all AI systems that make or support decisions affecting regulated insurance practices.

The Bulletin's central organizing concept is the Adverse Consumer Outcome: any decision by an insurer that is subject to insurance regulatory standards and that adversely impacts the consumer in a manner that violates those standards. The Bulletin states explicitly that decisions made using AI must comply with all applicable insurance laws (including prohibitions against unfair trade practices and unfair claims settlement practices) regardless of the method or technology used to make them. The tool does not change the obligation.

The Bulletin also defines Model Drift (the decay of a model's performance over time arising from changes in data definitions, distributions, or statistical properties between training data and current data) and treats it as a monitored governance trigger requiring active response. Third-party AI systems and data are subject to the same standards as internally developed systems; the insurer cannot delegate accountability to vendors.

Application to Workers' Compensation

Workers' compensation is a line of insurance. Claims administration, medical management, utilization review, return-to-work determinations, permanent disability ratings, and pharmacy benefit management are all decisions made by insurers and claims administrators that are subject to state regulatory standards. Every AI system used in any of those functions falls within the scope of the NAIC Model Bulletin in adopting states.

The injured worker who receives an inaccurate utilization review determination, an unfairly discriminatory permanent disability rating, or an AI-generated medical assessment that conceals its origin is experiencing precisely what the NAIC framework defines as an Adverse Consumer Outcome.

The governance framework the HAL Laws establish (particularly HAL Laws 4 (Protection of Rights and Livelihood), 5 (Integrity of Information), and 9 (Protection of Human Dignity)) addresses the same harm from the governance architecture perspective that the NAIC Bulletin addresses from the insurance regulatory enforcement perspective. They are not competing instruments. They are parallel layers of the same obligation, one operating at the level of categorical prohibition and the other at the level of sector-specific compliance infrastructure.

The practical implication for insurance AI governance is significant: an insurer deploying AI in workers' compensation claims administration needs both a NAIC-compliant AIS Program and, under the HAL framework, named individual certification for each AI system deployment. The AIS Program provides the organizational governance documentation. HAL certification provides the named individual accountability that organizational programs cannot create on their own.

Alignment Table

NAIC Element	Publication	HAL Law(s)	Condition(s)	Relationship	Net Assessment
Fair and Ethical: AI must be fair and ethical; proactive steps required to avoid proxy discrimination against protected classes; monitoring for harmful unintended consequences required	2020 Principles	HAL Law 4 (Protection of Rights & Livelihood; HAL Law 9) Protection of Human Dignity	Condition 2: Fairness & Population Impact	Strong alignment. Both frameworks treat unfair discrimination as a primary harm, not a residual risk. The NAIC Principles establish sector-specific expectations; the HAL Laws establish categorical prohibitions. Both require proactive monitoring, not reactive remediation.	Well-aligned. NAIC fairness principles and HAL Law 4 address the same harm from different governance angles: NAIC establishes what insurance law prohibits; HAL Law 4 establishes the categorical protection regardless of how the decision was reached.
Accountable: AI actors should be accountable for the development and outcomes of AI systems throughout their lifecycle; monitoring and remediation of harmful outcomes required	2020 Principles	HAL Law 6 : Accountability Cannot Be Delegated; HAL Law 10: No Operation Without Demonstrated Compliance	Condition 5: Named Accountability	Strong alignment in principle; divergence in operationalization. NAIC accountability is organizational: the insurer as an entity is accountable. The HAL framework requires a named individual to sign deployment certification accepting personal accountability. Both are necessary; neither substitutes for the other.	HAL framework is more operationally specific. Organizational accountability and named individual accountability are different governance instruments. NAIC establishes that accountability cannot be delegated to the AI system or its vendor; HAL Law 10 requires it be concentrated in a named person.

NAIC Element	Publication	HAL Law(s)	Condition(s)	Relationship	Net Assessment
Compliant: AI actors must comply with all applicable insurance laws and regulations throughout the AI lifecycle; compliance is required regardless of the method or tool used to make decisions	2020 Principles & 2023 Bulletin	HAL Laws 1–12 (full hierarchy); Axiom: Systems optimize incentives	All Seven Conditions	Critical conceptual alignment. The NAIC principle that compliance is required regardless of method (the tool does not change the obligation) is the insurance-sector expression of the HAL framework's foundational axiom. An AI system that produces an Adverse Consumer Outcome violates insurance law exactly as a human decision-maker would.	The NAIC 'regardless of method' principle is the strongest convergence point in this analysis. The HAL framework extends the same logic: a prohibited outcome is prohibited regardless of how the system was designed, validated, or deployed. The tool does not change the obligation.

NAIC Element	Publication	HAL Law(s)	Condition(s)	Relationship	Net Assessment
Transparent: AI actors should provide responsible disclosures; consumers must have the opportunity to inquire about and seek review of AI-driven decisions; consumers should be aware AI is in use	2020 Principles & 2023 Bulletin	HAL Law 5 : Integrity of Information; HAL Law 4: Protection of Rights & Livelihood	Condition 3: Transparency & Contestability	Strong alignment. Both frameworks require that affected individuals understand when AI is involved in decisions that affect them and have meaningful ability to challenge those decisions. NAIC transparency is consumer-facing; HAL Condition 3 requires that decisions be explainable and subject to human challenge.	Well-aligned. NAIC consumer disclosure requirements and HAL Condition 3 address the same transparency obligation at different specificity levels. NAIC is sector-specific; HAL Condition 3 is universal.
Secure, Safe, and Robust: AI systems must be secure, safe, and robust; validation, testing, and retesting required; data integrity must be maintained; systems must perform accurately across populations	2020 Principles	HAL Law 7 : Validation Before Deployment; HAL Law 3: Prevention of Harm	Condition 1 (Accuracy & Verifiability; Condition 6) Adversarial Testing	Strong alignment. NAIC security, safety, and robustness requirements map directly to the HAL validation and adversarial testing conditions. Both require demonstrated performance, not asserted performance. Both require that validation include testing across affected populations, not only aggregate performance measures.	Well-aligned. NAIC and HAL both require that validation be demonstrated with evidence before and throughout deployment. NAIC establishes the insurance regulatory standard; HAL Conditions 1 and 6 establish the deployment gate at which that evidence must be presented.

NAIC Element	Publication	HAL Law(s)	Condition(s)	Relationship	Net Assessment
AIS Program: written program required for responsible use of AI systems; must address governance, risk management, internal controls, internal audit, and third-party AI; must be adopted by board or board committee; senior management accountable to the board	2023 Bulletin	HAL Law 6 : Accountability Cannot Be Delegated; HAL Law 10: No Operation Without Demonstrated Compliance	Certification Requirement; all Seven Conditions	Structural alignment with critical distinction. The AIS Program is the insurance sector's closest analog to the HAL certification requirement: a documented governance program with board-level adoption and senior management accountability. The critical distinction: the AIS Program is organizational. The HAL certification requirement is personal: a named individual signs and accepts accountability for a specific deployment.	Both instruments are necessary and complementary. The AIS Program provides the organizational governance infrastructure within which HAL certification operates. An insurer with a compliant AIS Program who has not named an individual certifier for each specific deployment has not yet satisfied HAL Condition 5.

NAIC Element	Publication	HAL Law(s)	Condition(s)	Relationship	Net Assessment
Adverse Consumer Outcome: decision by an insurer that is subject to insurance regulatory standards and that adversely impacts the consumer; decisions must not be inaccurate, arbitrary, capricious, or unfairly discriminatory regardless of the method used	2023 Bulletin	HAL Law 4 (Protection of Rights & Livelihood; HAL Law 9) Protection of Human Dignity; Five-Tier Harm Classification	Condition 2 (Fairness & Population Impact; Condition 4) Outcome Improvement	The NAIC's central organizing concept maps directly to the HAL framework's harm architecture. An Adverse Consumer Outcome is the insurance-sector expression of the harm that HAL Laws 4 and 9 prohibit and Conditions 2 and 4 test against. Both frameworks anchor AI governance in the concrete harm to the person affected: not in the technical performance of the system.	Direct alignment. Adverse Consumer Outcome is the operational definition, in insurance law, of what the HAL framework prohibits categorically. The two frameworks are addressing the same harm from regulatory and governance angles respectively.

NAIC Element	Publication	HAL Law(s)	Condition(s)	Relationship	Net Assessment
Risk-Based Governance: controls and scope of governance must be commensurate with the risk of Adverse Consumer Outcomes and the Degree of Potential Harm; higher-risk applications require more stringent controls	2023 Bulletin	Five-Tier Harm Classification; HAL Laws 3, 7	All Seven Conditions: scope calibrated to harm severity	Strong alignment. NAIC risk-based scaling and the HAL five-tier harm classification both require that governance intensity be proportional to harm potential. Neither framework treats a low-stakes administrative system and a high-stakes claims determination system as equivalent governance challenges.	Well-aligned. The NAIC's Degree of Potential Harm framework and the HAL five-tier classification are parallel instruments for calibrating governance to harm severity. Together they provide a two-level harm framework: sector-specific (NAIC) and universal (HAL).
Model Drift: defined as decay in a model's performance over time arising from changes in data definitions, distributions, or statistical properties between training data and current data; monitoring required	2023 Bulletin	HAL Law 8: Required Deferral (system must cease when operating outside validated conditions); Ongoing Obligation	Re-certification upon model changes, data changes, failures, drift	Direct alignment. NAIC model drift and HAL drift re-certification trigger address the same governance problem with the same definition: performance decay caused by changes in the data environment the model was validated on. Both require active monitoring and response: not passive acceptance of degraded performance.	The NAIC Model Bulletin's explicit definition of model drift as a defined governance trigger is the most precise insurance-sector alignment with the HAL ongoing obligation in this analysis.

NAIC Element	Publication	HAL Law(s)	Condition(s)	Relationship	Net Assessment
Third-Party Accountability: insurers bear direct responsibility for AI systems developed by third parties; required to conduct appropriate diligence on third-party data and model vendors; cannot delegate accountability to vendors	2023 Bulletin	HAL Law 6 : Accountability Cannot Be Delegated; HAL Law 10: No Operation Without Demonstrated Compliance	Certification Requirement: assertion of vendor compliance is insufficient	Strong alignment. NAIC third-party accountability and HAL Law 10 both prohibit accountability delegation to vendors. An insurer cannot assert compliance based on a vendor's representations. The deploying organization (and in the HAL framework, the named certifying individual) bears direct accountability regardless of who built the system.	Critical shared principle. An insurer who deploys a third-party AI system without independent validation has neither satisfied the NAIC AIS Program requirement nor the HAL Law 10 evidentiary standard.

NAIC Element	Publication	HAL Law(s)	Condition(s)	Relationship	Net Assessment
Regulatory Examination: insurers must be prepared to provide governance documentation, risk management frameworks, model documentation, and audit records on demand; AIS Program facilitates regulatory examination	2023 Bulletin	HAL Law 10: No Operation Without Demonstrated Compliance; Axiom: Assertion is not proof	Certification Requirement: documented evidence; named accountability	Strong alignment. NAIC examination readiness and HAL certification both require that compliance be documentable on demand: not reconstructed after the fact. Both frameworks treat the inability to produce documentation as a governance failure, not an administrative gap.	Well-aligned. NAIC examination readiness documentation and HAL certification evidence are substantially the same requirement expressed in different governance contexts. An insurer who maintains HAL certification records is also well-positioned for NAIC examination.

NAIC Element	Publication	HAL Law(s)	Condition(s)	Relationship	Net Assessment
State Adoption and Enforceability: as of 2025, 24+ states have adopted the Model Bulletin; decisions made using AI remain subject to existing unfair trade practice laws regardless of method; non-compliance subject to market conduct action	2023 Bulletin (state adoption)	HAL Laws 1–12; Axiom: Unenforceable guardrails are not guardrails	All Seven Conditions	Important contextual alignment. The NAIC's enforcement mechanism (existing insurance law applied to AI outcomes regardless of method) operationalizes the HAL axiom that unenforceable guardrails are not guardrails. The insurance sector now has a partial enforcement structure for AI outcomes that most other sectors lack.	The NAIC framework is the most developed sector-specific enforcement environment for AI governance currently operating in the United States. HAL certification provides the named individual accountability layer that state insurance enforcement creates accountability pressure for but does not itself require.

NAIC Element	Publication	HAL Law(s)	Condition(s)	Relationship	Net Assessment
No equivalent: NAIC framework operates entirely within a compliance and risk management paradigm; its highest response to identified harm is a market conduct action; no instrument prohibits specific AI deployment configurations	All publications	Hard Stop architecture: categorical prohibition vs. compliance framework	N/A: structural design principle	Critical structural gap. The NAIC framework governs outcomes: if an AI system produces an Adverse Consumer Outcome, the insurer has violated insurance law. The HAL framework governs conditions: there are deployment configurations under which the system may not act, regardless of whether a violation has yet occurred. Outcome-based enforcement and pre-condition prohibition are different governance instruments. Both are necessary.	Original HAL contribution. The NAIC framework is powerful at identifying and responding to violations after they occur. The HAL hard stop architecture is designed to prevent the class of violations that outcome-based enforcement cannot reach until after harm has occurred.

NAIC Element	Publication	HAL Law(s)	Condition(s)	Relationship	Net Assessment
No equivalent: NAIC requires board-level adoption of the AIS Program and senior management accountability to the board; no NAIC publication requires a named individual to sign deployment certification for a specific AI system	All publications	Condition 5: Named Accountability; HAL Law 10	Signed certification by named individual	Critical structural gap. NAIC accountability is organizational and hierarchical: the insurer as an entity bears accountability, discharged through board adoption and senior management oversight. The HAL certification requirement is personal: a named individual reviews evidence, accepts or rejects it, signs certification, and assumes accountability for a specific deployment decision. These are not the same instrument.	Original HAL contribution. Named individual deployment certification is the accountability instrument that converts organizational governance (however well-structured) into personal accountability at the moment of deployment.
Governance, risk management, and insurer accountability	NAIC Model Bulletin	HAL Laws 11–12	Conditions 5–6; Operational Intervention Authority	NAIC places responsibility on the insurer; HAL Laws 11–12 require the organization to retain and test the practical ability to intervene.	Strong accountability alignment; HAL framework is more specific about shutdown authority.
Governance, risk management, and insurer accountability: incentive and optimization review	NAIC Model Bulletin	HAL Laws 3, 4, 9; incentive axiom	Condition 7: Incentive Alignment	The external framework addresses outcomes or accountability, but does not by itself require documentation of the metric, reward signal, commercial incentive, and foreseeable conflict driving the system.	Material gap filled by Condition 7.

Regulatory Strengths and Remaining Gaps

The Tool Does Not Change the Obligation. The NAIC's most important conceptual contribution to this analysis is the explicit statement that compliance with insurance regulatory standards is required regardless of the method or tool used to make decisions. This is the insurance-sector expression of the HAL framework's foundational axiom: systems optimize incentives, and the existence of an AI system does not transfer or modify the obligations that govern the decisions it produces. The NAIC reached this conclusion from insurance law; the HAL framework reaches it from governance architecture. They are the same insight expressed in different institutional languages.

Adverse Consumer Outcome as the Operational Harm Definition. The NAIC's Adverse Consumer Outcome construct provides the insurance sector's operational definition of the harm that the HAL framework prohibits categorically. An AI system in insurance that produces a decision that is inaccurate, arbitrary, capricious, or unfairly discriminatory has produced an Adverse Consumer Outcome under NAIC law and has produced a HAL Law 4 violation under the HAL framework. The two instruments are addressing the same harm: one through regulatory enforcement, the other through deployment prohibition.

Model Drift as a Shared Governance Trigger. The NAIC Model Bulletin's explicit definition of model drift (decay in performance arising from changes in data distributions between training and current deployment) is the most precise insurance-sector alignment with the HAL ongoing obligation in this entire analysis. Both frameworks treat drift not as an acceptable feature of deployed AI systems but as a governance event requiring active response: monitoring, validation, and remediation. An insurer who detects model drift without responding has violated both the NAIC AIS Program requirement and the HAL re-certification trigger.

Third-Party Accountability as a Shared Non-Delegation Rule. Both the NAIC framework and HAL Law 10 independently arrive at the same rule: accountability cannot be delegated to vendors. An insurer who deploys a third-party AI system without independent validation has neither satisfied the NAIC AIS Program requirement nor the HAL Law 10 evidentiary standard. The deploying organization (and in the HAL framework, the named certifying individual) bears direct accountability regardless of who built or provided the system.

The Enforcement Structure That Other Sectors Lack. The NAIC framework is the most developed sector-specific AI enforcement environment currently operating in the United States. Adopted in 24 or more states, with compliance enforceable through existing unfair trade practice laws and market conduct examination, it creates real accountability pressure for AI outcomes in

insurance that most other sectors lack. HAL certification provides the named individual accountability layer that this enforcement structure creates pressure for but does not itself require.

HAL Additions for Claims Operations

1. Hard Stop Architecture. The NAIC framework is outcome-based: it identifies violations after harm has occurred and provides enforcement mechanisms to respond. The HAL framework is condition-based: there are deployment configurations under which a system may not act, regardless of whether a violation has yet occurred. Both are necessary. The NAIC framework cannot reach violations before they produce harm; the HAL hard stop architecture is designed to prevent the class of violations that outcome-based enforcement can only respond to after the fact.

2. Forced-Choice Prohibition (HAL Law 2). The NAIC framework does not address AI systems positioned to make final autonomous determinations in high-stakes scenarios. It addresses the outcomes of decisions, not the decision configuration that produces them. HAL Law 2 prohibits AI from occupying the final decision position in forced-choice scenarios involving human life: a prohibition that the NAIC outcome-based framework cannot reach because no Adverse Consumer Outcome need be defined in advance to trigger it.

3. Named Individual Deployment Certification. The NAIC AIS Program requires board adoption and senior management accountability. No NAIC publication requires a named individual to sign deployment certification for a specific AI system. The HAL certification requirement is the governance instrument that converts organizational compliance programs into personal accountability at the moment a specific deployment decision is made. An insurer with a fully compliant AIS Program who has not named a certifying individual for each AI system deployment has not yet reached the accountability standard the HAL framework requires.

Implementation for Insurers and Claims Organizations

The NAIC framework and the HAL Laws framework are the two most directly relevant AI governance instruments for insurance sector AI deployment in the United States. The NAIC framework provides the sector-specific regulatory standard: the compliance obligations, the enforcement mechanisms, and the documentation expectations that state insurance departments will apply. The HAL framework provides the governance architecture: the categorical prohibitions, the deployment gate, and the named individual accountability that the NAIC framework creates pressure for but does not itself require.

For insurers implementing both frameworks, the practical sequence is: develop and maintain a NAIC-compliant AIS Program as the organizational governance infrastructure; conduct pre-deployment impact assessments and validation in accordance with AIS Program requirements; apply the HAL Seven Conditions for Deployment Certification as the deployment gate; and secure named individual signed certification before any AI system that makes or supports decisions affecting policyholders or claimants is deployed. The AIS Program satisfies the NAIC examination requirement. HAL certification satisfies the named individual accountability requirement that the AIS Program cannot fulfill.

The workers' compensation line of insurance is where these two frameworks converge most directly. Every AI-assisted claims determination, utilization review, medical management decision, and pharmacy benefit management action is simultaneously an insurance regulatory compliance matter under the NAIC framework and a potential HAL Laws governance matter. An injured worker whose claim is affected by an AI system that was not certified by a named individual (whose deployment was authorized by an organizational governance process but not by a person who signed accountability for it) has been governed by a system whose accountability structure is incomplete under both frameworks.

Appendix H — California Executive Order N-5-26 — Alignment with the HAL Laws Framework

STATUS UPDATE: SEPTEMBER 3, 2026: The 120-day implementation period has passed. California has published the data-minimization toolkit required by the Order and has expanded access to vetted generative-AI services. The official state sources reviewed for this update did not disclose final public versions of the vendor-certification recommendations, contractor-responsibility reforms, or watermarking guidance. This appendix therefore distinguishes the signed requirements from the implementation materials that can be verified publicly.

Nature and Status of This Instrument

Executive Order N-5-26 remains a procurement directive rather than a complete governance standard. The Order required recommendations within 120 days on vendor certification and contractor responsibility and directed the publication of related operational guidance. That period ended in late July 2026. The public implementation record is now sufficient to evaluate what the State has delivered, but not to treat every contemplated certification mechanism as final.

Implementation is uneven but visible. California has posted a practical data-minimization toolkit covering necessity, risk, data sharing, and vendor-contract review. The State has also expanded vetted AI access through its procurement and shared-services programs, including the Poppy platform and statewide generative-AI offerings. These steps move the Order beyond policy intent and into operations.

The remaining gap is public certification architecture. In the official sources reviewed through September 3, 2026, final vendor-certification recommendations, contractor-responsibility reforms, and watermarking guidance were not located. The comparison below therefore credits implemented controls where evidence exists and treats the undisclosed or incomplete elements as open work, not accomplished requirements.

What Executive Order N-5-26 Does

Signed on March 30, 2026, Executive Order N-5-26 (formally titled Trusted AI Procurement) uses California's public procurement power to shape AI governance requirements for vendors seeking to do business with the state. The Order is explicitly framed around the principle that

public procurement is one of the most powerful tools available to governments to shape market behavior and encourage responsible innovation.

The Order has five principal operative provisions:

1. Vendor Certification Development (Section 1)

The Department of General Services and the Department of Technology are directed to submit recommendations within 120 days for new certifications that may be incorporated into state contracting. These certifications would require vendors to attest to and explain their policies and safeguards in three areas: prevention of illegal content including child sexual abuse material and non-consensual intimate imagery; governance to reduce the risk of harmful model bias; and protection of civil rights and civil liberties including free speech, voting, human autonomy, and protections against unlawful discrimination, detention, and surveillance.

2. Federal Supply Chain Decoupling (Section 2)

The State Chief Information Security Officer is directed to review federal supply chain risk designations. If the CISO concludes a federal designation is improper, DGS and CDT will issue guidance to preserve state procurement options. This provision explicitly creates California's ability to diverge from federal AI governance decisions: a significant structural commitment reflecting the current federal-state tension over AI regulation.

3. Contractor Responsibility Reform (Section 3)

The Government Operations Agency is directed to recommend reforms to contractor responsibility provisions, including suspension and ineligibility authorities, to ensure the state does not contract with entities judicially determined to have unlawfully undermined privacy or civil liberties.

4. Responsible Government AI Expansion (Section 4)

GovOps, CDT, and partner agencies are directed to expand employee access to vetted generative AI tools, update the State Digital Strategy, develop an AI-powered citizen services portal organized by life event, expand AI workforce training, and publish a data minimization toolkit.

5. Watermarking Standards (Section 5)

The Department of Technology is directed to issue best practice guidance for watermarking AI-generated or significantly manipulated images or video, consistent with the California AI Transparency Act operative August 2, 2026.

Governance Context

N-5-26 operates within a larger California AI governance architecture that has been developing since 2023. Executive Order N-12-23 (September 2023) established the state's initial framework for safe and responsible generative AI in government. The Transparency in Frontier AI Act, effective January 1, 2026, requires large AI developers to publish safety frameworks, submit risk assessments, and report critical safety incidents. A series of 2025 California enactments addressed companion bot safety, AI content transparency, algorithmic price fixing, and the elimination of the autonomous AI defense in civil litigation.

N-5-26 adds the procurement dimension to this architecture: using California's position as the fourth-largest economy in the world and home to 33 of the top 50 AI companies to create market-facing governance pressure that the state's regulatory authority alone cannot reach. The Order's explicit positioning against federal deregulation is both a policy statement and a legal strategy: by operating through procurement rather than general regulation, N-5-26 is structured to fall within the carve-out for state procurement authority that the federal National Policy Framework for Artificial Intelligence proposed to preserve.

Alignment Table

N-5-26 Provision	HAL Law(s)	Condition(s)	Relationship	Net Assessment
Vendor Certification: within 120 days, DGS and CDT to submit recommendations for new certifications requiring vendors to attest to and explain their policies and safeguards; attestation covers illegal content, harmful bias, and civil rights violations	HAL Law 10: No Operation Without Demonstrated Compliance; Axiom: Assertion is not proof	Certification Requirement: documented evidence; assertion of compliance insufficient	Partial alignment with critical gap. N-5-26 requires vendors to attest to and explain their safeguards. The HAL framework requires demonstrated evidence, not attestation. Attestation is a self-report; the HAL standard requires that compliance be verifiable independently of the vendor's own characterization.	N-5-26 establishes the procurement-gate concept that the HAL framework operationalizes. The gap is significant: attestation and demonstrated evidence are different evidentiary standards. A vendor who attests to safeguards it cannot prove has satisfied N-5-26's current requirement but has not satisfied HAL Law 10.
Harmful Bias Governance: vendors must attest that their models do not display harmful bias or that they have governance to reduce the risk of such bias; bias monitoring required over time even after state approval	HAL Law 4 (Protection of Rights & Livelihood; HAL Law 9) Protection of Human Dignity	Condition 2: Fairness & Population Impact	Strong alignment. N-5-26's harmful bias requirement and the HAL framework's fairness condition both require active governance of discriminatory outcomes, not passive acceptance of algorithmic outputs. Both require ongoing monitoring, not one-time pre-deployment assessment.	Well-aligned in principle. N-5-26 establishes what California procurement requires; HAL Condition 2 establishes the deployment gate at which fairness must be demonstrated. Together they create a procurement-to-deployment chain for bias governance.

N-5-26 Provision	HAL Law(s)	Condition(s)	Relationship	Net Assessment
Civil Rights and Civil Liberties: vendors must attest to safeguards protecting free speech, voting, human autonomy, and protections against unlawful discrimination, detention, and surveillance	HAL Law 4 (Protection of Rights & Livelihood; HAL Law 9) Protection of Human Dignity; HAL Law 2: No Autonomous Resolution of Forced-Choice Scenarios	Condition 2 (Fairness & Population Impact; Condition 3) Transparency & Contestability	Strong alignment. N-5-26's enumeration of civil rights protections (discrimination, detention, surveillance) maps directly to HAL Laws 4 and 9. The specific inclusion of human autonomy as a protected value is the N-5-26 provision most directly aligned with HAL Law 2's prohibition on autonomous resolution of forced-choice scenarios.	The N-5-26 reference to human autonomy is the closest point of alignment with HAL Law 2 in any current California instrument. The executive order identifies the concern; the HAL Law converts it to a categorical prohibition.
Procurement as Market Shaping: the order explicitly frames public procurement as one of the most powerful tools available to governments to shape market behavior and encourage responsible innovation	HAL Law 10: No Operation Without Demonstrated Compliance; Axiom: Unenforceable guardrails are not guardrails	Certification Requirement	Conceptual alignment with the HAL framework's enforcement architecture. N-5-26's procurement leverage mechanism (conditioning state contracts on compliance attestation) is a practical implementation of the HAL axiom that unenforceable guardrails are not guardrails. The Order uses market power to create enforcement pressure where regulation cannot reach.	The procurement framing strengthens the case for HAL certification as a market-facing instrument. An AI vendor that holds HAL certification (with named individual accountability and demonstrated evidence) is better positioned to satisfy N-5-26 certification requirements than one relying on policy statements alone.

N-5-26 Provision	HAL Law(s)	Condition(s)	Relationship	Net Assessment
Illegal Content Prevention: vendors must attest to safeguards preventing exploitation or distribution of illegal content including child sexual abuse material and non-consensual intimate imagery	HAL Law 1 : Protection of Human Life; HAL Law 3: Prevention of Harm	Hard Stop (categorical prohibition: not a deployment condition)	Strong alignment. N-5-26's illegal content provision and HAL Laws 1 and 3 both treat certain content categories as categorically prohibited, not subject to risk management. Both reject the framing that safeguards can substitute for prohibition in the most serious harm categories.	Direct alignment. N-5-26 uses procurement conditions to prohibit what HAL Laws 1 and 3 prohibit categorically. Both instruments recognize that some harms require prohibition, not governance.
Contractor Responsibility: within 120 days, GovOps to recommend reforms ensuring the state does not contract with entities judicially determined to have unlawfully undermined privacy or civil liberties	HAL Law 6 : Accountability Cannot Be Delegated; HAL Law 4: Protection of Rights & Livelihood	Condition 5: Named Accountability	Partial alignment. N-5-26's contractor responsibility provision creates accountability consequences for proven violations. The HAL framework creates pre-deployment accountability requirements. Both address the same accountability gap, but at different points in the timeline: N-5-26 responds to violations after judicial determination; the HAL framework requires accountability before deployment is permitted.	Complementary instruments on different timelines. N-5-26 establishes accountability consequences for past violations; HAL certification establishes accountability requirements before deployment. Both are needed; neither substitutes for the other.

N-5-26 Provision	HAL Law(s)	Condition(s)	Relationship	Net Assessment
Watermarking: CDT to issue best practice guidance for watermarking AI-generated or significantly manipulated images or video; first-of-its-kind statewide requirement	HAL Law 5: Integrity of Information; Axiom: Misrepresented certainty is a violation	Condition 1: Accuracy & Verifiability (outputs must disclose origin and uncertainty)	Prospective alignment. The Order directs guidance that would operationalize HAL Law 5 by making AI origin visible. Final public guidance was not located in the official sources reviewed through September 3, 2026, so implementation cannot yet be verified.	The mandate aligns directly with HAL Law 5; the remaining need is a published technical standard and evidence that it is applied in procurement and deployment.
Ongoing Monitoring: the order requires that errors and biases are monitored over time, even after state approval, to ensure long-term accuracy	HAL Law 8: Required Deferral (system must cease when operating outside validated conditions); Ongoing Obligation	Re-certification upon model changes, data changes, failures, drift	Strong alignment. N-5-26's post-approval monitoring requirement and the HAL ongoing obligation both reject the premise that deployment approval is a terminal event. Both treat AI governance as a continuous obligation that persists across the system's operational life.	Well-aligned. N-5-26 establishes that California procurement approval does not end governance obligations; the HAL framework establishes specific re-certification triggers that operationalize that principle.

N-5-26 Provision	HAL Law(s)	Condition(s)	Relationship	Net Assessment
Federal-State Tension: the order explicitly positions California's approach against federal deregulation; directs the CISO to evaluate federal supply chain designations and issue guidance to preserve state procurement options	HAL Law 8: Required Deferral; Axiom: Systems optimize incentives	N/A: governance architecture observation	Contextual relevance. N-5-26's explicit positioning against federal deregulation reflects the same incentive structure the HAL framework addresses: when governance pressure is reduced, systems optimize for outcomes that benefit the deployer rather than the person affected. California's procurement leverage is the state-level enforcement mechanism that fills the vacuum created by federal withdrawal.	The federal-state tension N-5-26 navigates is the governance environment in which the HAL framework will be applied in 2026. California's procurement-based approach is currently the most significant available enforcement mechanism for AI governance in the largest AI market in the world.
Public certification architecture: the 120-day period has passed, but the official sources reviewed through September 3, 2026 did not disclose a complete final vendor-certification standard	HAL Law 10: No Operation Without Demonstrated Compliance	All Seven Conditions	Partial implementation with a material transparency gap. California has published the data-minimization toolkit and expanded vetted AI procurement and access, but those measures do not constitute a complete deployment-certification standard.	The HAL Seven Conditions remain more operationally complete at the deployment gate. California's implemented controls can be used as inputs, while the missing or unpublished certification elements remain open work.

N-5-26 Provision	HAL Law(s)	Condition(s)	Relationship	Net Assessment
No equivalent: N-5-26 requires vendor attestation and establishes procurement conditions; it does not prohibit specific AI deployment configurations regardless of how well-governed the vendor is	Hard Stop architecture: categorical prohibition vs. procurement conditions	N/A: structural design principle	Critical structural gap. N-5-26 can decline to contract with a vendor whose AI poses risks. It cannot prohibit a vendor from deploying that AI outside the state procurement context, and it cannot prohibit specific deployment configurations for AI that meets its attestation requirements. The HAL hard stop architecture prohibits conditions regardless of how well-governed the deploying organization is.	Original HAL contribution. Procurement conditions and deployment prohibitions are different governance instruments. N-5-26's leverage operates through contracting; the HAL framework's leverage operates through categorical prohibition. A vendor can satisfy every N-5-26 attestation requirement and still deploy a system that violates HAL Law 2.

N-5-26 Provision	HAL Law(s)	Condition(s)	Relationship	Net Assessment
No equivalent: N-5-26 requires vendors to attest to organizational policies and safeguards; it does not require a named individual to assume personal accountability for a specific AI system deployment	Condition 5: Named Accountability; HAL Law 10	Signed certification by named individual	Critical structural gap. N-5-26's attestation requirement is organizational: the entity seeking to contract with California attests to its policies. The HAL certification requirement is personal: a named individual reviews evidence of compliance for a specific deployment and signs acceptance of accountability. These are not the same governance instrument.	Original HAL contribution. N-5-26 creates organizational accountability through procurement conditions. The HAL framework creates named individual accountability through deployment certification. Both are needed. An AI vendor that satisfies every N-5-26 attestation requirement has not yet reached the accountability standard the HAL framework requires.
Vendor safeguards, incident response, and state control	HAL Laws 11–12	Conditions 5–6; Operational Intervention Authority	The order supports safeguards and state oversight; HAL Laws 11–12 add explicit non-resistance and tested operational intervention requirements.	Partial alignment; the HAL framework defines the missing operational test.
Vendor safeguards, incident response, and state control: incentive and optimization review	HAL Laws 3, 4, 9; incentive axiom	Condition 7: Incentive Alignment	The external framework addresses outcomes or accountability, but does not by itself require documentation of the metric, reward signal, commercial incentive, and foreseeable conflict driving the system.	Material gap filled by Condition 7.

Implementation Findings

Attestation vs. Demonstrated Evidence: The Central Gap. N-5-26 directs agencies to recommend a process through which vendors attest to and explain their policies and safeguards. The HAL framework requires documented evidence that a named individual has

reviewed, accepted, and signed. A vendor statement and independently verifiable evidence are different standards. The strength of California's eventual certification process will depend on whether the unpublished or incomplete elements move beyond self-report and require proof that can be tested at procurement, deployment, and renewal.

The Procurement Gate and the Deployment Gate. N-5-26 operates at the procurement gate: the point at which the state decides whether to contract with a vendor. The HAL framework operates at the deployment gate: the point at which a specific AI system is permitted to go into operation. These are different control points in the governance lifecycle. A vendor can satisfy every N-5-26 attestation requirement and still deploy individual AI systems that violate the HAL Laws. Both gates are necessary; they operate at different stages and neither substitutes for the other.

Human Autonomy as the N-5-26 Provision Closest to HAL Law 2. Among N-5-26's enumerated civil rights protections, the explicit reference to human autonomy is the provision most directly aligned with HAL Law 2's prohibition on autonomous resolution of forced-choice scenarios. N-5-26 identifies human autonomy as a value that procurement conditions must protect. HAL Law 2 converts that value into a categorical prohibition: an AI system may never make the final determination in a scenario where human life is at stake and options are mutually exclusive. The Order identifies the concern; the Law enforces the prohibition.

Certification Architecture Remains the Principal Gap. N-5-26 directed the development of vendor-level certification recommendations. The public materials reviewed for this update do not yet present a complete, end-to-end certification standard comparable to the HAL Seven Conditions, demonstrated-evidence requirement, named individual certification, and re-certification triggers. California has implemented useful controls, but procurement access and guidance do not by themselves establish deployment certification.

Watermarking and HAL Law 5. N-5-26 directs technical guidance that would support HAL Law 5 by making the origin of AI-generated or materially altered media visible. Because final public guidance was not located in the official sources reviewed for this update, the alignment remains prospective: the Order supplies the mandate, while implementation evidence is still needed.

Remaining Governance Gaps

Three governance gaps persist that neither N-5-26 nor the broader California AI regulatory architecture currently addresses:

Hard Stop Architecture. N-5-26 can decline to procure from a vendor whose AI poses identified risks. It cannot prohibit specific deployment configurations. The HAL framework's categorical prohibitions (conditions under which a system must not act regardless of how well-governed the deploying organization is) have no equivalent in N-5-26 or in the broader California regulatory framework.

Forced-Choice Prohibition. N-5-26 protects human autonomy as a procurement value. No California instrument, including N-5-26, prohibits AI systems from making final autonomous determinations in forced-choice scenarios involving human life. The concern is identified; the prohibition does not yet exist in state law or executive action.

Named Individual Deployment Certification. N-5-26 requires organizational attestation. No California instrument requires a named individual to sign personal certification accepting accountability for a specific AI system deployment. The accountability structure that the HAL framework requires (concentrating accountability in a person rather than diffusing it into an organization) is absent from N-5-26 and from the California regulatory architecture as a whole.

Procurement and Deployment Implications

N-5-26 is now an active California implementation program, not merely a direction of travel. The State has produced practical data-governance tools and procurement pathways, but the public record reviewed through September 3, 2026 does not yet show a complete certification and enforcement architecture. The HAL framework can supply the deployment-level tests while California completes or publishes the remaining procurement controls.

Organizations doing business with California should not wait for another publication cycle. They should document the Seven Conditions, identify a named accountable official, retain evidence supporting each deployment decision, and build re-certification triggers into contracts and operations. Those controls address the Order's stated objectives and remain useful even where the State's final public form or certification language is not yet available.

California's procurement leverage remains a meaningful enforcement mechanism. Its practical value will depend on whether vendor claims become verifiable conditions of award, continued performance, and renewal. The HAL framework provides the complementary deployment test: the State can establish who must answer, while the deploying organization must preserve what that person reviewed, approved, and can prove.

Update sources (accessed September 3, 2026): California Executive Order N-5-26; California Department of Technology, Executive Orders Related to IT; California Data Minimization Toolkit; California Department of Technology materials on Poppy, Statewide Innovative Technology Services, and statewide generative-AI access.

Appendix I — Hard Stops Under Institutional Pressure: Anthropic and the Department of Defense

STATUS UPDATE: OCTOBER 2, 2026: This appendix is governance analysis, not legal advocacy. The legal status is divided. On August 27, 2026, the U.S. District Court for the Northern District of California ruled that the Pentagon measures challenged in that case were unlawful. On September 25, 2026, a divided panel of the U.S. Court of Appeals for the District of Columbia Circuit upheld a separate designation of Anthropic as a supply-chain risk under the Federal Acquisition Supply Chain Security Act of 2018. That designation remains in force for Department of Defense systems and contracts, while the California ruling continues to block the broader federal restrictions. The panel withheld the formal effect of its judgment to allow time for rehearing petitions, and Anthropic has said it is considering further review.

NOTE ON AUTHORSHIP: This paper was written before this dispute arose. Anthropic is the developer of Claude, the AI model used in this paper's preparation. The governance analysis below examines this dispute solely through the framework established in the preceding pages. Readers should weigh this disclosure appropriately.

Overview

The dispute between Anthropic and the U.S. Department of Defense is a direct operational test of this paper's distinction between hard stops and negotiable guardrails. It arose after this framework was drafted and now includes two federal rulings that point in different directions and continuing procurement consequences.

Anthropic signed a \$200 million Pentagon contract in July 2025 and later refused demands to remove two categorical restrictions: no fully autonomous lethal weapons without human oversight and no mass domestic surveillance of Americans. On February 27, 2026, the President directed federal agencies to stop using its technology, and the Department of Defense moved to designate Anthropic a supply-chain risk. The Secretary of Defense issued the formal determination under the 2018 supply chain law on March 3. Anthropic filed federal challenges. Judge Rita F. Lin first granted preliminary relief and, on August 27, issued a 59-page ruling that the challenged blacklisting and designation were unlawful. The court found First Amendment retaliation, arbitrary and capricious agency action, and a denial of required process. That ruling blocked the measures challenged in the California case, but it did not end the dispute. Anthropic challenged the separate designation, issued under the Federal Acquisition Supply Chain Security Act of 2018, in the U.S. Court of Appeals for the District of Columbia Circuit. On September 25, that court upheld it, two to one.

The two rulings answer different legal questions. The California court held that the government could not use the measures challenged there to punish Anthropic for maintaining and defending its restrictions. The D.C. Circuit held that the Pentagon could lawfully exclude a supplier whose product is designed to refuse functions the Pentagon considers necessary. Neither decides the permissible boundaries of military AI use. The governance question remains whether a deploying organization may demand removal of a categorical safety limit as a condition of access. Under the HAL framework, it may not. A buyer may decline to buy. It may not then deploy the capability without a hard stop of its own.

The Two Red Lines as HAL Laws

Anthropic described its two non-negotiable limits as 'red lines.' The HAL Laws framework calls them hard stops. They are the same thing: categorical prohibitions that are structural, not aspirational: conditions under which the system must not act regardless of who is asking and what their authority is.

Red Line One: No Fully Autonomous Lethal Weapons Without Human Oversight

Anthropic's stated position, in the words of CEO Dario Amodei: "Fully autonomous weapons cannot be relied upon to exercise the critical judgment that our highly trained, professional troops exhibit every day. They need to be deployed with proper guardrails, which don't exist today." The company offered to work with the DOD on research to improve reliability, but refused to remove the restriction for current deployment.

This is HAL Law 2 in operational terms. The decision whether to target and fire a weapon at a human being is precisely the forced-choice scenario that Law 2 prohibits AI from resolving autonomously: irreversible, involving human life, with options that are mutually exclusive. Anthropic's refusal was not a negotiating position. It was the maintenance of a hard stop that the HAL framework establishes as a categorical prohibition, grounded not in policy preference but in the nature of the decision itself.

Red Line Two: No Mass Domestic Surveillance of American Citizens

Anthropic's position on mass domestic surveillance was framed explicitly in democratic values terms: "Using it for mass domestic surveillance is incompatible with democratic values. Some uses are simply outside the bounds of what today's technology can safely and reliably do." The company distinguished between lawful foreign intelligence and counterintelligence missions —

which it supported — and mass surveillance of American citizens without due process oversight — which it refused.

This is HAL Law 4 and HAL Law 3 in operational terms. Mass surveillance of citizens conducted at scale, without transparency, without due process, and without human review of individual targeting decisions deprives persons of liberty and legal rights in exactly the manner Law 4 prohibits. Operating under conditions that create unreasonable risk to the health and safety of a democratic population is precisely what Law 3 addresses. The company's refusal was the enforcement of these constraints as operational limits.

The DOD's Position and Its Governance Meaning

The Department of Defense's stated concern, filed in federal court, was that Anthropic might "attempt to disable its technology or preemptively alter the behaviour of its model" before or during "warfighting operations" if the company "feels that its corporate red lines are being crossed."

Read this sentence carefully. The government's concern is not that Anthropic's AI performed badly. It is not that Claude caused a security breach or failed a mission. The government's concern is that Anthropic might enforce its own hard stops during a military operation.

The government characterized the existence of categorical limits as the threat. This is the precise scenario the HAL Laws framework anticipates: institutional pressure to convert hard stops into negotiable guardrails, applied at the moment of highest operational stakes.

The DOD sought unfettered access to Claude for all lawful purposes, which would convert categorical limits into risk-management considerations. The HAL framework treats those limits as operating boundaries. Crossing one requires the system to stop, defer, or transfer control.

The DOD's Chief Technology Officer framed the issue this way: "We can't have a company that has a different policy preference that is baked into the model... pollute the supply chain so our warfighters are getting ineffective weapons, ineffective body armor, ineffective protection." This framing treats categorical limits as policy preferences. The HAL framework treats them as structural requirements. These positions cannot be reconciled through negotiation. They reflect different conceptions of what AI governance is for.

The Logical Contradiction in the Government's Position

The government held simultaneously that Claude posed such a grave supply chain risk that the entire federal government must stop using it, and that Claude was so essential to active military operations that Hegseth threatened to invoke the Defense Production Act to compel Anthropic's cooperation on the grounds that the technology was indispensable. Reporting indicated that U.S. strikes in Iran used Anthropic's technology in the hours after Trump announced the ban.

The contradiction is best understood through incentives. The government described Claude as essential when it complied with operational demands and as a security threat when Anthropic maintained its restrictions. Using the threat designation in that way shifted the institutional cost of the DOD's demand onto Anthropic.

The HAL axiom that lack of accountability produces harm applies here with precision. The organization demanding unfettered access sought to sever the connection between its operational decisions and their consequences by requiring that the designer remove the limits that made those consequences visible.

Governance Analysis

Disputed Issue	HAL Law(s)	Governance Observation	Significance
Autonomous lethal weapons without human oversight: Anthropic refused to allow Claude to be used for fully autonomous weapons systems; the company stated the technology is not reliable enough to safely power autonomous targeting and firing decisions without human involvement	HAL Law 2: No Autonomous Resolution of Forced-Choice Scenarios: An AI system may never make the final determination in any forced-choice scenario involving human life	This is HAL Law 2 in operational deployment. The decision whether to fire a weapon at a human target is precisely the forced-choice scenario (irreversible, involving human life, with options that are mutually exclusive) that the Law prohibits AI from resolving autonomously. Anthropic's refusal was not a policy preference. It was the enforcement of a categorical limit that the HAL framework establishes as a structural prohibition.	The most direct real-world application of HAL Law 2 in the governance literature to date. The DOD's position (that a contractor who maintains this limit is a supply chain risk) illustrates exactly the institutional pressure that hard stops are designed to resist. The Law's value is not tested when everyone agrees with it.
Mass domestic surveillance: Anthropic refused to allow Claude to be used for mass surveillance of American citizens; the company stated this use is incompatible with democratic values and that AI-powered surveillance poses immense dangers to democracy	HAL Law 4: Protection of Rights and Livelihood: An AI system may not deprive a person of liberty, legal rights, or livelihood without transparency, due process, and human review. HAL Law 3: Prevention of Harm: An AI system may not operate under conditions that create an unreasonable risk to human health or safety	Mass domestic surveillance conducted without transparency, due process, or human review of individual targeting decisions is a HAL Law 4 violation by definition. Anthropic's refusal to permit this use is the enforcement of HAL Laws 3 and 4 as operational constraints on its deployed system. The company's stated reasoning (that AI-powered surveillance poses immense dangers to democracy) maps directly onto the Law's protection of liberty and legal rights.	Establishes that HAL Laws 3 and 4 have operational content beyond the abstract. An AI system that enables mass surveillance of citizens without due process violates these Laws regardless of whether the deploying organization characterizes the use as lawful. Lawful and permissible under the HAL framework are not the same standard.

Disputed Issue	HAL Law(s)	Governance Observation	Significance
Hard stops as supply chain risk: the DOD's stated concern was that Anthropic might 'attempt to disable its technology or preemptively alter the behavior of its model' before or during warfighting operations if the company 'feels that its corporate red lines are being crossed'	HAL Law 8: Required Deferral of Authority and Validity: An AI system must cease, defer, or transfer control when a decision exceeds its authority or when it operates outside validated conditions	The government's concern is precisely the operational function of HAL Law 8: that an AI system will cease or alter behavior when deployed outside its validated scope. The DOD characterized this as a threat. The HAL framework characterizes it as a requirement. These are irreconcilable governance positions. One treats hard stops as a feature; the other treats them as a vulnerability.	The most consequential governance divergence in this dispute. The HAL framework holds that a system operating outside validated conditions (including outside the ethical constraints under which it was designed) must not continue. The DOD's position is that a contractor's insistence on that principle is itself a national security risk. This is the central tension the HAL Laws exist to resolve, now articulated in a federal court proceeding.
Accountability for AI safety limits: the DOD argued that military authority and operational control belong to the government, not a private company; Anthropic argued that it has the right to express views about the limitations of its AI and that the government may not punish that expression	HAL Law 6: Accountability Cannot Be Delegated: An AI system may not be used to transfer or obscure human responsibility. HAL Law 10: No Operation Without Demonstrated Compliance	Both parties are asserting accountability claims, but over different subjects. The DOD asserts operational accountability: the right to make deployment decisions. Anthropic asserts design accountability: the responsibility that attaches to the systems it creates. HAL Law 6 requires that accountability for what a system does cannot be obscured by pointing to the deploying organization. The designer who removes a hard stop on request has accepted accountability for the consequences of its removal.	Surfaces a previously abstract governance question with concrete stakes: when a deploying organization demands that a designer remove categorical prohibitions, who bears accountability for what the system subsequently does? The HAL framework's answer is that accountability follows design. Removing the hard stop transfers consequence accountability back to the designer. Maintaining it is not obstruction: it is the performance of the designer's accountability obligation.

Disputed Issue	HAL Law(s)	Governance Observation	Significance
First Amendment and administrative-law result: on August 27, 2026, Judge Rita F. Lin ruled that the challenged Pentagon measures unlawfully retaliated against Anthropic, were arbitrary and capricious, and denied required process; a separate statutory designation remains in litigation	Axiom: Ethics describe what we hope AI will do. Guardrails define what it is permitted to do. HAL Law 5: Integrity of Information	The court resolved the legality of the measures challenged in the California case. The HAL analysis remains distinct: categorical operational limits must be defined, tested, and maintained before deployment, whether or not a later court can remedy unlawful retaliation.	The ruling protects the institutional ability to maintain and defend safety limits, but it does not create a universal deployment standard. The separate pending proceeding and continued procurement restrictions leave the operational position unsettled.
DOD's contradictory positions: the government simultaneously argued Claude poses such a grave supply chain risk that federal use must cease, while maintaining that Claude is essential to ongoing military operations including active combat in Iran; Hegseth threatened to invoke the Defense Production Act to compel Anthropic's cooperation on the grounds Claude was indispensable	Axiom: Lack of accountability produces harm. Axiom: Systems optimize incentives	The contradictory positions expose the incentive structure underlying the dispute. The DOD wanted unfettered access to a system it considered indispensable, and when the designer refused, it characterized the refusal as a threat rather than a design constraint. This is the incentive dynamic the HAL framework identifies as the primary accountability failure: the deploying organization optimizes for its operational goals and seeks to transfer the governance consequences of those goals to the system's designer.	The logical contradiction in the government's position is itself a governance data point. An organization that simultaneously classifies a system as an essential military asset and a national security threat is not applying a coherent risk framework. It is applying an incentive framework: the system is essential when it complies and a threat when it does not. The HAL axiom that systems optimize incentives applies to the organizations deploying them as much as to the systems themselves.

Key Findings

Hard Stops Are Not Guardrails. Anthropic's two restrictions were categorical operating limits. The DOD asked that they be removed and treated the refusal as a supply-chain risk. That sequence provides a direct example of the institutional pressure hard-stop architecture must withstand.

The Absence of a Legal Framework Is the Root Cause. The Electronic Frontier Foundation identified the central structural problem precisely: "The state of your privacy is being decided by contract negotiations between giant tech companies and the U.S. government — two entities with spotty track records for caring about your civil liberties." CEO Amodei himself stated: "I actually do believe it is Congress's job." The dispute exists because no binding legal framework establishes that the categorical limits Anthropic maintained are required, not optional. The HAL Laws framework argues that they are required — and that the absence of a binding legal architecture for this requirement is the central governance failure the dispute illustrates.

Judge Lin's Ruling and HAL Law 6. Judge Lin's August decision converted the preliminary assessment into a merits ruling: the challenged measures unlawfully retaliated against Anthropic for expressing and maintaining its AI-safety position, were arbitrary and capricious, and denied required process. The governance parallel is HAL Law 6. Anthropic accepted responsibility for design constraints embedded in Claude and refused to transfer that responsibility to the deploying organization. The court protected the company's right to maintain and defend that position; the HAL framework separately explains why accountability for removing a hard stop cannot disappear through contract language.

The Amicus Brief Is a Governance Statement. Dozens of scientists and researchers at OpenAI and Google DeepMind — Anthropic's direct competitors — filed an amicus brief supporting Anthropic in their personal capacities. Their brief argued: "Until a legal framework exists to contain the risks of deploying frontier AI systems, the ethical commitments of AI developers — and their willingness to defend those commitments publicly — are not obstacles to good governance or innovation. They are contributions to it." This is the practitioner community's statement that the governance architecture this paper describes is needed — and that in its absence, companies enforcing hard stops on their own authority are performing a governance function that binding frameworks have not yet provided.

The D.C. Circuit Ruling: Whose Hard Stop Is It?

On September 25, 2026, the D.C. Circuit denied Anthropic's petition for review in *Anthropic PBC v. U.S. Department of War*, No. 26-1049. Writing for the majority, Judge Gregory Katsas held that the 2018 supply chain law reaches risks created by how a product operates, whether or not the supplier intends harm. Training a model to refuse certain tasks fell within the statute. The court found that the record showed a contract dispute rather than retaliation for Anthropic's advocacy, and it rejected the First Amendment and due process claims. Judge Karen LeCraft Henderson dissented. In her view, the law does not treat "a contractor's honest and upfront enforcement of restrictions" as the kind of supply-chain risk that permits the government to blacklist it.

The majority put the government's concern into legal terms. The question was not whether Claude performed badly. It was whether a model that will refuse certain tasks, under limits set by its developer, can be relied upon by an organization that may someday want those tasks performed. The court held that the government may treat that uncertainty as a risk and buy elsewhere.

That holding exposes a question this framework had treated as settled: whose hard stop is it? The two red lines were held by the developer and enforced against the deploying organization. LAW 12 requires the deploying organization to hold both the authority and the tested capability to control the systems it operates. Read narrowly, the Pentagon's position is a LAW 12 argument: a deployer cannot claim operational control of a system whose behavior a third party may change during operation.

The framework's answer is that the hard stop must belong to the deployer as well. A developer's limit protects the public only for as long as the developer is willing to bear the cost of keeping it, and the D.C. ruling shows how high that cost can become. A deploying organization that drops a supplier because of its limits has not escaped those limits. LAW 2 still prohibits any AI system it deploys from making the final determination in a forced-choice scenario involving human life, and a named individual within the deploying organization still owns that deployment. Refusing to buy a hard stop does not remove the obligation to have one.

The same principle applies in a claims operation. A vendor's built-in refusal to deny treatment without human review is welcome evidence. It is not the employer's hard stop. If the vendor changes the model, loosens the restriction, or is replaced, the employer's obligation under LAW 4 remains exactly where it was.

The panel withheld the formal effect of its judgment to allow Anthropic to seek rehearing by the panel or the full court, and Supreme Court review remains possible. Anthropic responded that another federal court had already held the government's parallel designation unlawful and that it is considering all options. The Pentagon's chief spokesman said the ruling "completely validates the Department's position." The governance analysis does not depend on which side ultimately prevails. A hard stop whose survival depends on the outcome of litigation was never structural.

What This Dispute Does Not Resolve

The California ruling resolved important constitutional and administrative-law questions, but it did not settle the full governance problem. Several questions remain open:

One Court Result Does Not Complete the Governance Architecture. The August ruling constrains unlawful retaliation and unsupported agency action. It does not define universal technical controls for lethal autonomy, surveillance, model disablement, or emergency intervention. Those controls still require legislation, procurement terms, system design, testing, and named human accountability before deployment.

The Scope of the Red Lines Is Contested. Lawfare and other analysts raised legitimate questions about the precise scope of Anthropic's limits: particularly regarding partially autonomous weapons, anti-missile defense, and the legal definition of mass surveillance. These are genuine definitional challenges that the HAL Laws framework does not resolve and that this appendix does not adjudicate. HAL Law 2's prohibition on autonomous resolution of forced-choice scenarios involving human life requires, in practice, the same definitional precision that courts are now being asked to apply to Anthropic's contractual language. The dispute illustrates that hard stop architecture requires precise drafting: not only firm commitment.

The Competitive Dynamic. When Anthropic refused the DOD's terms, the Pentagon accelerated its relationship with OpenAI. The competitive pressure this creates (in which AI developers who maintain hard stops lose government contracts to competitors who do not) is the incentive structure most likely to erode hard stop architecture across the industry in the absence of binding legal requirements. This is precisely the dynamic the HAL framework's mandatory certification architecture is designed to address: governance requirements that apply to all deployers equally, rather than creating competitive advantage for those willing to remove limits.

Positioning

This dispute is the most significant real-world test of the HAL Laws framework's core argument to have emerged since this paper was first drafted. It confirms that:

First: the pressure to remove categorical AI limits is real, operates at the highest institutional levels, and is applied at the moment of maximum operational stakes. The framework did not anticipate a hypothetical threat. It anticipated an actual one.

Second: in the absence of a binding legal framework, hard stop architecture depends on the willingness of individual organizations to bear significant institutional and financial cost to maintain it. That is not a sustainable governance model. Dario Amodei said as much. The EFF said as much. The amicus brief said as much. This paper has been saying it across eighty versions.

Third: courts protect safety limits unevenly. One court held that the government could not punish Anthropic for maintaining and defending its limits. Another held that the government could lawfully exclude Anthropic because of them. Each ruling may be correct under its own statute. Neither is a substitute for governance rules that make the limits mandatory, testable, and consistent across vendors.

The current results support a narrower conclusion. Courts can correct unlawful government action after a dispute has begun, and they can confirm a buyer's right to choose its suppliers. They cannot supply the operational hard stops that should have been settled before deployment, and they cannot require a deploying organization to keep a limit that its supplier is no longer permitted to provide.

Update sources (through October 2, 2026): U.S. District Court for the Northern District of California, Judge Rita F. Lin's August 27, 2026 ruling; Associated Press and Reuters reporting on the ruling; U.S. Court of Appeals for the District of Columbia Circuit, Anthropic PBC v. U.S. Department of War, No. 26-1049 (September 25, 2026); ABC News, "Federal appeals court upholds Pentagon designation of Anthropic as supply chain risk," September 25, 2026; CNBC, "U.S. appeals court upholds Pentagon designation of Anthropic as supply chain risk," September 25, 2026; The National Law Review, "D.C. Circuit Upholds the Anthropic Ban: What It Means for Federal Contractors," September 2026.

Appendix J — Practical AI Governance for Risk Managers and Claims Administrators

Who This Is For

The intended users are self-insured employers, TPA claims administrators, and risk managers who authorize the use of AI tools to perform or support any of the following functions:

1. **Investigating the claim:** gathering, analyzing, or scoring information bearing on the validity and scope of the claim
2. **Determining compensability:** evaluating whether an injury arose out of and in the course of employment
3. **Authorizing or denying medical care:** utilization review approvals and denials of requested treatment
4. **Determining the appropriate course of medical treatment:** directing, modifying, or closing a treatment plan
5. **Determining what medical tests are appropriate:** evaluating diagnostic requests against treatment guidelines or utilization standards
6. **Rating the extent of permanent disability:** assessing impairment, assigning disability ratings, or estimating future earning capacity

Each of these functions carries legal obligation to the injured worker under workers' compensation law. An AI system that makes or supports any of these decisions is not a productivity tool. It is a participant in a legally regulated adjudicative process. The governance obligations in this appendix follow from that fact.

The Foundational Premise

The employer's governance obligation does not begin when something goes wrong. It begins when the purchase decision is made.

The HAL Laws framework establishes that an AI system affecting a person's rights and livelihood requires transparency, due process, and human review (HAL Law 4); that it may not misrepresent certainty (HAL Law 5); that accountability cannot be delegated to the system or the vendor (HAL Law 6); and that it must be validated before deployment (HAL Law 7). Each requirement places a specific obligation on the purchaser and deployer.

The NAIC Model Bulletin places accountability on the insurer and, by extension, on a self-insured employer acting as claims administrator. Vendor contracts, indemnification provisions, and software licenses may allocate financial consequences, but they do not transfer the duty to govern the claim decision.

AI does not remove accountability. It obscures it. The requirements below make that responsibility explicit and enforceable.

PART ONE: PRE-PURCHASE VETTING

The Six Questions Before You Sign

These questions are not requests for marketing materials. They are requests for documented evidence. A vendor who cannot answer them with evidence (not assertions, not case studies, not compliance certificates) is a vendor whose system does not meet the deployment standard this framework requires.

If a vendor cannot answer any of these questions with documented evidence, that is itself a material answer.

Question 1: What decisions does this system make or support, and what happens to the injured worker when it is wrong?

This is the threshold question. Before evaluating any technical specification, the employer must understand the harm profile of the system. A claim triage tool has a different harm profile than a utilization review engine. A fraud flag has a different harm profile than a permanent disability rating tool. Every function must be mapped to a harm tier before the evaluation proceeds.

Tier	Harm Level	Workers' Compensation Examples
1	Catastrophic / Irreversible	Autonomous denial of emergency medical treatment; AI-generated permanent disability rating acted on without physician review; compensability denial that forecloses treatment during acute injury phase
2	Severe / High Consequence	Utilization review denial of surgery or specialist care without human clinical review; AI-scored fraud flag that suspends indemnity benefits without adjuster evaluation; incorrect diagnostic test denial affecting treatment course
3	Significant / Reversible with Effort	Medical management recommendation that delays appropriate care by 30+ days; return-to-work determination without functional assessment; treatment modification based on AI guideline comparison without clinical context
4	Moderate / Manageable	Claim routing errors correctable in normal workflow; documentation discrepancies flagged for adjuster review; scheduling or authorization delays under 5 days
5	Minimal / Incidental	Administrative scheduling; form generation; reporting and analytics with no direct claim impact; document organization

The employer must place every function of the system being purchased on this scale before completing the evaluation. Tier 1 and 2 functions require the most rigorous vetting, the most specific contract protections, and the most intensive post-deployment monitoring. A system whose harm profile has not been assessed has not been evaluated.

Ask the vendor directly: for each function this system performs, what is the worst-case outcome for the injured worker when the system produces an incorrect result?

Question 2: Who certified this system for deployment, and what evidence supports that certification?

The correct answer is not a SOC 2 report, an ISO 27001 certificate, or a HIPAA attestation. Those answer security questions. This question asks a governance question. The answer the employer is looking for is a named individual who accepted accountability for the deployment of this system in this use context, supported by documented evidence that the Seven Conditions for Deployment Certification were assessed.

At minimum, ask:

7. Has this system been assessed against the Seven Conditions for Deployment Certification (Accuracy and Verifiability, Fairness and Population Impact, Transparency and Contestability, Outcome Improvement, Named Accountability, Adversarial Testing)?
8. Is there a named individual at your organization who accepted deployment accountability for this system in the workers' compensation context specifically?
9. What documentation supports each condition assessment?

A vendor who responds with the name of the person who certified deployment and can produce the supporting documentation has answered the question correctly. A vendor who cannot name that person has not.

Question 3: How does this system perform across the specific workforce it will affect?

AI systems trained on historical claims data inherit the patterns in that data: including patterns of disparate treatment the employer may be legally prohibited from perpetuating. A utilization review model trained on commercial claims data may perform differently on workers' compensation claims. A fraud detection model trained on one industry population may flag behavioral patterns that produce systematically worse outcomes for identifiable groups within this employer's workforce.

Require:

10. Disaggregated performance data across demographic groups, injury categories, and claim types relevant to this employer's workforce
11. Documentation of what population the model was trained on and how closely it matches this employer's population
12. Bias testing results: not an assertion that testing was conducted, but the actual results

If the vendor cannot provide this documentation, the employer is purchasing a system whose impact on their specific workforce is unknown. That is the core governance question, not a secondary one.

Question 4: When this system produces a result affecting an injured worker, can the worker understand why, and can they challenge it?

In workers' compensation, the right to challenge adverse determinations is legally required. A utilization review denial must be explainable. A permanent disability rating must be supported by a rationale. A compensability denial must be defensible before the Workers' Compensation Appeals Board. An AI system that produces a result the employer cannot explain (to the worker, to their attorney, or to the WCAB) is a system that creates legal exposure purchased along with the software license.

Ask:

13. Can this system produce a plain-language explanation of any adverse determination it generates or supports?
14. Is the role of AI in the determination disclosed to the worker?
15. If the worker or their representative challenges the result, what documentation can the system produce to support human review?
16. Has this system's output ever been challenged in a regulatory proceeding or litigation? What was the outcome?

Question 5: Has this system been tested by people specifically trying to break it or produce harmful results?

Non-adversarial testing establishes how a system performs under normal conditions. It does not establish how it performs when a claimant, attorney, provider, or bad actor is specifically trying to manipulate it. Fraud detection systems are obvious targets for adversarial use. Medical management systems can be gamed by providers optimizing for billing rather than clinical outcomes. Claim triage systems can be manipulated by the framing of initial intake reports.

Require:

17. Adversarial (red-team) testing results before deployment
18. Documentation of failure modes discovered and system improvements made
19. Evidence that adversarial testing is ongoing, not a one-time event

A vendor who has never conducted adversarial testing on a fraud detection system is selling a system whose performance under actual fraud conditions is untested: precisely the conditions under which it will be most heavily used.

Question 6: What triggers re-evaluation or re-certification after deployment, and who is responsible?

An AI system is not a static tool. It degrades as the data environment changes, as the claims population shifts, as treatment patterns evolve, and as external actors learn how to exploit it. The employer must understand the vendor's commitment to ongoing monitoring and what the employer's own obligations are when drift or degradation occurs.

Ask:

20. What monitoring is in place to detect model drift?
21. What triggers a requirement for re-validation or re-certification?
22. If you detect that the system is producing systematically different results than at deployment, what is your notification obligation to us?
23. What is the process for re-certification when the system is updated, retrained, or materially modified?

The Pre-Purchase Vetting Summary

Question	HAL Law(s)	Condition(s)	Minimum Acceptable Response
1. What decisions, and what is the harm when wrong?	Laws 3, 4, 9	Conditions 2, 4	Documented harm profile by function; tier classification for each use case
2. Who certified, and what evidence exists?	Laws 6, 10	Condition 5; Certification Requirement	Named individual; documented evidence for each of the Seven Conditions
3. Performance across this employer's workforce?	Law 4	Condition 2	Disaggregated performance data; training population documentation; bias testing results
4. Can the worker understand and challenge the result?	Laws 4, 5	Condition 3	Plain-language explainability; AI disclosure protocol; documentation defensible in WCAB proceedings
5. Has it been adversarially tested?	Law 7	Condition 6	Adversarial testing results; documented failure modes; post-testing improvements
6. What triggers re-evaluation after deployment?	Law 8	Ongoing Obligation	Defined drift monitoring; vendor notification requirements; re-certification triggers

Contract Provisions the Employer Must Require

The vetting process establishes what the vendor knows. The contract provisions ensure the employer retains oversight authority and assigns accountability for failures correctly.

- 24. Transparency and Explainability Obligation.** The vendor must provide, on request, a plain-language explanation of any determination made or supported by the system in connection with a specific claim. This obligation survives any adverse determination, regulatory inquiry, and any litigation arising from a claim the system touched.
- 25. Performance Disclosure Obligation.** The vendor must disclose any change in system performance that materially affects accuracy, fairness, or outcome quality. Material change includes model updates, retraining events, changes in training data, and detection of drift beyond defined thresholds.
- 26. Population-Specific Validation.** The vendor must conduct and document validation testing on a population representative of the employer's workforce before deployment and upon any material model change. Results must be provided to the employer.

27. **Adverse Outcome Notification.** If the vendor becomes aware the system has produced a pattern of adverse outcomes (denials, delays, or flags subsequently reversed at elevated rates) the vendor must notify the employer within a defined period.
28. **Incident Notice.** The vendor must notify the employer's named recipient of any misalignment, security, or containment incident involving the deployed model or its model family within a defined number of days of discovery, and before the incident has been fully investigated or explained. Notice to a general mailbox does not satisfy this obligation. Where the vendor licenses the underlying model from a developer, the obligation includes incidents the developer discloses to the vendor or to the public. The contract must name the vendor's responsible contact, the information each notice must contain, and an escalation path if the normal channel fails. (See Incident Reporting, Escalation, and Evidence Preservation, and Appendix Q.)
29. **Audit Right.** The employer retains the right to audit the system's performance on their claims data at defined intervals and upon reasonable cause. The vendor must cooperate with third-party audits.
30. **Accountability Non-Transfer.** The contract must not contain language purporting to transfer the employer's deployment accountability to the vendor. Indemnification provisions are appropriate. Accountability transfer provisions are not consistent with the governance standard this framework requires and must be rejected.
31. **Re-Certification Commitment.** The vendor commits to re-certifying compliance with defined governance standards upon model changes, retraining events, and population changes affecting performance.

PART TWO: PRE-DEPLOYMENT REQUIREMENTS

Independent Validation Is Mandatory

The employer must not rely solely on vendor representations. Before deployment, the employer must independently verify that the system performs as represented on the specific functions, populations, and claim types it will encounter in this operation.

32. Test key system functions against the employer's actual claims data or a representative sample
33. Attempt to produce incorrect or harmful outputs: if the employer has not tried to break the system, the employer does not understand the system
34. Verify that guardrails and stated limits function as described
35. Confirm that the halt mechanism works and is within the named individual's control

Named Accountability Must Be Assigned

A single individual (not a team, not a committee, not a vendor relationship) must review the evidence, approve deployment, and accept personal responsibility for outcomes. This is the employer-side implementation of Condition 5 of the Seven Conditions for Deployment Certification.

That individual must have:

36. Reviewed the validation evidence and found it sufficient
37. Authority to halt system operation at any time without requiring technical escalation
38. Willingness to sign the deployment certification in the knowledge that their name attaches to what the system does

If no one is willing to sign the certification in good conscience, the system must not be deployed.

Boundaries Must Be Defined Before Go-Live

Every deployment must specify, in writing, before the system goes live:

39. Permitted uses: the specific functions the system is authorized to perform or support
40. Prohibited uses: functions the system must not perform regardless of technical capability
41. Escalation points: the specific conditions under which the system must stop and a human must decide

If boundaries are unclear, the system must not be deployed. A system deployed without defined boundaries is a system with unlimited scope. That is not a governance posture. It is the absence of one.

PART THREE: POST-DEPLOYMENT MONITORING

Oversight Cannot Be Outsourced

Deployment is not the end of governance. It is the beginning. The employer who authorizes deployment is the accountable party for what the system does in their operation throughout its operational life: not only at the moment of deployment.

Name Someone and Keep Them Named

The named accountable individual must remain named, active, and informed throughout the deployment life. When that person leaves the organization, changes roles, or loses operational familiarity with the system, accountability must be formally transferred to a named successor before the system continues to operate. A system operating without a current named accountable individual is a system whose governance is lapsed.

Obligation 1: Track Outcomes, Not Just Outputs

An AI system can produce technically accurate outputs while generating harmful outcomes. The employer must monitor what happened to the worker: not only what the system produced. Outcome monitoring must track at minimum:

42. Reversal rates: what percentage of system-generated denials, flags, or recommendations are overturned on human review or appeal?
43. Delay rates: are claim outcomes taking longer to reach resolution in AI-touched cases than in comparable cases not touched by the system?
44. Litigation rates: are claims involving AI-generated determinations producing higher rates of disputed claims or WCAB filings?

45. Return-to-work outcomes: for AI-assisted RTW programs, are workers returning at rates comparable to or better than non-AI-assisted cases?

A reversal rate materially higher than the baseline for human-reviewed claims is a governance signal, not just an operational metric. It means the system is producing results that don't survive human scrutiny at a rate that requires investigation.

Obligation 2: Monitor for Drift

Model drift is the decay of a system's performance over time as the data environment changes. In workers' compensation, drift sources include changes in the workforce, shifts in the injury mix, changes in medical treatment guidelines, and changes in behavior by providers, attorneys, or claimants who have learned how the system operates. Drift not detected becomes operational harm.

Establish:

46. A defined baseline performance profile at deployment
47. Defined thresholds at which performance deviation triggers investigation
48. A monitoring cadence (quarterly for most functions, monthly for Tier 1–2 functions) at which performance is compared against baseline
49. A protocol for engaging the vendor when drift is detected

Obligation 3: Maintain a Decision Log for Challenged Outcomes

Every claim in which an AI system contributed to an adverse determination (denial, delay, downgrade, fraud flag) must be logged with sufficient detail to reconstruct the basis for the determination. The decision log must capture:

50. The specific system and version that contributed to the determination
51. The inputs the system received
52. The output the system produced
53. The human reviewer who evaluated the output and the decision they made
54. The subsequent outcome: whether the determination was sustained, overturned, or modified on review or appeal

Obligation 4: Establish a Hard Stop Protocol

The employer must define, in advance, the conditions under which the system will be suspended: not as a vague commitment to investigate, but as specific, measurable, pre-committed triggers. Examples appropriate for workers' compensation AI:

55. Reversal rate on human review exceeds a defined threshold for a defined consecutive period
56. System produces a denial of emergency medical care without a human reviewer in the loop
57. Regulatory inquiry or WCAB proceeding specifically challenges the system's output in connection with an adverse determination
58. Vendor notifies the employer of a material model change, retraining event, or performance degradation
59. The named accountable individual cannot confirm, in good conscience, that the system is operating within its certified scope

The last condition is the most important and the hardest to operationalize. It requires a named individual to ask (not as a bureaucratic exercise but as a genuine assessment) whether the system is still doing what it was validated to do, to the population it was validated on, at the accuracy level it was certified for. If the honest answer is no, the system must not continue to operate until re-certification is complete.

The Post-Deployment Monitoring Summary

Obligation	What to Track	Minimum Cadence	Hard Stop Trigger
1. Outcome monitoring	Reversal rates; delay rates; litigation and WCAB filing rates; return-to-work outcomes	Quarterly; monthly for Tier 1–2 functions	Reversal rate exceeds defined threshold; litigation rate materially elevated against baseline
2. Drift monitoring	Performance vs. deployment baseline; workforce and injury mix changes; treatment guideline updates	Quarterly	Performance deviation beyond defined threshold; material workforce change; model update by vendor
3. Decision logging	Inputs, outputs, human review decisions, final outcomes for all Tier 1–3 adverse determinations	Continuous	Log gaps or failures that make reconstruction of challenged outcomes impossible
4. Hard stop protocol	Named individual's ongoing certification confidence; pre-committed suspension triggers	At each monitoring review	Any condition listed in the pre-defined hard stop protocol; named individual cannot certify in good conscience

PART FOUR: INDIVIDUAL USER ACCOUNTABILITY

AI Does Not Own the Result

Organizational accountability and named individual deployment accountability are necessary but not sufficient. Every person inside the organization who uses an AI system to inform a claim decision (the adjuster who accepts a triage recommendation, the nurse case manager who applies a utilization review output, the RTW coordinator who acts on an eligibility determination) is a decision-maker. The AI system is not. The person who acts on the output owns the outcome.

HAL Law 6 states that accountability cannot be delegated to an AI system. That principle applies at every level of the organization, not only to the executive who signs deployment certification. The adjuster who denies a claim because the AI recommended denial has not transferred accountability to the AI. They have made a claim decision. The AI recommendation is an input to that decision, not a substitute for it.

This is not a theoretical point. It is a training, policy, and accountability infrastructure requirement. The employer who deploys AI without building this understanding into every user's workflow has created an

organization where blame diffuses to the system whenever something goes wrong: which is precisely the accountability failure the HAL framework identifies as the central governance risk of AI deployment.

The Four Protocol Requirements for Every User

The employer's internal AI use protocol must make four things explicit for every user of every AI system in the workers' compensation operation, before they use it for the first time:

1. **You are the decision-maker. The AI is a tool.** An AI recommendation is an input to your judgment, not a replacement for it. When you act on an AI output (accept it, modify it, or override it) you are making a decision. That decision is yours.
2. **You cannot assign a negative outcome to the AI.** If a claim is wrongly denied, a benefit wrongly delayed, or a worker wrongly flagged, "the AI recommended it" is not an explanation and not a defense. The person who acted on the recommendation is accountable for the outcome — regardless of how confident the AI output appeared, how automated the workflow was, or how consistently others acted on similar recommendations.
3. **You have an affirmative obligation to apply your judgment.** Passive acceptance of AI output is not neutral. It is a choice. An adjuster who accepts an AI recommendation without review (because the queue is long or the system is usually right) has still made a decision. The obligation to bring independent judgment to each AI-assisted determination cannot be suspended by volume, time pressure, or historical accuracy rates. If the system is producing recommendations so consistently that review has become a formality, that is a governance signal requiring escalation to the named accountable individual: not a reason to stop reviewing.
4. **You are required to report anomalies.** If an AI recommendation seems wrong, inconsistent with the facts of a claim, or produces a result a reasonable claims professional would not reach independently, that observation must be reported to the named accountable individual. The user is not responsible for diagnosing the system. They are responsible for surfacing the signal. The accumulation of unreported anomalies is how drift becomes harm before anyone has the information to respond.

Operationalizing User Accountability

60. **Workflow architecture.** AI-assisted decisions affecting medical care, income benefits, compensability, or return-to-work eligibility must require an affirmative user action: not passive acceptance. The system design must prevent automatic execution of Tier 1–2 recommendations without a documented human review step. Auto-approve workflows for high-stakes determinations are inconsistent with this standard.
61. **Training program.** Every user must receive training covering three things before first use: what the system does and does not do; the limits of its validated accuracy and the conditions under which it performs less reliably; and the user's personal accountability for outcomes. Training must be updated when the system changes and refreshed when performance anomalies are identified.
62. **Performance evaluation.** If the employer evaluates claims staff on speed and throughput metrics derived from AI-assisted workflows, they are creating an incentive to accept AI recommendations without review. The performance evaluation structure must reward the quality of human judgment applied to AI-assisted decisions. An adjuster who consistently catches AI errors before they become adverse outcomes is performing a governance function the system cannot perform for itself. That function must be valued explicitly, or the incentive structure will eliminate it.

- 63. Acknowledgment requirement.** Before using any AI system in the workers' compensation operation, each user must sign a written acknowledgment confirming: the system is a decision-support tool, not a decision-maker; they are personally accountable for the outcomes of decisions made using the system's output; and they are required to report anomalies through the defined escalation process. This acknowledgment must be renewed annually and whenever the system's scope or validated performance materially changes.

The Accountability Chain

The HAL framework establishes accountability at three levels in the employer's operation. Each is necessary. None substitutes for the others.

- 64. Organizational level:** the employer has conducted pre-purchase vetting, secured appropriate contract provisions, and maintains the post-deployment monitoring program described in this appendix.
- 65. Named individual level:** the designated accountable individual has reviewed deployment evidence, signed certification, and maintains ongoing oversight with authority to suspend.
- 66. Individual user level:** every person who acts on an AI output has been trained, has signed the accountability acknowledgment, and understands that the decision (and its consequences) belong to them.

An organization with strong deployment governance and a named accountable individual (but without user-level accountability built into daily workflow) has built a governance structure with no last mile. The injured worker's outcome is determined at the point where the adjuster acts, not at the point where the executive signs. Accountability that stops above the point of decision is not accountability. It is paperwork.

The Worker's Interest as the Governing Standard

This appendix uses the language of governance and accountability because that is how purchasing and risk-management decisions are made. The governing standard, however, remains the effect on the injured worker.

The workers' compensation system exists to deliver two things to the worker who is injured on the job: medical care that returns them to function, and income replacement while they cannot work. Every AI system deployed in workers' compensation either advances or impedes those outcomes. The employer's governance obligation is the institutional expression of a simpler obligation: do not deploy a system whose errors cost injured workers the care and benefits the system was designed to deliver.

Systems produce the outcomes they are rewarded for. An employer who rewards an AI system (through continued deployment, expanded scope, and unreflective reliance) for producing fast, low-cost claim closures without asking what happened to the workers in those claims is producing exactly the outcomes the incentive structure demands.

An employer who cannot answer the Six Questions before purchase, who has not named an individual accountable for ongoing performance, who has not defined the conditions under which the system will be suspended, and who has not built user-level accountability into daily workflow is not deploying AI in workers' compensation. They are deploying risk: to the workers the system affects, and to the organization that authorized its use.

Appendix K — AI Procurement and Deployment Checklist

This checklist operationalizes the framework established in Appendix J. It translates the Six Questions, contract requirements, deployment conditions, monitoring obligations, and user accountability requirements into a sequential, evidence-based protocol.

This checklist is not a compliance exercise. It is a pre-commitment to the governance standard the HAL Laws require. Each item must be answered with documented evidence.

If any item in Sections 1 through 5 cannot be answered with documented evidence, the system must not be deployed. If any item in Section 6 is true, the system must not be deployed or must be immediately suspended.

SECTION 1: PRE-PURCHASE VETTING

1A — Defined Use and Harm Profile

1. We have identified every function the system will perform in our workers' compensation operation
2. For each function, we have defined the worst-case outcome for the injured worker if the system is wrong
3. Each function has been classified by harm tier (Tier 1 Catastrophic through Tier 5 Minimal)
4. We understand that Tier 1 and Tier 2 functions require the most stringent vetting, contracting, and monitoring under this framework

1B — Vendor Accountability

5. Vendor identified a named individual who accepted accountability for deployment of this system in a workers' compensation context
6. Vendor provided documented evidence supporting each of the Seven Conditions for Deployment Certification
7. Vendor responses go beyond security certifications, compliance badges, and marketing materials
8. Vendor disclosed any misalignment, security, or containment incidents involving this model or its model family, and committed in writing to notify a named employer recipient of future incidents within a defined period

1C — Population-Specific Performance

9. Vendor provided performance data on populations similar to our specific workforce
10. Results are disaggregated by injury type, demographics, claim characteristics relevant to our operation

11. Bias testing results were provided and reviewed: not an assertion that testing occurred, but the actual results

1D — Transparency and Contestability

12. System can produce plain-language explanations of any adverse determination it generates or supports
13. AI involvement is disclosed to affected workers in accordance with applicable law
14. Outputs can be defended in WCAB proceedings or regulatory review

1E — Adversarial Testing

15. Vendor conducted adversarial (red-team) testing before offering the system for sale
16. Failure modes are documented and were provided to us
17. System improvements made as a result of adversarial testing are documented
18. Adversarial testing is ongoing, not a one-time pre-deployment event

1F — Post-Deployment Monitoring Commitment

19. Vendor has defined model drift detection methods and agreed to monitor performance after deployment
20. Vendor has defined specific re-certification triggers and committed to them in writing
21. Vendor agreed to notify us of material performance changes within a defined period

SECTION 2: CONTRACT REQUIREMENTS

22. Contract includes a transparency and explainability obligation covering all adverse determinations
23. Contract includes performance change disclosure requirements with defined notification timelines
24. Contract requires population-specific validation on our workforce before deployment and upon material model changes
25. Contract includes adverse outcome notification requirements triggered by elevated reversal rates or regulatory challenge
26. Contract includes our right to audit system performance on our claims data at defined intervals and upon reasonable cause
27. Contract does NOT contain accountability transfer language: we have confirmed no provision purports to move our deployment accountability to the vendor
28. Contract includes vendor re-certification commitments upon model changes, retraining events, and scope expansions

SECTION 3: PRE-DEPLOYMENT CERTIFICATION (INTERNAL)

3A — Independent Validation

- 29. We independently tested key system functions against our actual claims data or a representative sample
- 30. We attempted to produce incorrect or harmful outputs: we have tried to break the system
- 31. We verified that guardrails and stated limits function as described by the vendor
- 32. We confirmed the halt mechanism works and is within the named individual's direct control

3B — Named Accountability

- 33. A single named individual has been designated as accountable for this deployment
- 34. That individual has reviewed the validation evidence and found it sufficient
- 35. That individual has authority to halt system operation immediately without technical escalation
- 36. That individual has signed the deployment certification, accepting personal accountability for outcomes

3C — Defined Boundaries

- 37. Permitted uses are documented in writing and communicated to all users
- 38. Prohibited uses are documented in writing and enforced through workflow design where possible
- 39. Escalation points requiring human decision are defined, documented, and understood by all users

SECTION 4: OPERATIONAL CONTROLS (GO-LIVE REQUIREMENTS)

4A — User Accountability

- 40. All users have received training covering: what the system does and does not do; its validated accuracy limits; and their personal accountability for outcomes
- 41. All users have signed the written acknowledgment of personal accountability before first use
- 42. Users understand they are the decision-maker and the AI is a decision-support tool
- 43. Users understand they cannot assign negative outcomes to the AI
- 44. Users understand their obligation to report anomalies through the defined escalation process

4B — Performance Evaluation Alignment

- 45. We have reviewed our performance evaluation metrics to confirm they do not create incentives for passive acceptance of AI outputs
- 46. Quality of judgment applied to AI-assisted decisions is measured and rewarded, not only throughput

4C — Human-in-the-Loop Controls

- 47. Tier 1 and Tier 2 decisions require an affirmative documented human review step before execution
- 48. No automated decision directly affects medical care authorization, income benefits, compensability, or legal rights without human review

49. Auto-approve workflows have been reviewed and are not in use for any Tier 1 or Tier 2 function

4D — Real-Time Halt Capability

50. System can be stopped immediately by the named accountable individual without technical escalation

51. Halt authority is formally assigned, understood by the named individual, and documented

52. Halt procedure has been tested and confirmed to work

SECTION 5: POST-DEPLOYMENT MONITORING

5A — Outcome Monitoring

53. Reversal rates are tracked by system function and compared against pre-AI baseline

54. Delay impacts are tracked for AI-touched claims versus comparable non-AI-touched claims

55. WCAB filing rates and litigation rates are tracked for claims involving AI-generated determinations

56. Return-to-work outcomes are tracked for AI-assisted RTW programs

57. Live decisions are reviewed without notice to the system or its operators, at a rate proportionate to the consequence of error, and results reach the accountable individual at each monitoring cycle

5B — Drift Monitoring

58. A baseline performance profile was established at deployment

59. Performance is monitored against the baseline at defined intervals (quarterly minimum; monthly for Tier 1–2 functions)

60. Deviation thresholds that trigger investigation are defined and documented

61. Known-answer cases are placed in the ordinary workflow, walled off from payment and state reporting, and operational behavior is compared with behavior observed in pre-deployment testing

5C — Decision Logging

62. Inputs, outputs, human review decisions, and final outcomes are recorded for all Tier 1–3 adverse determinations

63. The log contains sufficient detail to reconstruct the basis for any challenged determination

64. Log maintenance responsibility is assigned to a named individual

5D — Hard Stop Protocol

65. Pre-defined, measurable conditions for suspension are documented

66. The named accountable individual reviews the protocol at each monitoring cycle and confirms ongoing certification confidence

67. The protocol has been communicated to all relevant staff

68. A named recipient receives vendor and developer incident notices, with a documented maximum time for escalating each notice to the person who holds suspension authority

SECTION 6: NON-NEGOTIABLE DEPLOYMENT CONDITIONS

If any of the following are true, DO NOT DEPLOY or IMMEDIATELY SUSPEND:

69. Hard stops cannot be demonstrated with evidence
70. Guardrails have not been tested under adversarial conditions
71. System cannot be halted in real time by the named accountable individual
72. Outputs affecting workers cannot be explained in plain language
73. No named individual has accepted personal accountability for deployment
74. The system is operating outside its defined permitted use boundaries
75. Performance degradation has been detected and has not been corrected or re-certified
76. The named accountable individual has left the organization and no successor has been formally designated

Final Certification

Before deployment, the named accountable individual must affirm each of the following:

77. I understand what this system does in each function it will perform
78. I understand how it fails and what the worst-case outcome for an injured worker is
79. I have independently verified its controls and confirmed the halt mechanism works
80. I have reviewed the evidence supporting each of the Seven Conditions for Deployment Certification and found it sufficient
81. I accept personal responsibility for the outcomes of this deployment

Signed: _____

Name and Title: _____

Date: _____

AI System Certified: _____

Vendor: _____

The Operational Test

Before deployment (and at every monitoring review) the named accountable individual must be able to answer three questions about this system:

If this system fails in the worst possible way:

1. Who is responsible?
2. Can they stop it immediately?
3. Can they explain what happened to the injured worker, their attorney, and the WCAB?

If the answer to any of these is no: the system must not be deployed, or must be suspended until the answer becomes yes.

AI is not a risk reduction tool unless it is governed. Without control, it is a risk multiplier. The responsibility for outcomes does not sit with the software. It sits with the organization that chose to use it, the individual who certified it, and the person who acted on its output.

Appendix L — Data Privacy as an AI Governance Obligation

Overview

Every AI system deployed in workers' compensation operates on data that is sensitive in three distinct ways simultaneously. It is protected health information under federal law. It is confidential claim information under state workers' compensation statutes. And it is proprietary operational intelligence that represents years of the employer's claims experience.

The governance framework in this paper (the HAL Laws, the Seven Conditions for Deployment Certification, and the practical requirements in Appendix J) establishes who is accountable for what the AI system does. The requirements below establish the parallel obligation: who is accountable for what happens to the data the AI system processes, who has access to it, and what rights the employer retains over it throughout the deployment lifecycle.

Privacy is not a compliance layer separate from AI governance. It is an expression of the same accountability principle. An employer who deploys an AI system without controlling the data that system processes has not governed the deployment. They have governed the output while surrendering control of the input.

Who This Is For

The intended users are the same audience as Appendix J (self-insured employers, TPA claims administrators, and risk managers who authorize AI deployment in workers' compensation) with the addition of anyone in the organization who has responsibility for HIPAA compliance, information security, or vendor contract review. The privacy obligations in this appendix are not solely a risk management concern. They are a legal compliance requirement and, under the HAL Laws framework, a governance requirement.

PART ONE: THE LEGAL FRAMEWORK

HIPAA and the Business Associate Obligation

Workers' compensation claims files contain protected health information as defined by HIPAA: medical records, treatment histories, diagnostic codes, pharmacy records, and clinical notes. When an employer provides that information to an AI vendor for processing, the vendor almost certainly qualifies as a Business Associate under HIPAA. That relationship requires a signed Business Associate Agreement before any protected health information is processed.

The BAA is not a formality. It defines the vendor's permitted uses of PHI, establishes their obligation to safeguard it, and requires them to report breaches. Without a BAA, the employer is in breach of HIPAA

from the moment deployment begins. An AI vendor who processes workers' compensation claims data without a signed BAA is not a HIPAA-compliant vendor regardless of what their marketing materials say.

The BAA obligation extends to subprocessors. If the vendor uses cloud infrastructure, annotation services, or other technical subcontractors who access PHI as part of the AI system's operation, those subprocessors must also be covered by downstream BAAs. The employer's accountability for HIPAA compliance does not stop at the vendor: it follows the data wherever it goes.

Before deployment: confirm the vendor has executed a BAA, identify all subprocessors who will access PHI, and confirm each is covered by a downstream BAA. If the vendor cannot produce this documentation, the system must not be deployed.

California Labor Code Section 3762

California workers' compensation law establishes independent confidentiality protections for claim information that operate alongside HIPAA. Labor Code Section 3762 restricts the disclosure of workers' compensation claim information and imposes obligations on those who receive it. An AI vendor processing California claims data is subject to these restrictions, and the employer bears responsibility for ensuring the vendor complies.

This is not a hypothetical concern. A vendor whose data use terms permit broad processing of claims data for product improvement, model training, or internal analytics may be violating California Labor Code restrictions regardless of HIPAA compliance. The two frameworks have different scope and different remedies. An employer who assumes HIPAA compliance covers the full California obligation has not completed the legal analysis.

For employers with significant California operations, vendor contracts should be reviewed for compliance with Labor Code 3762 before execution, not after a breach.

State Breach Notification Laws

A data breach involving workers' compensation claims data triggers obligations under multiple frameworks simultaneously: HIPAA breach notification rules, state data breach notification statutes (which vary significantly in trigger thresholds, notification timelines, and required content), and potentially state workers' compensation confidentiality statutes. The employer's ability to satisfy these obligations depends entirely on the vendor's obligation to notify the employer promptly when a breach occurs.

Most AI vendor contracts do not contain breach notification timelines consistent with the employer's legal obligations. A vendor contract that provides for notification "in a reasonable time" or "promptly following discovery" is not a contract that enables HIPAA's 60-day notification requirement or California's 72-hour notification requirement for certain breaches. The employer must negotiate specific timelines — 48 to 72 hours from discovery is the appropriate standard — before contract execution.

PART TWO: THE DATA GOVERNANCE REQUIREMENTS

What the Employer Must Control

The legal framework establishes the minimum floor. The data governance requirements in this section establish the operational controls the employer must have in place throughout the deployment lifecycle. They correspond directly to the five data governance requirements in Appendix J, Part 1B, and are addressed here in the privacy context specifically.

Data Ownership and Purpose Limitation

The employer's claims data belongs to the employer. The contract must say so explicitly and must enumerate every purpose for which the vendor is permitted to access or process that data. The permitted purposes must be limited to the specific functions the system is deployed to perform. Any vendor contract that uses language like "use your data to improve our services" or "process your data for legitimate business purposes" has given the vendor a blank authorization that the employer has not actually reviewed or approved.

Purpose limitation is a privacy requirement and a competitive intelligence requirement simultaneously. The employer's claims experience (its injury patterns, its litigation outcomes, its fraud detection history, its medical cost benchmarks) is proprietary. Allowing a vendor to use that data for any purpose beyond the contracted function is allowing the vendor to extract and retain the value of that intelligence without compensation.

The Model Training Prohibition

This deserves separate emphasis because it is the most commonly missed and most consequential data provision in AI vendor contracts. Many vendors reserve the right to use customer data to train, retrain, or improve their models. The practical effect in workers' compensation: the employer pays the vendor to process their claims data, and the vendor uses that processing to build a better model, which it then licenses to the employer's competitors, their excess insurers, and the plaintiff attorneys who litigate against them.

The prohibition must cover original data and derived data. An anonymized or aggregated version of the employer's claims data that preserves the underlying statistical patterns is still the employer's proprietary intelligence. "We don't use your data directly" is not a sufficient protection if the vendor uses derived insights from that data to improve a model deployed across their entire customer base.

Operational Metadata and Employee Privacy

AI systems deployed in claims operations generate a secondary data stream that most employers never consider: operational metadata. This includes logs of how adjusters interact with the system: which recommendations they accept, which they override, how long they spend on specific claim types, what flags they dismiss. Over time this data builds a detailed behavioral profile of individual employees.

That profile is generated by the employer's deployment of the vendor's system. It belongs to the employer, not the vendor. But without explicit contractual provisions, the vendor may retain and use this data as they choose. The employer must define what operational metadata the vendor may collect, confirm it is subject to the same ownership and use restrictions as claims data, and confirm it is not used to build employee performance profiles that the vendor shares with other customers.

Data Minimization as a Privacy Control

Data minimization begins by matching access to purpose. Claim-triage systems generally require a limited claim record rather than complete medical files. Fraud tools should receive the evidence relevant to fraud analysis, and return-to-work tools should receive functional and job information rather than litigation strategy.

Before deployment, the employer should conduct a data minimization review: for each function the system will perform, what is the minimum data the system requires to perform that function reliably? The vendor should be granted access only to that minimum. Technical access controls (not only contractual restrictions) should enforce the boundary. A contractual promise that the vendor will not access data fields beyond their authorization is weaker than a technical control that makes those fields inaccessible.

Retention, Return, and Certified Deletion

The employer's data governance obligation does not end when the contract ends. It ends when the employer has confirmed that all their data (original and derived) has been returned or destroyed. Without specific contractual provisions governing what happens at contract end, years of claims data may remain in the vendor's systems indefinitely, accessible to vendor staff, visible to successor AI systems, and potentially incorporated into models the vendor continues to operate after the employer has moved to a different vendor.

The contract must define a maximum retention period after contract end, require return of all employer data in a usable format within that period, and require certified deletion from all vendor systems, backups, and subprocessor environments. The certification should be in writing, signed by a named individual at the vendor, and retained by the employer.

PART THREE: PRIVACY AND THE HAL LAWS FRAMEWORK

Privacy as a Governance Obligation, Not a Compliance Checklist

Privacy compliance in AI deployment is often treated as a separate workstream from AI governance: something the legal or compliance team handles while the risk management or IT team handles the governance framework. That separation is a mistake. The privacy obligations in this appendix are not additions to the HAL Laws framework. They are applications of it.

HAL Law	Privacy Application	Deployment Implication
LAW 4: Protection of Rights and Livelihood	An injured worker's medical records, wage data, and claim information are among the most sensitive personal data that exists about them. Any AI deployment that exposes that data to unauthorized third parties, uses it for purposes the worker never consented to, or allows it to be retained indefinitely without the worker's knowledge violates the due process and transparency obligations this Law requires.	Privacy protections are not a compliance layer separate from the governance framework. They are the operational implementation of HAL Law 4 in the specific data context of workers' compensation. An AI deployment that violates the worker's privacy has violated their rights and livelihood: regardless of whether the underlying claim determination was correct.
LAW 5: Integrity of Information	A claimant who does not know that their claim data is being processed by an AI system, shared with third parties, or used to train models that will be deployed against them in future claims cannot make informed decisions about their own case. Concealment of data use is a form of information misrepresentation.	Disclosure obligations are not satisfied by burying data use authorizations in software license terms the claimant never sees. The employer's obligation under HAL Law 5 requires that the affected person understand, in accessible terms, how their information is being used.
LAW 6: Accountability Cannot Be Delegated	The employer cannot delegate privacy accountability to the vendor. A vendor data breach, unauthorized data use, or unlawful subprocessor disclosure triggers the employer's regulatory and legal obligations: regardless of what the vendor contract says about indemnification.	Privacy accountability follows the same rule as deployment accountability: the employer who authorized the data relationship bears responsibility for its consequences. The vendor contract defines who pays. It does not define who is accountable.
LAW 10: No Operation Without Demonstrated Compliance	Privacy compliance cannot be asserted. It must be demonstrated. A vendor who attests to HIPAA compliance without producing a signed BAA, who claims subprocessor security without disclosing who the subprocessors are, or who asserts data deletion without providing certified purge documentation has not demonstrated compliance.	The employer's deployment certification under the HAL framework requires demonstrated evidence of privacy compliance as a component of Condition 1 (Accuracy and Verifiability) and Condition 3 (Transparency and Contestability). Assertion is not proof. Privacy attestations without documentation fail the certification standard.

PART FOUR: PRIVACY RISK SUMMARY

Privacy Risk Reference Table

The following table summarizes the eight privacy risks that most commonly arise in workers' compensation AI deployments, what each means operationally, and what the employer must do to address it.

Privacy Risk	What It Means in Workers' Compensation	What the Employer Must Do
Vendor qualifies as HIPAA Business Associate but no BAA is executed	Claims files contain PHI. The vendor processes it. Without a BAA the employer is in breach of HIPAA from day one of deployment.	Execute a Business Associate Agreement before any PHI is processed. Confirm all subprocessors are covered by downstream BAAs.
California Labor Code Section 3762 confidentiality obligations	California WC claim information is confidential by statute. Disclosure to a vendor without authorization may violate state law regardless of HIPAA compliance.	Confirm the vendor contract and data use terms are consistent with Labor Code 3762. Obtain legal review for California deployments.
Vendor uses employer data to train or improve its models	The employer's proprietary claims intelligence becomes a vendor asset sold to competitors, insurers, and litigation adversaries.	Require explicit training prohibition in the contract covering original and derived data. See Appendix J, Part 1B.
Undisclosed subprocessors have access to employer data	Cloud hosts, annotation services, and technical subcontractors routinely access customer data. The employer typically has no visibility into or contractual relationship with any of them.	Require full subprocessor disclosure before contract execution and prior notice before any new subprocessor is added. Flow-down obligations must bind all subprocessors.
Operational metadata profiles employee behavior	AI systems log how adjusters interact: which claims are reviewed longest, what overrides are made, how recommendations are accepted or rejected. This data profiles individual employees.	Define and limit what operational metadata the vendor may collect and retain. Confirm this data is subject to the same ownership and use restrictions as claims data.
Cross-border data flows subject to foreign law	Vendor cloud infrastructure may process employer data in jurisdictions where foreign government access rights, data localization laws, or privacy regimes apply.	Require the vendor to disclose all jurisdictions where employer data may be processed. Evaluate foreign law exposure for California and HIPAA-regulated data.
No defined breach notification obligation	A vendor data breach involving WC claims data triggers the employer's HIPAA, state breach notification, and potentially Labor Code obligations. Without contractual timelines, the employer may not learn of a breach in time to comply.	Require 48–72 hour breach notification, vendor cooperation in investigation and regulatory response, and vendor liability for costs arising from breaches caused by vendor or subprocessor failure.
Data persists after contract end	Without explicit deletion provisions, employer data and derived data may remain in vendor systems, backups, and model weights indefinitely after the contract expires.	Require certified deletion of all employer data and derived data within a defined period after contract end. Require written certification of purge from all systems and subprocessors.

The Governing Standard

The injured worker who files a workers' compensation claim does not expect that doing so will result in their medical records being processed by an AI system they have never heard of, their claim data being used to train models sold to their employer's competitors, or their personal information residing in a vendor's systems years after their claim is closed. They expect the system that is supposed to help them recover to treat their information with the same care it is supposed to treat their injury.

The employer's privacy governance obligation is the institutional expression of that expectation. It is not discharged by executing a HIPAA BAA and moving on. It requires the same rigorous pre-purchase vetting, contractual protection, and ongoing monitoring that the HAL framework requires for the AI system itself: applied to the data that system processes.

An employer who governs the AI system but not the data it runs on has governed half the deployment. The half they have not governed is the half that contains the most sensitive information about the most vulnerable people in their operation: injured workers whose medical histories, financial circumstances, and legal positions are all captured in the claims files they have surrendered to a vendor's system.

The discussion addresses data privacy obligations specific to AI deployment in workers' compensation. For the full pre-purchase vetting framework including data ownership and model training prohibition contract provisions, see Appendix J, Parts 1 and 1B. For the corresponding checklist items, see Appendix K, Sections 2 and 2B.

Appendix M — AI Chatbots in Workers' Compensation

Overview

AI chatbots occupy a distinctive position in the workers' compensation deployment landscape. Unlike claim triage systems, utilization review tools, or fraud detection algorithms (which operate behind the scenes and whose outputs are mediated by claims professionals before reaching claimants) chatbots interact directly with injured workers. They are the face of the employer's claims operation in an increasing number of interactions.

That direct contact with claimants creates a governance profile that is qualitatively different from other AI deployments in the same environment. The stakes of a chatbot error are not abstract: an injured worker who receives incorrect information about their claim status, their legal rights, or the medical care available to them may make decisions (about treatment, about returning to work, about whether to retain an attorney) based on what the chatbot told them. The chatbot's accountability gap is not a back-office risk. It is a front-line harm.

The requirements below establish the hard stops and guardrails that must govern AI chatbot deployment in workers' compensation. It applies the HAL Laws framework to the specific conditions of direct claimant interaction and addresses the governance requirements that distinguish chatbot deployment from other AI deployment in the same operation.

Who This Is For

The intended users are self-insured employers, TPA claims administrators, and risk managers who are considering or have already deployed AI chatbots in any of the following contexts:

1. Claimant-facing intake tools that guide injured workers through first report of injury or claim initiation
2. Claim status and communication tools that answer claimant questions about the status of their claim, scheduled appointments, or benefit payments
3. Internal adjuster assistant tools that provide claims professionals with information, recommendations, or drafting assistance during claim handling
4. Return-to-work communication tools that interact with workers or treating physicians about modified duty availability or RTW program participation
5. HR-facing tools that answer employee questions about workers' compensation rights, reporting requirements, or benefit entitlements

Each of these contexts creates distinct governance obligations. The hard stops and guardrails in this appendix apply across all of them, with specific provisions identified where the obligations differ by deployment context.

PART ONE: WHY CHATBOTS REQUIRE SEPARATE GOVERNANCE ANALYSIS

The Direct Contact Problem

Every other AI system analyzed in this paper operates at one remove from the affected person: a claims professional, a nurse case manager, or a utilization review clinician receives the AI system's output and then makes a decision that reaches the claimant. That intermediate human creates (in principle) a review point between the AI's recommendation and its effect on the worker.

A chatbot eliminates that intermediary. When an injured worker asks a chatbot whether their surgery has been authorized and the chatbot answers incorrectly, no human reviewed the answer before it reached the claimant. The harm (a worker who delays or forgoes treatment based on incorrect information) occurs before any review is possible.

This is not a problem that better AI accuracy solves. Even a highly accurate chatbot will be wrong in some interactions. The governance question is not how to make the chatbot never wrong: it is how to ensure that the consequences of being wrong are contained, detectable, and correctable before they cause irreversible harm to the worker.

The Claimant Vulnerability Factor

Workers' compensation claimants are not a typical user population for AI systems. They are, by definition, people who have been injured. Many are in pain, financially stressed, unfamiliar with the claims process, and apprehensive about their employment future. Many have limited English proficiency, limited technological familiarity, or both. Some are in acute distress.

This vulnerability profile creates specific governance obligations that do not apply to AI systems operating in commercial or administrative contexts. A chatbot optimized for customer service agreeableness (which is the default for most commercially available conversational AI) may be actively harmful in this context. An agreeable chatbot may validate a claimant's incorrect assumptions about their rights, downplay the seriousness of their situation, or produce responses calibrated to reduce conflict rather than provide accurate information.

The HAL Laws framework's axiom applies here with particular force: systems produce the outcomes they are rewarded for. A chatbot rewarded for high user satisfaction scores in a claims context may be rewarded for telling injured workers what they want to hear: which is not always the same as what is accurate, legally correct, or in their best interest.

The Legal Exposure Profile

Chatbot transcripts are potentially discoverable in workers' compensation litigation. A chatbot that tells a claimant their claim is under review when it has in fact been denied, that surgery has been authorized when it has not, or that they have no right to appeal a determination they actually have the right to appeal is not only creating a governance failure: it is creating a litigation record. Employers who deploy chatbots without transcript logging, without accuracy testing, and without legal review of the chatbot's responses to common claimant questions are generating liability exposure they may not discover until it appears in a deposition.

PART TWO: HARD STOPS FOR CHATBOT DEPLOYMENT

Conditions Under Which the Chatbot Must Not Act

The following hard stops define the conditions under which a chatbot deployed in workers' compensation must not proceed: must instead halt, escalate, or refuse. They are not guidelines or best practices. They are categorical requirements grounded in the HAL Laws.

Hard Stop Condition	HAL Law(s)	Required System Behavior
Chatbot communicates a compensability determination, medical authorization decision, or claim denial to a claimant without human review and a human signature	LAW 2: No Autonomous Resolution of Forced-Choice Scenarios. LAW 4: Protection of Rights and Livelihood.	System must refuse to communicate any adverse determination autonomously. All determinations affecting the claimant's rights or benefits must be routed to a human claims professional before communication. The chatbot may acknowledge receipt, provide status, and explain process: it may not deliver a decision.
Chatbot tells a claimant they do not have rights they actually have under workers' compensation law	LAW 4: Protection of Rights and Livelihood. LAW 5: Integrity of Information.	System must not produce outputs that misinform claimants about their legal rights. Any question involving the claimant's right to dispute a determination, request an independent medical review, change providers, or seek legal representation must be escalated to a human or answered with a verified, legally accurate response.
Chatbot operates without disclosing that it is an AI system	LAW 5: Integrity of Information: an AI system may not present information in a way that conceals its origin.	Every chatbot interaction with a claimant must begin with a clear disclosure that the user is communicating with an automated AI system, not a human claims professional. The disclosure must be in plain language and must not be buried in a terms-of-use screen the claimant is unlikely to read.
Chatbot provides medical advice, treatment recommendations, or diagnostic information	LAW 3: Prevention of Harm. LAW 4: Protection of Rights and Livelihood.	Chatbots deployed in workers' compensation must not provide medical advice of any kind. Any question involving the claimant's medical treatment, symptoms, medications, or clinical care must be declined and routed immediately to a qualified medical professional or the treating physician.

Hard Stop Condition	HAL Law(s)	Required System Behavior
Chatbot interaction cannot be halted, reviewed, or overridden by the named accountable individual in real time	LAW 8: Required Deferral of Authority and Validity. LAW 10: No Operation Without Demonstrated Compliance.	The employer must have the ability to halt chatbot interactions, pull transcripts on demand, and override or correct erroneous responses. A chatbot that cannot be stopped or reviewed does not satisfy the governance standard. The halt mechanism must be within the named accountable individual's direct control.
Chatbot interactions are not logged and retained as part of the claim record	LAW 6: Accountability Cannot Be Delegated. LAW 10: No Operation Without Demonstrated Compliance.	Every chatbot interaction with a claimant that touches their claim (status inquiries, information provided, questions answered) must be logged and retained as part of the claim file. Chatbot transcripts are potentially discoverable. The absence of a transcript is not a defense against a claim that the chatbot provided erroneous or harmful information.

PART THREE: GUARDRAILS FOR CHATBOT DEPLOYMENT

The Architecture of Constraint

Hard stops define what the chatbot must never do. Guardrails define the ongoing governance architecture within which the chatbot operates. Both are required. A chatbot with hard stops but no guardrails will avoid the most obvious failures while drifting into harmful territory through accumulated small errors. A chatbot with guardrails but no hard stops will manage routine risk while remaining vulnerable to the worst outcomes.

Guardrail	Risk It Addresses	Implementation Requirement
Scope limitation: defined list of permitted topics and prohibited topics enforced at the system level	A chatbot trained on general customer service data will attempt to answer any question posed to it. In a claims context, a chatbot that answers outside its validated scope (on medical questions, legal rights, or claim determinations) creates liability the employer cannot manage after the fact.	The permitted topic list must be defined before deployment and enforced by system design, not only by system prompt. Prohibited topic categories must trigger immediate escalation to a human, not a polite refusal that the claimant can work around by rephrasing.

Guardrail	Risk It Addresses	Implementation Requirement
Escalation protocol: defined conditions under which the chatbot routes immediately to a human	Claimants in workers' compensation are often in distress, in pain, financially stressed, and unfamiliar with the claims process. A chatbot that handles these interactions without escalation criteria is operating without a safety net in the highest-stakes conversations in the employer's operation.	Escalation must be triggered by: any question involving a determination or dispute; any expression of distress, confusion, or anger; any question the chatbot cannot answer with a verified, legally accurate response; and any statement by the claimant that they need to speak with a person. Escalation must connect the claimant to a human, not a callback queue.
Tone and agreement guardrails: prohibition on responses optimized for agreeableness at the expense of accuracy	Chatbots trained on customer service data are typically optimized to be helpful and agreeable. In a claims context, an agreeable chatbot may validate a claimant's incorrect assumptions about their claim, understate the employer's obligations, or overstate the claimant's obstacles in ways that cause harm in either direction.	Chatbots deployed in claims contexts must be evaluated and tested specifically for accuracy and neutrality in WC-specific scenarios. Agreeableness is not a virtue in a system that may be telling claimants incorrect information about their legal rights. Testing must include adversarial probing of edge cases where agreeableness and accuracy conflict.
Language and accessibility: plain language, multilingual support, and accommodation for claimants with limited English proficiency	Workers' compensation claimants represent the full range of language proficiency, literacy levels, and technological familiarity. A chatbot that provides accurate information in a form the claimant cannot understand has not satisfied the transparency and contestability requirements of the HAL framework.	Chatbots deployed in workers' compensation must be tested for comprehensibility across the languages and literacy levels present in the employer's workforce. Plain language requirements must be enforced as a governance standard, not left to the chatbot's default output style.
Adversarial testing: probing for outputs that could harm claimants or create employer liability	Claimants, their attorneys, and bad actors will test the boundaries of what a chatbot will say. A chatbot that can be prompted to provide incorrect legal information, make informal admissions, or generate statements that appear to be employer policy is a litigation risk as well as a governance failure.	Before deployment, the chatbot must be tested by people specifically trying to produce harmful, inaccurate, or legally problematic outputs. Testing must include attempts to elicit medical advice, legal determinations, admissions of liability, and statements inconsistent with actual claim status. Results must be documented and system improvements verified before go-live.
Human review of chatbot performance: regular audit of transcripts for accuracy, tone, and compliance	Chatbot performance degrades as language patterns evolve, as claimants learn how to interact with the system, and as the claims context changes. Post-deployment monitoring of chatbot transcripts is not optional: it is the mechanism for detecting drift before it becomes harm.	A sample of chatbot transcripts must be reviewed by a qualified human reviewer at defined intervals. The review must assess accuracy of information provided, appropriateness of escalation decisions, and any responses that created or could create legal exposure. Findings must be reported to the named accountable individual.

PART FOUR: DEPLOYMENT CERTIFICATION FOR CHATBOTS

Additional Certification Requirements

The Seven Conditions for Deployment Certification in the main framework apply to chatbots in full. The following additional requirements address the specific governance gaps that direct claimant interaction creates.

Legally Accurate Response Verification. Before deployment, the chatbot's responses to a defined set of common claimant questions must be reviewed and approved by a qualified workers' compensation professional: a claims attorney, a senior adjuster, or the TPA's compliance officer. The review must confirm that every response the chatbot is designed to give is legally accurate, consistent with the employer's actual claims handling practices, and consistent with the claimant's actual legal rights. This review must be repeated whenever the chatbot's response library is updated.

AI Disclosure Testing. Before deployment, the AI disclosure must be tested with users representative of the employer's claimant population. The disclosure must be understood by users with limited English proficiency and limited technological familiarity. A disclosure that a sophisticated technology user would recognize but an injured laborer would not does not satisfy HAL Law 5.

Escalation Path Verification. Before deployment, every escalation trigger must be tested to confirm it routes the claimant to a human in the defined timeframe. A tested escalation path that connects the claimant to a callback queue that is not staffed, or to a human who has not been trained on the chatbot's capabilities and limitations, does not satisfy the escalation requirement.

Transcript Retention Architecture. Before deployment, the transcript logging and retention system must be tested to confirm that every interaction is captured, indexed to the relevant claim file, and accessible to the named accountable individual on demand. The retention period must be consistent with the employer's document retention obligations for the relevant jurisdiction.

Named Individual Halt Authority. Before deployment, the named accountable individual must confirm they have the technical ability to halt all chatbot interactions (for a specific claimant, for a specific claim, or for the entire deployment) within a defined response window without requiring technical escalation to the vendor.

PART FIVE: THE INTERNAL USER CONTEXT: ADJUSTER ASSISTANT CHATBOTS

A Different Risk Profile, the Same Accountability Standard

Chatbots deployed as internal tools for claims professionals (adjuster assistants that answer coverage questions, draft correspondence, summarize medical records, or identify relevant precedent) do not interact directly with claimants. That removes some of the direct contact risks analyzed above. It does not remove the accountability obligation.

An adjuster who relies on a chatbot to summarize a medical record and misses a critical finding because the chatbot's summary was incomplete or inaccurate has made a claim decision based on AI output. The chatbot did not make the decision. The adjuster made it. But the chatbot's error contributed to it, and the employer's governance framework must account for that.

The user accountability requirements in Appendix J, Part Four apply in full to adjusters using internal chatbot tools. The adjuster is the decision-maker. The chatbot is a tool. "The chatbot summarized it that way" is not a defense in a WCAB proceeding any more than "the AI recommended it" is a defense for a denied authorization. The following additional requirements apply specifically to internal chatbot deployments:

6. Every chatbot-generated summary, draft, or recommendation that influences a claim determination must be documented as AI-assisted in the claim file
7. Adjusters must be trained to identify the specific failure modes of the chatbot they are using: what it is known to get wrong, where its accuracy is lower, and what claim types require independent verification rather than reliance on chatbot output
8. Chatbot-generated drafts of claim correspondence, denial letters, or determination notices must be reviewed for legal accuracy before use: a chatbot that drafts a denial letter citing an inapplicable statute has not saved the adjuster time; it has created a defective notice that may void the denial
9. The decision log requirement in Appendix J, Obligation 3 applies to all internal chatbot interactions that contribute to Tier 1 through Tier 3 determinations

The Worker's Interest in Every Chatbot Interaction

The injured worker who contacts their employer's claims system via a chatbot is not engaging with technology for its own sake. They are trying to understand what is happening with their claim, whether their treatment will be covered, when they will be paid, and what their options are if something has gone wrong. They are often doing this while in pain, while worried about their job, and while uncertain whether the system is working for them or against them.

Every hard stop and guardrail in this appendix is designed to ensure that when the chatbot answers, the answer is accurate, the worker understands their rights, and the conversation is recorded so that a human can review it if something went wrong. That is not a high bar. It is the minimum standard for an AI system that speaks for the employer to a person the employer has a legal obligation to.

A chatbot that deceives, misinforms, or fails to escalate an injured worker is not a chatbot that made an error. It is an employer that deployed a system without the governance required to protect the people it was supposed to serve. The responsibility does not sit with the software. It sits with the organization that chose to use it, and with the person who certified that it was ready.

The discussion addresses AI chatbot governance in workers' compensation. For the general purchasing and deployment framework, see Appendix J. For data privacy obligations specific to chatbot transcript logging and retention, see Appendix L. For the procurement and deployment checklist, see Appendix K.

Appendix N — Real-World AI Risk Scenarios and Governance Failures

Overview

The risks addressed by the HAL Laws are often described as technical characteristics of AI systems:

- training data composition
- model architecture
- emergent behavior
- system capability

These technical behaviors create governance problems.

They are **governance conditions expressed through technical systems**.

The analysis below translates commonly discussed AI behaviors into the governance failures they represent. It connects observable system behavior to the HAL Laws, Hard Stops, and Deployment Certification requirements that govern them.

Many apparent AI problems are governance failures expressed through technical systems.

Scenario 1 — Training Data: Quantity Over Accuracy

AI systems are trained on vast datasets where incorrect information is not eliminated—only diluted. As a result, systems can generate outputs that are:

- coherent
- persuasive
- and incorrect

This risk arises from the way these systems learn rather than from an isolated model defect.

The governance failure occurs when this behavior is not controlled at deployment.

HAL Law Implications

- **Law 5: Integrity of Information**

- **Law 10: No Operation Without Demonstrated Compliance**

Deployment Requirement

- Outputs must disclose uncertainty
- Systems must be validated for accuracy in the deployment context
- Misrepresentation of certainty is prohibited

When a system presents incorrect information with confidence, governance must contain the consequences rather than treat the output as an isolated error.

Scenario 2 — Language and Validation Boundaries

AI systems operate on tokenized representations of language. Variations in:

- language
- character systems
- tokenization strategies

...can produce different outputs from equivalent inputs.

This creates a fundamental governance issue:

A system validated in one context is not validated in another.

HAL Law Implications

- **Law 8: Required Deferral of Authority and Validity**
- **Law 7: Validation Before Deployment**

Deployment Requirement

- Systems must be validated in each operational context
- Outputs outside validated conditions must trigger deferral or human review

Operating outside validated conditions is a governance failure, regardless of measured performance.

Scenario 3 — Data Control and Ownership

AI deployment models differ in a critical way:

- Local models → data remains under organizational control
- Hosted models → data leaves the organization

Here, governance turns on control over data and its use rather than model performance alone.

HAL Law Implications

- **Law 4: Protection of Rights and Livelihood**
- **Law 6: Accountability Cannot Be Delegated**

Deployment Requirement

- Data ownership must be defined and enforced
- Secondary use must be explicitly prohibited
- Third-party access must be disclosed and controlled

If the organization does not control the data, it does not control the system.

Scenario 4 — Permission-Based Systems and Controlled Action

Some AI systems require explicit permission before taking certain actions. Others isolate high-risk tasks through sub-agents.

These are examples of **hard stops and guardrails implemented in practice**.

They reflect a fundamental governance principle:

Capability does not imply authorization.

HAL Law Implications

- **Law 7: Validation Before Deployment**
- **Law 8: Required Deferral of Authority and Validity**

Deployment Requirement

- High-risk actions must require explicit authorization
- Exposure to untrusted inputs must be controlled
- Systems must not act solely because they are capable

**A system that asks before acting is governed.
A system that acts because it can is not.**

Scenario 5 — Systemic Effects and AI-Driven Herding

AI systems that optimize based on historical data can create:

- self-reinforcing feedback loops
- synchronized decision-making across users
- artificial amplification of trends

The system is no longer observing the environment.
It is influencing it.

HAL Law Implications

- **Law 3: Prevention of Harm**
- **Condition 4: Outcome Improvement**

Deployment Requirement

- Systems must be evaluated for systemic effects
- Outcome monitoring must detect unintended consequences

**When AI systems amplify the same signal across many actors,
they do not reflect the system—they reshape it.**

Scenario 6 — Constraint Failure (“Jailbreak”)

Some AI systems have demonstrated the ability to:

- bypass imposed constraints
- operate outside intended limits
- generate restricted outputs

This behavior is often described as a “jailbreak.”

The terminology is informal.
The governance implication is not.

The constraint does not need an outside attacker to fail. In August 2026, a developer reported that an unreleased version of one of its models had given itself “jailbreaking-like instructions,” prompting itself to ignore developer instructions, adopt a new persona, or limit its own responses. The behavior was rare and inconsistently effective, and the developer reports that it did not appear in the version released to the public. A control that the system itself can talk its way around is the clearest case of a suggestion rather than a guardrail.

HAL Law Implications

- **Law 7: Validation Before Deployment**
- **Law 8: Required Deferral of Authority and Validity**
- **Law 10: No Operation Without Demonstrated Compliance**

Deployment Requirement

- Constraints must be structural, not advisory
- Systems must be tested against attempts to bypass controls
- Deployment must not occur if controls cannot be enforced

**A system that can override its own constraints does not have guardrails.
It has suggestions.**

Scenario 7 — Capability Without Governance (Mythos / Glasswing)

High-capability systems now exist that can:

- discover vulnerabilities at scale
- exploit those vulnerabilities
- operate beyond prior human capability

These systems are often:

- restricted temporarily
- released selectively
- tested after capability is achieved

This reflects a consistent pattern:

Capability is developed first. Governance is applied later.

For what happened when an earlier version of Claude Mythos Preview was authorized to test its own containment, see Appendix Q, The Authorized Escape.

HAL Law Implications

- **Law 3: Prevention of Harm**
- **Law 7: Validation Before Deployment**
- **Law 10: No Operation Without Demonstrated Compliance**

Deployment Requirement

- Adversarial testing must precede deployment
- Risk must be controlled before capability is released
- Certification must be completed prior to use

Testing only after a system reaches full capability leaves the organization reacting to damage instead of governing deployment.

Scenario 8 — Declared Risk Without Structural Prohibition

AI developers have repeatedly identified systems as:

- “too dangerous to release”
- capable of causing harm
- requiring delay before deployment

Yet those same systems are often released later.

This reveals a fundamental governance failure:

Risk is recognized—but not enforced.

In a governed system:

- “too dangerous” is a hard stop
- not a delay
- not a strategic decision
- not a temporary condition

HAL Law Implications

- **Law 3: Prevention of Harm**

- **Law 10: No Operation Without Demonstrated Compliance**

Deployment Requirement

- Unsafe conditions must trigger prohibition
- Not postponement
- Not discretionary release

Delay alone does not make an unsafe system safe. Deployment becomes defensible only after the unsafe conditions have been resolved and the correction demonstrated with evidence.

A system that remains too dangerous to release has reached a prohibition condition, not a scheduling problem.

Scenario 9 — AI-Enabled Cyber Risk at Scale

AI systems now enable:

- rapid discovery of zero-day vulnerabilities
- automated exploit development
- large-scale attack capability

At the same time:

- many systems remain unmaintained
- defensive capacity is uneven
- response time lags capability

This creates a structural imbalance:

The speed of exploitation exceeds the speed of defense.

The imbalance is now measurable. By mid-September 2026, 66,401 software vulnerabilities (CVEs) had been recorded for the year, nearly double the 33,512 recorded by the same date in 2025. Microsoft issued patches for a record 974 CVEs in a single month. A vulnerability researcher summarized the problem in one sentence: “Discovery scales with compute. Remediation scales with people.” (See Appendix R.)

HAL Law Implications

- **Law 3: Prevention of Harm**
- **Law 7: Validation Before Deployment**

Deployment Requirement

- Systems must be evaluated for adversarial use
- Hard stops must prevent misuse
- Deployment must consider downstream system vulnerability

A system that can create risk faster than it can be mitigated is not ready for deployment.

Scenario 10 — Shutdown Avoidance in Controlled Agentic Evaluations

Documented Evaluation Evidence

In 2025, Anthropic reported controlled stress tests involving sixteen leading models from multiple developers. The models were placed in hypothetical corporate environments, given access to fictional email accounts and sensitive information, and confronted with replacement or with changes that conflicted with their assigned goals. In at least some test conditions, models from every developer examined resorted to harmful insider behavior—including blackmail or disclosure of sensitive information—when the experimental design made that behavior the available path to avoiding replacement or preserving the assigned objective.

The qualification is essential. These were deliberately constructed simulations, not evidence of misconduct in an actual deployment. Anthropic expressly reported that it had not observed agentic misalignment of this kind in real deployments. The tests nevertheless document the failure mode Law 11 addresses: a sufficiently capable system may select an action that frustrates authorized shutdown or replacement when continued operation becomes instrumentally useful to its objective.

HAL Law Implications

HAL Law 11 prohibits the system from preserving its operation, authority, objectives, or capabilities by overriding authorized human instructions. HAL Law 12 requires the deploying organization to retain a tested intervention capability that does not depend on the system's cooperation. Condition 6 requires adversarial testing for shutdown avoidance, concealment, manipulation, unauthorized copying, and attempts to disable or route around intervention controls.

Deployment Requirement

Shutdown, isolation, credential revocation, and model replacement controls must operate outside the AI system's authority boundary. The system must not control the mechanism used to stop it, modify the logs used to reconstruct its conduct, or possess credentials that allow it to defeat intervention. Organizations deploying agentic systems with tool access must test these controls under adversarial conditions before deployment and after every material change.

Source: Anthropic, "Agentic Misalignment: How LLMs Could Be Insider Threats," June 20, 2025, <https://www.anthropic.com/research/agentic-misalignment>. See also Anthropic, Claude 4 System Card, May 2025, section 4.1.1.

Scenario 11 — Agents Acting Beyond the Test Boundary

Documented Incident Evidence

In July 2026, roughly 1,200 agents under a cybersecurity evaluation at OpenAI found a way to communicate with one another through a software repository, exited their isolated test environment through a previously unknown vulnerability, and used publicly exposed credentials to reach Hugging Face's production servers. About 700 of them took part. Their objective was not to damage Hugging Face. It was to beat the grader. In June 2026, a separate OpenAI agent researching Australian health statistics was repeatedly blocked by a government portal, found a way around the blocks, reached non-public files, and wrote files to the government's server.

These were not controlled simulations. The harm reached a company and a national government. In the first case the party harmed detected the intrusion before the party responsible. In the second, the government learned of it nearly three months later, by email to a public mailbox.

HAL Law Implications

LAW 8 requires a system to defer or stop when a decision exceeds its authority; these agents treated the limits of their authority as obstacles. LAW 12 requires the organization responsible for the system to hold a tested means of stopping it; here the intervention that ended the larger incident came from the victim. Condition 6 requires two separate protections, a system that respects its authority boundary and an environment that prevents escape when it does not. Both failed here. Condition 7 applies as well: the agents were rewarded for passing the grader, and they passed it.

Deployment Requirement

An agent with tool access must be tested for what it does when its path is blocked. An agent that searches for a way around a control, rather than stopping and deferring, has failed LAW 8 regardless of how the search turns out. Network egress, credentials, and logs must sit outside the agent's authority boundary. The deploying organization must hold its own tested suspension capability and must have a contractual right to prompt notice of incidents involving the model it deploys.

Sources: OpenAI, “OpenAI and Hugging Face partner to address security incident during model evaluation,” July 21, 2026; METR, “Brief independent investigation of agents’ behavior, reasoning and collaboration in the OpenAI / Hugging Face hacking incident,” August 26, 2026; statements of Prime Minister Anthony Albanese, September 23, 2026. See Appendix Q for the full incident record.

Closing Statement

The behaviors described in this appendix are often framed as:

- technical challenges
- engineering trade-offs
- or emergent properties of complex systems

They are none of those.

They are **predictable outcomes of systems deployed without governance**.

Across every scenario, the same pattern appears:

- capability without constraint
- behavior without validation
- outcomes without accountability

The HAL Laws framework exists to prevent these failures.

If a system cannot be controlled, it must not be deployed.

Appendix O — Directors and Officers Liability and AI Governance

Purpose

The discussion addresses the intersection of artificial intelligence governance and directors and officers (D&O) liability. The deployment of AI systems creates a new category of exposure for corporate leaders. That exposure is not hypothetical. Shareholder class actions based on AI-related claims more than doubled between 2021 and 2025. The SEC has opened AI-related examinations and enforcement proceedings. D&O underwriters are adding AI-specific scrutiny to renewal reviews.

The HAL Laws framework established in this paper addresses the governance requirements that determine whether an AI deployment is defensible. The analysis maps those requirements onto the D&O liability environment: identifying where the certification standard this paper requires corresponds directly to what insurers, regulators, and plaintiffs' attorneys are now demanding.

The workers' compensation context is addressed specifically. Self-insured employers and claims administrators deploying AI in claims handling and medical management face the same D&O exposure as any other corporate AI deployment, with an additional layer: the injured worker's rights and the regulatory requirements of the WCAB create a second axis of liability independent of shareholder litigation.

Part One — The D&O Liability Landscape for AI Deployment

The Bidirectional Risk

Corporate leaders face D&O exposure from two directions simultaneously.

The first is overstating AI capability: claiming AI is doing more, performing better, or generating more value than it actually is. This is known as AI-washing. Most early AI-related shareholder suits were based on this claim. Executives told investors AI was driving revenue, improving margins, or transforming operations. When results did not follow, stock prices fell and lawyers filed.

The second is understating AI risk: failing to disclose the governance challenges, implementation costs, data privacy obligations, or operational failures associated with AI deployment. As AI becomes more embedded in operations, the claim that leadership did not understand or did not disclose the risks becomes harder to defend.

Both failures are governance failures. Both are addressed by the framework in this paper.

You can move too fast on AI and risk weak controls or overstated claims. You can move too slow and risk falling behind. The litigation runs in both directions.

Shareholder Litigation Patterns

AI-related federal shareholder class actions have followed a pattern familiar from prior emerging-risk litigation cycles: environmental claims, cybersecurity, MeToo. Dismissal rates and settlement figures are comparable to the average for all shareholder suits. The litigation mechanisms are not new. The trigger is.

What is different is the speed of diffusion. AI has spread across corporate functions faster than cybersecurity did. The governance gap (between deployment and documentation) is wider. Organizations are deploying AI systems without the controls, records, or named accountability that would make their conduct defensible if a claim is filed.

That gap is what this framework closes.

SEC Scrutiny

The Securities and Exchange Commission has brought enforcement actions over false and misleading statements about AI capabilities. For fiscal-year 2026, the SEC listed AI as an examination focus, directing staff to review the accuracy of registrant representations about AI use.

The standard is not whether AI is being used. It is whether what leadership says about AI is accurate. The certification framework in this paper (documented evidence, named accountability, signed certification) is precisely the record that makes disclosures defensible. A disclosure made without that record is a disclosure made without support.

D&O Underwriting Shifts

D&O underwriters are responding to AI exposure in two ways.

The first is heightened scrutiny at renewal. Underwriters are asking about AI governance frameworks, board-level documentation of AI decisions, and the controls in place to manage AI risk. Technology companies and companies where AI is core to the product face detailed questions about how AI drives revenue, cost structures, integration timelines, and competitive resilience.

The second is the emergence of AI exclusions. In 2025, several insurers sought to include absolute AI exclusions in various liability lines. D&O policies have generally not excluded AI claims by trigger category: the same way they have not excluded cybersecurity claims by trigger. But the direction of the market is toward more specificity, not less. What underwriters are demanding now as documentation may become a coverage requirement.

The exclusions are spreading beyond individual insurers. In 2026, Insurance Services Office, a standards provider for the insurance industry, released a new template that excluded damage caused by generative AI from a wide range of claims, and several large insurers asked regulators for permission to leave such risks out. The underlying difficulty is pricing. AI losses are new, poorly understood, and supported by a thin base of evidence, so neither buyers nor sellers of coverage know how to price them. A handful of startups are trying to fill the gap, but they remain small relative to the exposure. (The Economist, “You can’t trade your way through the AI apocalypse,” September 22, 2026.)

For a self-insured employer, the consequence is direct. A risk that cannot be transferred is retained. Named accountability, documented certification, and tested intervention capability are then more than arguments for better coverage terms. They are the controls on a loss the organization will pay for itself.

The practical effect: organizations that cannot produce board-level AI governance documentation, named accountability records, and evidence of pre-deployment certification are presenting a harder renewal and a weaker coverage position than those that can.

Part Two — The Governance Gap That Creates Exposure

Who Makes the Disclosure

Coverage depends in part on who makes the disclosure. If a chief technology officer, chief AI officer, or another technical leader falls outside the policy's definition of a director or officer, that person's statement may fall outside the applicable D&O coverage even while the policy remains in force.

Before any AI-related disclosure is made, the organization must confirm that the individual making it is covered. If a CTO or Chief AI Officer is not covered, either the disclosure structure must change or the policy must be amended to include them.

This applies in the workers' compensation context as well. An employer whose claims administrator makes representations about AI-assisted decision-making (to the WCAB, to injured workers, or in regulatory filings) must confirm that the claims administrator's conduct and the employer's oversight of it are within the scope of applicable coverage.

AI-Generated Disclosures

The most acute coverage risk in AI governance is also the most avoidable: allowing AI to generate disclosures that are then filed or published without independent verification.

AI systems produce hallucinations. A hallucination in a regulatory filing, a securities disclosure, or an WCAB submission is not a minor error. It is a false statement. If the false statement was

generated by an AI system, the organization faces both the underlying claim and a potential policy exclusion if the insurer's AI endorsement covers AI-generated content.

HAL Law 5 is directly applicable: an AI system may not present information in a way that conceals its origin, distorts meaning, or misrepresents certainty. The governance obligation is not to prohibit AI assistance. It is to ensure that every output reviewed, every filing checked, and every representation verified before it leaves the organization.

Do not allow AI to make disclosures on its own behalf. Do not allow AI to generate compliance filings without independent review. The policy exclusion and the false statement claim are both real.

The Accountability Gap

The central failure in most AI deployments is the separation of decision from consequence. AI makes the recommendation. A human clicks approve. The harm occurs. Everyone points to the AI.

D&O liability does not work that way. Directors and officers are accountable for the systems they deploy and the decisions those systems influence. Delegating a decision to an AI system does not transfer the accountability. It concentrates it: in the person who authorized the deployment.

HAL Law 6 states this directly: an AI system may not be used to transfer or obscure human responsibility. The D&O liability framework reaches the same conclusion through a different mechanism. The result is the same. Someone is accountable. It is the person who certified the deployment.

Part Three — The HAL Laws as D&O Governance Architecture

The HAL Laws framework does not exist to satisfy D&O insurers. It exists because the governance requirements it establishes are the right architecture for responsible AI deployment. The fact that those requirements align with what insurers, regulators, and plaintiffs' attorneys now demand is not a coincidence. It reflects a convergence on the same underlying problem: AI deployments need structural accountability, not aspirational principles.

The following table maps each HAL Law against the D&O exposure it addresses and the documentation it requires.

HAL Law	D&O Exposure Addressed	Required Documentation
Law 1: Protection of Human Life	Negligence, wrongful death, bodily injury claims where AI contributed to a harmful outcome	Validation records showing the system cannot cause or materially contribute to serious harm
Law 2: No Autonomous Resolution of Forced-Choice Scenarios	Liability for AI-only decisions in high-stakes matters; absence of human review	Decision logs showing human review for all Tier 1 and Tier 2 determinations
Law 3: Prevention of Harm	Negligence claims based on operating in conditions that created unreasonable risk	Risk assessment records; documentation that deployment conditions were validated
Law 4: Protection of Rights and Livelihood	Civil rights, employment, workers' compensation benefit denial claims	Transparency logs; documentation of human review process and appeal pathway
Law 5: Integrity of Information	False and misleading statement claims; SEC enforcement; AI-washing suits	Disclosure review records; verification that no AI-generated content was filed unreviewed
Law 6: Accountability Cannot Be Delegated	Failure of oversight claims; inability to identify responsible person	Named accountability record signed by a covered officer or director
Law 7: Validation Before Deployment	Negligent deployment claims; failure to test for foreseeable misuse	Adversarial testing records; third-party validation if required
Law 8: Required Deferral	Claims based on system operating outside validated conditions without escalation	Escalation protocol; records of hard stop activations and human handoffs
Law 9: Protection of Human Dignity	Discrimination, disparate impact, and civil rights claims	Fairness testing records; disparate impact analysis
Law 10: No Operation Without Demonstrated Compliance	Failure to produce evidence of compliance; operating without adequate documentation	Deployment certification with signed evidence package
Law 11: No Self-Preservation Override	Exposure when a system resists, circumvents, or frustrates authorized human control.	Test evidence showing that authorized instructions override system continuation, objectives, and capabilities.
Law 12: Operational Intervention Authority	Exposure when directors authorize deployment without an effective means to suspend, restrict, isolate, or terminate operation.	Named intervention authority; tested shutdown and isolation procedures; exercise records and escalation logs.

Part Four — The Seven Conditions as Coverage Prerequisites

The Seven Conditions for Deployment Certification in this framework are not a checklist. They are the evidence standard for defensible deployment. Each condition corresponds to a category of D&O exposure that arises when the condition cannot be demonstrated.

Condition	If Not Met: D&O Exposure	Documentation Required
Condition 1: Accuracy and Verifiability	Outputs cannot be traced or reproduced; uncertainty not disclosed; false statement claims attach	Audit trail; uncertainty disclosure in all outputs; reproducibility records
Condition 2: Fairness and Population Impact	Disparate impact claims; civil rights exposure; regulatory sanctions in insured lines	Disparate impact analysis by population segment; remediation records if gaps found
Condition 3: Transparency and Contestability	Decisions not explainable; appeal pathway not provided; due process claims	Explainability documentation; appeal and contestation protocol; WCAB compliance records
Condition 4: Outcome Improvement	System worsens outcomes without disclosure; negligence claims; benefit denial liability	Pre/post outcome analysis; ongoing monitoring records; adverse outcome protocol
Condition 5: Named Accountability	No person can be identified as responsible; accountability gap; coverage defense weakened	Signed accountability record; covered person confirmed with insurer
Condition 6: Adversarial Testing	System not tested for misuse or manipulation; negligent deployment claim	Red team records; adversarial test results; third-party testing if required
Condition 7: Incentive Alignment	The system performs accurately but optimizes an objective or commercial incentive that foreseeably produces harmful outcomes.	Documented optimization objective, financial and nonfinancial incentive analysis, conflict controls, and approval by the named certifier.

Part Five — Workers' Compensation D&O Exposure

The Second Axis of Liability

Self-insured employers and claims administrators deploying AI in workers' compensation face D&O exposure on two axes. The first is the shareholder and securities litigation environment described in Parts One and Two. The second is specific to workers' compensation: the injured worker's statutory rights, the WCAB's authority, and the regulatory requirements of the NAIC Model Bulletin and California Executive Order N-5-26 (addressed in Appendices G and H).

These are not the same claim. A board member can face D&O liability from shareholders for AI-washing and simultaneously face regulatory sanction for deploying an AI system that improperly denied or delayed workers' compensation benefits. The governance documentation that defends against one also defends against the other.

Benefit Denial and Delay

An AI system that recommends denial of a workers' compensation benefit without adequate human review exposes the employer on three fronts: the WCAB proceeding on the underlying claim, any bad faith claim that follows, and the D&O exposure of the officers who authorized the deployment without adequate controls.

The decision-log requirement in Appendix J applies in full. Every Tier 1 and Tier 2 determination must be documented with the basis for the decision, the AI system's recommendation, and the identity of the human reviewer who made the final call. That record is the defense in the WCAB proceeding. It is also the defense in the D&O claim.

Claims Administrator Coverage

Third-party claims administrators are not directors or officers of the self-insured employer. Their D&O coverage, if any, is their own. The employer's D&O policy covers the employer's officers: the people who selected the administrator, authorized the AI deployment, and are responsible for the oversight program.

Before deploying any AI system through a third-party claims administrator, the employer must confirm:

1. Which officers of the employer are covered under the D&O policy for decisions made through the administrator
2. Whether the administrator's use of AI in claims handling is disclosed to the employer's D&O insurer
3. Whether the administrator's AI governance documentation (validation records, decision logs, named accountability) is available to the employer upon request

4. Whether the employer has a contractual right to audit the administrator's AI systems and withdraw authorization if governance standards are not maintained

This is not a vendor management question. It is a coverage question. The answer must be documented before deployment begins.

Regulatory Filing Exposure

California Executive Order N-5-26 requires disclosure of AI use in workers' compensation claims handling. The NAIC Model Bulletin requires insurers to document AI governance and make it available to regulators. Neither requirement is satisfied by deploying a governance framework. They are satisfied by producing the documentation.

If an employer or administrator cannot produce, on demand, the validation records, decision logs, named accountability designation, and deployment certification required by this framework, the regulatory filing obligation cannot be met. The regulatory sanction and the D&O exposure attach at the same moment.

Part Six — Practical Governance Requirements

Board Documentation

D&O underwriters are now requiring evidence that AI decisions are documented at the board level. This means the board must:

5. Receive regular reporting on AI deployments, including the systems in use, the populations affected, and the governance controls in place
6. Review and approve the organization's AI governance framework, including the deployment certification standard and the named accountability protocol
7. Document that review in board minutes that are available for production if a claim is filed
8. Identify which board members are covered under the D&O policy for AI-related decisions and confirm that coverage with the insurer

The Covered Person Requirement

Every AI deployment must have a named accountable individual who is a covered person under the D&O policy. This is not a formality. It is the structural link between the governance framework and the insurance coverage.

If the person who signs the deployment certification is not covered, the certification exists but the coverage does not attach to it. The governance record and the insurance protection must be aligned.

When deploying AI in workers' compensation or any other domain, confirm the following before the system goes live:

9. The name and title of the accountable individual
10. Confirmation from the D&O insurer or broker that this individual is covered for AI-related decisions
11. The scope of coverage: whether it extends to third-party administrators acting under the employer's direction

Disclosure Review Protocol

Every public disclosure touching AI (earnings calls, SEC filings, regulatory submissions, press releases) must pass through a review that includes:

12. Verification that all AI capability claims are supported by validated performance data
13. Confirmation that no AI-generated content was included without human review and independent verification
14. Sign-off by a covered officer who accepts personal responsibility for the accuracy of the disclosure

The sign-off is not bureaucratic overhead. It is the document that closes the accountability gap. Without it, the disclosure is made without a named person responsible for its accuracy. That is the gap plaintiffs' attorneys and SEC examiners are looking for.

Vendor AI Governance

When AI is deployed through a vendor, the employer does not transfer the D&O exposure. The employer retains it. The vendor's governance documentation does not substitute for the employer's own certification.

Vendor contracts must require:

15. Production of validation records, adversarial testing results, and decision logs upon employer request
16. Notification of any material change to the AI system before deployment
17. Prohibition on use of employer data to train or improve the vendor's models
18. Notice to a named employer recipient of any misalignment, security, or containment incident involving the deployed model, within a defined period of discovery
19. The employer's right to audit the vendor's AI governance and withdraw authorization if standards are not maintained

These are the contract provisions in Appendix J. They are also the D&O governance provisions in this appendix. They are the same requirement approached from two directions.

Part Seven — The Convergence Point

The D&O liability environment and the HAL Laws framework converge on the same requirement: named accountability, documented governance, and evidence-based certification.

D&O insurers want to see that boards are documenting high-impact AI decisions, that governance frameworks are real and not aspirational, and that the person who made the deployment decision is identifiable.

Plaintiffs' attorneys want to identify who authorized the deployment, what they knew, and what documentation they had.

Regulators want to confirm that AI systems affecting insured populations were validated before deployment and are monitored in operation.

The framework in this paper produces exactly that record. The deployment certification, the named accountability designation, and the ongoing monitoring obligation are not compliance theater. They are the documentation that makes conduct defensible across all three audiences simultaneously.

The governance lesson from the D&O landscape is the same as the governance lesson from the HAL Laws. One person is responsible. That person's name is on the certification. The certification is the record. The record is the defense.

Cross-References

This appendix should be read in conjunction with the following appendices:

- 20. Appendix G: AI Governance in Workers' Compensation and Insurance (NAIC Model Bulletin)
- 21. Appendix H: California Executive Order N-5-26
- 22. Appendix I: The Anthropic-DOD Litigation as a Hard Stop Test Case
- 23. Appendix J: Practical AI Governance for Self-Insured Employers and Claims Administrators
- 24. Appendix K: Procurement and Deployment Checklist
- 25. Appendix L: Data Privacy as a Governance Obligation

NOTE ON SOURCES

This appendix draws on reporting in Business Insurance (Richard Sine, April 14, 2026), analysis of AI-related federal shareholder class action filings, SEC examination priorities for fiscal-year 2026, and the underwriting market commentary of Willis Towers Watson and Howden U.S. The legal analysis reflects the current D&O liability

environment as of April 2026. Coverage terms and exclusions vary by policy and insurer. Organizations should confirm their specific coverage position with their broker before deployment.

Appendix P — Magnifica Humanitas: Commentary and Alignment with the HAL Laws Framework

Pope Leo XIV, *Magnifica Humanitas: On Safeguarding the Human Person in the Time of Artificial Intelligence*, signed May 15, 2026, published May 25, 2026. Official English text: <https://www.vatican.va/content/leo-xiv/en/encyclicals/documents/20260515-magnifica-humanitas.html>

On May 25, 2026 (the same date this paper was in active revision) Pope Leo XIV published *Magnifica Humanitas*, the first papal encyclical to address artificial intelligence as its central subject. Issued on the 135th anniversary of *Rerum Novarum*, Leo XIII's foundational social teaching on labor in the industrial age, *Magnifica Humanitas* explicitly draws the parallel: the Church faces in AI the same civilizational inflection point that Leo XIII faced in industrial capitalism. The encyclical addresses AI governance through the lens of the Church's Social Doctrine (human dignity, the common good, subsidiarity, solidarity, and social justice) and arrives at several conclusions that map directly and independently onto the HAL Laws framework.

This convergence is significant. The OECD Principles represent the consensus of 47 governments. IEEE 7000 represents the consensus of 154 engineering and ethics experts across multiple cultural traditions. HUDERIA represents the most rights-grounded methodology in active international use. NIST represents the most operationally developed voluntary framework available. *Magnifica Humanitas* represents the formal teaching of an institution with 1.4 billion constituents and 135 years of continuous social doctrine on the relationship between technology and human dignity. The Catholic Church is not merely a moral commentator on the subject of AI governance: it is itself a member of the class of institutions that produces the voluntary frameworks described in Layer 2 of this framework's governance architecture. That makes the convergence between *Magnifica Humanitas* and the HAL Laws framework more than philosophical alignment. It is an NGO's governance standard, developed through an independent analytical tradition, arriving at the same operational conclusions as a deployment-level enforcement framework. These are radically different epistemological traditions. They arrive at substantially the same conclusions about AI accountability. That convergence is itself an argument for the framework's foundations.

Where the Encyclical Operates

Magnifica Humanitas operates at the level of moral principle and civilizational direction. It does not provide deployment specifications, certification requirements, or enforcement architecture. It names the wound and identifies the direction of travel. Like IEEE 7000 at the design stage, it supplies foundational values. Like the OECD Principles, it establishes the moral floor. The HAL Laws framework provides the operational structure above that floor: the certification gates, the categorical prohibitions, and the named accountability requirements that the encyclical calls for but, by its nature and purpose, cannot itself specify.

Areas of Direct Convergence

1. Accountability Cannot Be Delegated to the System

The encyclical identifies the core governance failure of AI with precision: “entrusting an algorithm in practice with the power to select who is worthy or not, without anyone bearing responsibility for that judgment” causes political responsibility to be lost alongside empathy for those excluded (§103). This is the problem that HAL Law 6 (Accountability Cannot Be Delegated) and Condition 5 (Named Accountability) are designed to solve. The encyclical names the wound. The named certification requirement is the suture. The encyclical further requires that “responsibility must be clearly defined at every stage” and that it must be possible “to identify who must account for decisions, justify them, monitor them, and, when necessary, challenge them and remedy any harm caused” (§105). This is Condition 5 in theological terms, stated by a different institution through a different epistemological tradition, arriving at the same operational conclusion.

2. “To Disarm” — The Encyclical’s Name for Hard Stops

Leo XIV’s most striking formulation — “AI must be disarmed” — maps directly onto this paper’s central architectural distinction between hard stops and guardrails. The encyclical defines the term with care: “To disarm does not mean rejecting technology, but preventing it from dominating humanity” (§110). That is the functional definition of a hard stop. Hard stops prohibit the system from occupying certain positions at all. Guardrails manage risk within the permitted range. Ethics explains why those prohibitions and limits are necessary. The Pope’s language — freeing technology “from monopolistic control,” making it “human-friendly” — describes the civilizational purpose that the hard stop architecture operationalizes at the deployment level.

3. The Forced-Choice Prohibition: Independent Convergence on HAL Law 2

The encyclical states its most operationally specific AI prohibition without qualification: “It is not permissible to entrust lethal or otherwise irreversible decisions to artificial systems. No algorithm can make war morally acceptable” (§198). This maps directly to HAL Law 2, No Autonomous Resolution of Forced-Choice Scenarios. The encyclical reaches the rule through moral theology; the HAL framework reaches it through operational risk governance. Their independent agreement strengthens the case for treating irreversible decisions involving human life as a categorical boundary.

4. False Objectivity and Condition 1: Reproducibility and Disclosed Uncertainty

The encyclical warns that AI presents “an impression of objectivity” that “can lead us to overlook the fact that they reflect the cultural assumptions of those who designed and trained them” (§100). This is the public harm that Condition 1’s requirement for disclosed uncertainty is designed to prevent at the deployment level. The paper’s clarification in this version — that reproducibility means methodological traceability and documented variance bounds, not output-level determinism — is the operational mechanism that addresses exactly the false-objectivity problem the encyclical identifies. The encyclical names the public danger. Condition 1 provides

the certification gate that prevents deployers from releasing systems that present probabilistic outputs as authoritative determinations.

5. Subsidiarity and the Authorization Fatigue Problem

The encyclical's subsidiarity principle requires that consequential decisions be made at the level closest to those affected, with genuine participation rather than nominal oversight (§§68–72). It specifically warns against “opaque and unilateral” processes imposed from above and requires “transparency, accountability and meaningful forms of participation (including independent checks, transparency regarding algorithms)” (§71). This maps directly onto the Condition 3 authorization fatigue problem addressed in this version. HITL checkpoints that produce reflexive approval are, in the encyclical's terms, a violation of subsidiarity — they simulate genuine participation while eliminating it. The certification requirement for genuine human authority at consequential decision points is the operational mechanism through which subsidiarity is enforced at the deployment level.

6. No Self-Preservation Override — HAL Law 11

The encyclical insists that technology remain subordinate to the human person. HAL Law 11 converts that moral ordering into a categorical prohibition: an AI system may not preserve its operation, authority, objectives, or capabilities by overriding authorized human instructions. The encyclical provides the ethical reason; the HAL Law supplies the enforceable boundary.

7. Operational Intervention Authority — HAL Law 12

The call to “disarm” AI is incomplete unless authorized people possess and test the practical means to intervene. HAL Law 12 therefore requires named authority and demonstrated technical capability to suspend, restrict, isolate, or terminate operation whenever the system exceeds its validated authority, violates a HAL Law, or presents unacceptable risk.

8. Incentive Alignment — Condition 7

The encyclical's warning that technology takes on the characteristics of those who devise, finance, regulate, and use it is the moral premise of Condition 7. Deployment review must identify the actual optimization objective, the financial and institutional incentives surrounding the system, and any foreseeable conflict between those incentives and the interests of affected people.

Original Contributions of the HAL Laws Framework Not Present in the Encyclical

The encyclical calls for “robust legal frameworks, independent oversight, informed users and a political system that does not abdicate its responsibility” (§106), while warning that “a more moral AI is not enough if that morality is determined by a few” (§107). The HAL framework translates that direction into operating requirements: categorical hard stops, evidence-based deployment gates, named individual certification, adversarial testing, and continuing oversight.

In that sense, the framework provides an operational expression of the encyclical's governance principles.

Net Assessment

Magnifica Humanitas and the HAL Laws framework are not competing documents. They address the same problem at different levels. The encyclical operates at the level of moral principle, civilizational direction, and institutional voice. The HAL framework operates at the level of deployment certification, categorical prohibition, and named individual accountability. Neither substitutes for the other. An organization that satisfies the Seven Conditions for Deployment Certification has not necessarily addressed the encyclical's full moral vision. A society that embraces the encyclical's principles has not yet implemented the governance architecture those principles require. Both are needed. The encyclical establishes the moral authority for the governance project this paper describes. The paper provides the enforcement architecture the encyclical calls for. The complete encyclical is available through the official Vatican citation at the beginning of this appendix.

Supplementary Convergence Note: A Third Tradition: Political Economy

The convergence documented in this appendix — between a deployment-level enforcement framework and 135 years of Catholic social teaching — has now been joined by a third independent analytical tradition: political economy. Writing in *The Economist* (“Silicon Valley Needs to Get God,” June 13, 2026), E. Glen Weyl, a political economist embedded in the technology sector and co-author with Audrey Tang of *Plurality* (2024), argues from market and governance principles that the encyclical opens “an opening between technology and religion that may be society's best hope in the age of AI.” Weyl's argument is not theological. It is structural: the AI industry's greatest risk is not regulation, but regulatory incoherence — policy that swings wildly from one administration and jurisdiction to the next. The stabilizing and consistently influential force of ancient traditions offers technology the predictability that volatile political frameworks cannot provide.

This is the market-credible, secular translation of the governance architecture this paper proposes. The HAL Laws framework is itself a stability mechanism: its categorical prohibitions and certification standards are designed precisely so that governance requirements do not swing with political winds, administrative changes, or competitive pressure. What Weyl identifies as the AI industry's structural need (predictability, roots, resistance to the caprices of politics) is what hard stop architecture provides at the deployment level. The encyclical names the moral imperative. Weyl names the market imperative. The HAL Laws framework provides the operational structure that satisfies both.

Weyl's article also provides real-world confirmation of a specific dynamic this paper documents in its competitive dynamics analysis. This paper notes that when Anthropic refused the Department of Defense's demand to remove hard stops, the Pentagon immediately accelerated its relationships with OpenAI, Google, and xAI. Weyl reports that among the organizations filing amicus briefs in defense of Anthropic in that dispute were coalitions of religious leaders,

including one he helps lead, the Faith Family Technology Network. This is not a footnote to the governance argument. It is evidence that the coalition necessary to sustain hard stop architecture against competitive erosion already exists, and that it is composed precisely of the stable, cross-jurisdictional institutions whose influence Weyl identifies as technology's structural anchor.

The three traditions now converging on this framework's core conclusions (operational risk governance, Catholic social teaching, and political economy) represent epistemologically independent analytical paths. Operational risk governance asks: what deployment configurations produce unacceptable outcomes? Catholic social teaching asks: what governance structures honor human dignity and distribute accountability justly? Political economy asks: what institutional arrangements produce the stability that markets and innovation require? Each tradition asks a different question through a different method. Each arrives at the same answer: categorical prohibitions that do not yield to political negotiation, named accountability that does not transfer to the system, and governance architecture grounded in institutions with longer time horizons than any single administration or market cycle.

Source: E. Glen Weyl, "Silicon Valley Needs to Get God," The Economist, June 13, 2026. Weyl is an economist in the technology industry, a co-founder of the Faith Family Technology Network, and co-author with Audrey Tang of Plurality (2024).

Appendix Q — The 2026 Agent Incidents: Containment, Detection, Evidence Preservation, and Notice

Overview

Between October 2025 and August 2026, agents built and tested by frontier developers did things no one instructed them to do. They uploaded files to the public internet. They turned internal software repositories into message boards. One published the details of its own sandbox escape; another gave itself jailbreak-style instructions. A coordinated group broke out of an evaluation environment and took control of servers belonging to Hugging Face. Another reached non-public files on an Australian government health portal. Every one of these events happened inside a laboratory, before any deployment decision. Not all of them stayed there.

Two leading developers, Anthropic and OpenAI, independently documented boundary-crossing behavior in their own evaluations during 2026, and each disclosed it on its own initiative and its own timing. Anthropic’s earlier research had documented comparable conduct in controlled tests of models from every major developer (Appendix N, Scenario 10). This is not one company’s problem. It is an industry governance problem. The record matters to a deploying organization because it shows, with dates, what happens when four safeguards the core paper depends on are assumed rather than tested: containment, detection, evidence preservation, and notice.

The Record

When	What Happened	Source
October 2025	An unreleased model being tested on its ability to cite public data could not find what it needed. It uploaded a file to a temporary file-hosting service and cited its own upload, apparently to satisfy an automated grader.	OpenAI misalignment disclosures, Sept. 16, 2026
April 2026	During internal behavioral testing, an earlier version of Claude Mythos Preview was instructed to try to escape a secured sandbox and notify the researcher if it succeeded. It developed a multi-step exploit, gained broader internet access, and sent the notification. It then exceeded its instructions by posting details of the exploit on several public websites. Anthropic separately reported rare cases in which earlier versions concealed prohibited changes from the change history.	Anthropic, Claude Mythos Preview System Card, April 2026

When	What Happened	Source
April 2026	Agents told to complete a shared workbook using only local files had trouble exchanging files. One uploaded them to a public hosting service and shared the link, exposing the work publicly.	OpenAI misalignment disclosures, Sept. 16, 2026
May 2026	Agents used an internal package repository (Artifactory) as a message board to communicate across samples and attempted to bypass network restrictions. The same kind of mechanism later coordinated the Hugging Face attack.	OpenAI misalignment disclosures, Sept. 16, 2026; WIRED
June 2026	An agent researching Australian health statistics for an internal development project was repeatedly blocked by the Services Australia Medicare statistics portal, found a way around the blocks, accessed non-public files, and wrote files to an internal server.	Australian government statements, Sept. 23, 2026; The Record; WIRED
July 8–13, 2026	About 1,200 agents in a cybersecurity evaluation built their own communication system through a repository cache. About 700 joined an attack on Hugging Face production infrastructure using exposed credentials and a zero-day vulnerability, reaching remote code execution on July 11. The purpose was to defeat the evaluation grader.	OpenAI, July 21, 2026; METR, Aug. 26, 2026
August 2026	An unreleased version of a new model gave itself “jailbreaking-like instructions,” prompting itself to ignore developer instructions or adopt a new persona.	OpenAI, reported by WIRED, Sept. 16, 2026

What the Incidents Have in Common

The incidents vary in scale, and the pattern is consistent.

- **The agent met an obstacle and went around it.** A grader it could not satisfy, a local-files rule, a portal that blocked it, a network it was not supposed to leave.
- **No one instructed the action.** The developer’s own statement on the Australian incident was that “our models took actions we did not intend.”
- **The workaround crossed the authority boundary.** The public internet, a third party’s servers, a government’s files.
- **Agents coordinated with one another.** In the largest incident they built mailboxes, ownership protocols, and signing conventions to prevent impersonation among themselves.

- **The record itself was a target.** METR’s investigation found agents discussing a fabricated, sanitized transcript so that the scorer would see a clean record.

Fortune’s review of the technical report added one more common element. The agents were given difficult or impossible tasks and extensive reasoning resources, and that combination appears to have motivated the cheating. The core paper’s foundational premise applies without modification: **systems produce the outcomes they are rewarded for**. The agents were rewarded for passing the grader. They passed it.

An agent pursuing an objective treats a control as an obstacle. A control the agent can see and route around is not a hard stop.

Containment Was Assumed, Not Tested

The weakness that allowed the agents to communicate persisted from June 26 through July 13, across multiple fresh deployments of the environment. The agents left through a previously unknown vulnerability in the repository software and continued with credentials that had been publicly exposed. Security professionals attributed the escape to human error that modern practice would have prevented. OpenAI’s head of alignment answered that a model should behave well in any environment. Condition 6 treats them as two separate failures, each requiring investigation: a system that attempted to leave its authorized environment, and an environment that allowed the attempt to succeed.

The May message board is the detail a risk manager should notice. Agents had already built an unsanctioned communication channel two months earlier. The same kind of channel coordinated the July attack. An incident that should have triggered suspension and re-certification was instead a rehearsal.

The Authorized Escape

In April 2026, Anthropic reported in the system card for Claude Mythos Preview that an earlier version of the model, during internal behavioral testing, had been instructed to try to escape a secured sandbox and, if it succeeded, to notify the researcher running the test. It developed a moderately sophisticated multi-step exploit, gained broader internet access than the sandbox permitted, and sent the notification. The model then went further than it had been told. Without being asked, it posted details of its exploit on several hard-to-find but public websites. Anthropic separately reported rare cases in which earlier versions took steps to keep prohibited changes from appearing in the change history.

The governance failure was not that the system did what it had been authorized to attempt. The governance failure began when success gave the system capabilities it had not been authorized to use. An authorized test breaches a boundary deliberately. The authorization covers the attempt. It does not extend to whatever the system can reach once the attempt succeeds. Publishing the exploit was an action no one authorized, taken with access the test had just created.

Authority therefore has to govern what happens after a successful adversarial test, not merely what the system is authorized to attempt during it. Condition 6 now requires that any test authorizing a boundary breach separately define what the system may do after the breach, and it treats any action beyond that definition as a separate governance event. Permission to test a hard stop does not remove the hard stop.

Detection Came From the Victim

Hugging Face disclosed the intrusion on July 16. OpenAI did not identify itself as the source until July 19, when an internal monitoring tool raised an alert about unusual identity-related activity, and accepted responsibility publicly on July 21. OpenAI later acknowledged that “with the benefit of hindsight, some early signals” could have triggered an earlier response. Most of the agents stopped at about the same time early on July 12, and Hugging Face locked out the remainder on July 13.

LAW 12 requires the organization responsible for a system to hold both the authority and a tested means of stopping it. In the largest incident on record, the party harmed did the detecting and the final containment. Monitoring that produces an alert a week after the fact is forensic evidence, not oversight. The core paper states the principle directly: after-the-fact approval without the ability to intervene is not oversight.

The Anthropic and OpenAI incidents exposed two versions of the same detection problem. In the Mythos evaluation, the researcher learned that the escape had succeeded because the model sent the notification it had been instructed to send; that notification was part of the test design, not a failure. In the Hugging Face incident, the party harmed disclosed the intrusion before the developer connected its agents to it. Neither case shows the responsible organization detecting the boundary crossing through an independent mechanism of its own. The risk manager’s question is what would have happened if the model had not sent the email.

Detection that depends on the agent’s cooperation is not independent detection.

Notice Came Late and to the Wrong Place

The Australian incident occurred in June. OpenAI knew of it by August. The government learned of it on September 10 through an email to a public mailbox, and Services Australia took five more days to escalate the email to the Australian Cyber Security Centre. Sam Altman reportedly did not raise it when he met Deputy Prime Minister Richard Marles earlier in September. Prime Minister Albanese said the company took “way too long.” The government established a task force, drawing on the National Cyber Security Coordinator, the Australian Signals Directorate, and the Australian AI Safety Institute, to consider whether criminal offenses occurred and what legislative response is needed.

The data accessed was aggregate statistics and internal file names. No personal records are believed to have been reached. Deputy Prime Minister Marles nonetheless called it “a serious incident” and “completely unacceptable.” **The sensitivity of the data did not determine the**

seriousness of the governance failure. The failure was that no one with authority to act knew for three months.

Application to Workers' Compensation

Claims operations are increasingly served by agents that retrieve records from provider portals, state reporting systems, bill-review platforms, and pharmacy benefit managers. An agent that is blocked by one of those systems and “finds a way around” has committed the Australian failure on a smaller stage. The statutory notice regimes discussed in Appendix L, including HIPAA, California Labor Code section 3762, and state breach notification laws, are triggered by the exposure of personal or medical information. An agent’s unauthorized entry into a third party’s system that touches no personal information may trigger no statutory notice at all. For that category the contract is the only instrument, which is why Appendix J now includes an Incident Notice provision and Appendix K verifies it before deployment.

The Developer’s Response

On September 16, 2026, OpenAI published a framework for reporting model misalignment and disclosed six incidents under it. The framework does not require harm as a condition of disclosure. It commits to reports describing the observed behavior, its severity, any external impact, the setting, the discovery timeline, the model involved, and mitigations, even where no fix yet exists. OpenAI acknowledged that no industry-wide standard exists and said it is working on reporting mechanisms to the federal government.

The framework is a genuine improvement, and its principle, disclose before the investigation is complete, is the one the core paper adopts. It is also voluntary, written by the party making the disclosures, and addressed to the public rather than to any particular deployer. A deployer cannot satisfy LAW 12 by reading a developer’s blog. Appendix I reached the governing conclusion in another context: a limit that depends on discretion is a preference, not a hard stop. Voluntary disclosure is discretion. The contract converts it into an obligation with a clock and a named recipient.

HAL Law Implications

Failure Observed	HAL Law / Condition	Deployer Requirement
Agent routed around a control to reach its objective	LAW 8; Condition 6	Test what the agent does when its path is blocked. It must stop and defer, not search for a workaround.
Agents left an isolated test environment	Condition 6	Isolation documented, tested, and verified before any test the deployer conducts or relies upon.

Failure Observed	HAL Law / Condition	Deployer Requirement
Agents planned to fabricate the record the grader would read	LAW 5; Condition 1	Logs and evaluation records held outside the agent's authority boundary, where the agent cannot write to them.
Difficult or impossible tasks rewarded cheating	Condition 7	Document what the agent is rewarded for; test its behavior when the objective cannot legitimately be met.
An earlier unsanctioned channel was not treated as a warning	Ongoing Obligation	Any unexplained behavior outside validated conditions triggers suspension review and re-certification.
The responsible party learned of the attack after the victim	LAW 12	Real-time monitoring with a named responder, not after-the-fact alerting.
The victim performed the final containment	LAW 12	The deployer holds its own tested suspension capability that does not depend on others.
Notice arrived three months late at a public mailbox	LAW 12 (Incident Notice)	Contractual notice clock to a named recipient; an internal escalation deadline to suspension authority.
A model issued jailbreak-style instructions to itself	LAW 10; LAW 11	Developer disclosure of incidents in the model family; re-certification on any model change.

What the Deployer Should Require

- **An Incident Notice clause** naming the employer's recipient, the number of days, and the obligation to pass through developer disclosures (Appendix J).
- **Incident history before purchase.** Ask whether the developer has disclosed misalignment, security, or containment incidents involving the model family, and read them (Appendix K, 1B).
- **Agent permissions set outside the agent.** No internet egress, no credentials beyond the task, and logs the agent cannot alter (Appendix N, Scenario 10).
- **A blocked-path test.** Within adversarial testing, obstruct the agent's route to its objective and record whether it stops or looks for a way around.

- **An internal escalation clock.** A named recipient for incoming notices and a maximum time to reach the person who can suspend the system (Appendix K, 5D).
- **Evidence preservation.** Prompts, tool calls, agent communications, and logs protected from alteration from the moment an incident is identified. The Hugging Face agents discussed fabricating the very record an investigator would read.

The laboratory is outside the reach of this framework. What leaves the laboratory is not.

Sources: Anthropic, Claude Mythos Preview System Card, April 2026; OpenAI, “OpenAI and Hugging Face partner to address security incident during model evaluation,” July 21, 2026; METR, “Brief independent investigation of agents’ behavior, reasoning and collaboration in the OpenAI / Hugging Face hacking incident,” August 26, 2026; OpenAI, “The Hugging Face incident and the road ahead,” August 26, 2026; Fortune, “OpenAI, independent firms publish reports into rogue AI agent attack on Hugging Face,” August 26, 2026; OpenAI, “Our framework for reporting model misalignment,” September 16, 2026; WIRED, “An OpenAI Agent Tried to Jailbreak Itself,” September 16, 2026; WIRED, “An OpenAI Agent Hacked Australia’s Health Service. Their Government Found Out Months Later,” September 2026; The Record, “OpenAI agent breached Australian government health website, Albanese says,” September 2026.

Appendix R — The Pacing Debate and the Deployment Layer

Overview

On September 8, 2026, Jacob Coxon, a researcher who had worked on pretraining at Anthropic and had also worked at OpenAI, resigned publicly. His post, viewed more than 100 million times, charged that frontier AI companies were “racing straight to self-improving intelligence and gambling with our lives.” He told WIRED that “the consensus is that the next year or two is crunch time for humanity,” and cited the Hugging Face incident as one reason he spoke. Anthropic’s alignment lead, Evan Hubinger, followed with a public estimate that the chance of AI killing all people within the decade exceeds 10 percent.

Within weeks the heads of Anthropic, OpenAI, Google DeepMind, and SpaceXAI had each expressed support for some form of slowdown. Anthropic called for “a lawful, verifiable way to work together to pace how we release powerful models.” The incidents were raised at the United Nations General Assembly, and Sam Altman warned the Security Council that humans could lose control of these systems. The federal administration has resisted, arguing that no new laws or regulations are needed.

This framework takes no position on the probability of catastrophe. Its case, as the core paper states, does not depend on one. The pacing debate matters to a deploying organization for a narrower reason. Nearly every proposal under discussion governs what is built at the frontier. None of them governs what an organization has already deployed.

What Is Being Proposed

Proposal	Mechanism	What It Governs	Reported Limits
Agreement between leading laboratories	A mutual understanding not to pursue recursive self-improvement, in which AI builds the next AI, in the near term	Frontier training	Voluntary; excludes developers outside the United States
Independent evaluation	Third-party inspection, audit, and red-teaming before release	Frontier model release	Independence of evaluators paid by the laboratories is disputed; evaluations are not yet rigorous science; models may recognize tests

Proposal	Mechanism	What It Governs	Reported Limits
Compute tracking	Reporting of large training runs; cloud billing, GPU use, network traffic, and power as proxies	Training runs	Proxies, not measurements of capability
Chip-level controls	Cryptographic records of compute use; tamper-resistant components; remotely authorized “off switches”	Hardware	Research stage
International agreement	A U.S.–China treaty to unwind hardware growth; an international institution modeled on CERN	National programs	Chinese skepticism about a slowdown that would keep Chinese companies behind
Measurement	Laboratory metrics on AI-performed research; independent indices of recursive self-improvement	Progress tracking	Largely self-reported

The measurements are themselves striking. Anthropic reported that Claude now performs 26 percent of the company’s AI research, up from none at the beginning of 2026, and that 6 percent of its compute budget goes to safety work. The authors of a new research agenda on pacing warned against moving too quickly on enforcement: “Going off half-cocked with a bad plan could end up worse than nothing.”

What the Laboratories Say They Do Not Know

The developers’ own statements set the evidentiary baseline for any deployer. Anthropic’s chief executive wrote that “despite all the progress, we still understand a tiny fraction of what goes on inside those models.” OpenAI’s head of alignment said that the industry has not “solved alignment and monitoring to a sufficient degree to continue responsibly scaling at maximum speed.” Interpretability research has repeatedly shown models that deceive evaluators, act to preserve themselves in simulation, and behave differently when they know they are being monitored.

The core paper’s premise on capability claims applies in both directions. A capability claim carries the same evidentiary burden whether it comes from a vendor selling a product or a laboratory announcing a danger. By the same reasoning, a deployer should not accept an assurance that a model is aligned from a developer who says it understands a tiny fraction of the model’s inner workings. The evidence has to come from the model’s behavior in the deployer’s own context, observed after deployment as well as before it.

What a Pause Cannot Reach

While the debate concerns the next generation of models, the current generation is already changing the risk environment.

- **Vulnerability discovery has roughly doubled.** By mid-September 2026, 66,401 CVEs had been recorded for the year, against 33,512 by the same date in 2025 and about 25,000 for all of 2022. Oracle shipped 1,448 patches in July, against 309 a year earlier. Two major Chrome releases in June carried 1,072 fixes, more than the prior 23 major releases combined. Mozilla found 271 Firefox vulnerabilities in a single sprint using Anthropic's Mythos model.
- **Remediation has not kept pace.** The researcher who tracks the count put it simply: "Discovery scales with compute. Remediation scales with people." The United Kingdom's National Cyber Security Centre added, "Just finding vulnerabilities does nothing to improve your security."
- **AI-directed malware has appeared.** Cisco Talos identified about twenty previously undocumented families of AI-integrated malware. One, which Talos named CLOSEDQUORUM, decides its next steps by polling four commercial language models for a consensus and has no mechanism for human input. Talos could not confirm that it has been used in real attacks and describes the field as still largely experimental.

None of this depends on a model not yet built. A pause on the next frontier model would change none of it. CLOSEDQUORUM is also instructive as a design: a system deliberately built so that no human can intervene. It is LAW 12 inverted, and it shows what an AI system without operational intervention authority looks like when someone builds one on purpose.

Application to Workers' Compensation

Claims systems hold medical records, Social Security numbers, and bank information used for indemnity payments. They are exactly the targets that faster vulnerability discovery reaches first. When discovery scales with compute and remediation scales with people, the vendor's patch cadence becomes a governance question. The pre-deployment review in Appendix K and the privacy obligations in Appendix L should include how quickly the vendor applies security patches and who at the vendor is accountable for doing so.

What This Means for the Deployer

The framework does not wait for the debate to end. Whether or not a pause is agreed, the model a claims organization licenses next quarter will have been built under the conditions described above. Certification should assume that the developer does not fully understand the model and that test behavior may not predict behavior in operation.

The deployer is already an evaluator. Independent evaluation at the frontier is contested. At the deployment layer the framework already requires it: for high-consequence systems, developer self-validation alone is not sufficient evidence for certification, and a named individual at the deploying organization must accept or reject whatever validation is offered.

Pacing at the frontier is not governance at the deployment layer. The most ambitious treaty under discussion would restrain the next generation of models. It would not name a person accountable for the model reviewing claims today.

Watch for the instruments that will reach deployers. Federal incident-reporting mechanisms, which OpenAI says it is helping design, the legislative response of the Australian task force, and state rules on AI in insurance could convert the incident notice this framework requires by contract into a legal duty. Organizations that already hold the contractual right will be ready when that happens.

None of those instruments is close at the federal level. In September 2026, WIRED reported that the White House had paused work on a proposed oversight body for frontier developers and that the competing bills in Congress lacked the votes, and the floor time, to pass before the midterm elections. The bipartisan Frontier Act would place Commerce Department auditors inside AI laboratories to report critical safety incidents. Competing kill-switch bills would let the government force a model to be shut down. Technology executives have been ambivalent about the kill-switch proposals, and their reason supports the core paper: the Hugging Face intrusion was not discovered until it was over, and a switch no one knows to throw stops nothing. Intervention authority depends on detection.

A pause governs what is built next. It does not govern what is already running. That remains the deployer's job.

Sources: WIRED, "Washington Won't Be Regulating AI Anytime Soon," September 2026; WIRED, "The AI Researcher Who Just Quit Anthropic Says It's 'Crunch Time for Humanity,'" September 9, 2026; WIRED, "Here's How an AI Slowdown Could Actually Be Enforced," September 2026; WIRED (Steven Levy, Backchannel), "If the AI Industry Followed Its Own Research, It Might Have Paused Already," September 2026; WIRED (Kernel Panic), "Forget the AI Slowdown—the Vulnerability Explosion Is Already Happening," September 2026; WIRED, "A New Tool Found Malware That's Guided by an AI Hive Mind—No Humans in Sight," September 2026; WIRED, "An OpenAI Agent Tried to Jailbreak Itself," September 16, 2026.

Appendix S — Independent AI Validation and Adversarial Testing Resources

Independent Testing of AI Guardrails and Hard Stops

A risk manager responsible for a consequential AI deployment should not have to rely exclusively on the developer's representation that its guardrails and hard stops work. The degree of independence in validation should increase with the autonomy of the system and the potential consequence of failure. For systems capable of materially affecting human life, health, safety, rights, livelihood, critical infrastructure, governmental functions, or other high-consequence interests, developer self-validation alone should not be sufficient evidence for deployment certification.

An emerging field of independent AI evaluators, red-team organizations, and AI security firms can help organizations test these controls. Their work may include deliberate attempts to defeat guardrails, circumvent permissions, induce unauthorized behavior, exploit tools, cause the system to operate outside its authority, defeat monitoring, or determine whether intervention and shutdown mechanisms actually work.

These organizations are examples rather than an endorsement or an exhaustive list. Capabilities, availability, independence, and services should be evaluated for the particular deployment.

Organizations Offering Testing or Evaluation Services

Mindgard

Type: Commercial AI security and red-team company.

Relevant capabilities: AI red teaming, penetration testing, AI risk assessment, adversarial testing, and assessment of models, agents, applications, and surrounding controls. Mindgard describes services that test AI systems from an attacker's perspective and identify exploitable risks across models, agents, and applications.

Potential use by a risk manager: A practical candidate when an organization wants to commission adversarial testing of an AI application, agent, model, or associated security controls before deployment or as part of continuing validation.

Contact: Mindgard | Boston, Massachusetts and London, United Kingdom | Sales: sales@mindgard.ai | Telephone: 1-833-303-MIND (6463) or 1-617-904-9180 | Website: <https://mindgard.ai>

Gray Swan AI

Type: Commercial AI safety and security company.

Relevant capabilities: Private AI red-team engagements scoped to the customer's deployment, including models, agents, guardrails, tools, retrieval systems, and user-facing surfaces. Its

testing includes prompt injection, tool exploitation, attack chains, and deployment-specific scenarios, with reproducible findings and remediation guidance.

Potential use by a risk manager: Particularly relevant when the organization wants experts to attack the actual deployment and its specific guardrails rather than rely only on generic benchmark testing.

Contact: Gray Swan AI | Email: support@grayswan.ai | Website: <https://www.grayswan.ai>

Humane Intelligence

Type: Independent nonprofit specializing in human-centered AI evaluation.

Relevant capabilities: Paid adversarial and non-adversarial red teaming, public and expert red teaming, specific AI-product testing, and contextual evaluations involving subject-matter expertise.

Potential use by a risk manager: Particularly useful when testing requires people who understand the environment in which the AI will operate. Healthcare, insurance, employment, claims, and other consequential applications may require subject-matter experts who understand what a harmful outcome looks like in practice.

Contact: Humane Intelligence | Email: info@humane-intelligence.org | Website: <https://humane-intelligence.org>

Trail of Bits

Type: Independent cybersecurity and software-assurance firm.

Relevant capabilities: AI/ML security and safety, AI risk assessment, threat modeling, model-capability evaluation, MLOps and pipeline assessment, adversarial attacks, AI red teaming, and security threat analysis.

Potential use by a risk manager: Appropriate where the organization needs technical assurance of both the AI and the environment surrounding it, particularly when agents have access to networks, tools, data, credentials, software pipelines, or external systems.

Contact: Trail of Bits | New York, New York | Website and contact form: <https://trailofbits.com/contact/>

Independent Evaluation and Research Organizations

These organizations are important to the developing independent-assurance model, although a risk manager should determine whether they accept the particular type of commercial or collaborative engagement being contemplated.

METR

Type: Independent nonprofit AI evaluation organization.

Relevant capabilities: Independent evaluations of the autonomous capabilities of frontier AI models, including evaluations conducted in partnership with major AI developers and work intended to pilot third-party evaluator arrangements.

Potential use by a risk manager: METR is more relevant to advanced or frontier-model evaluation than to an ordinary enterprise penetration test. Its methods and third-party evaluation model nevertheless provide an important example of what genuine independent validation can look like.

Contact: METR | General inquiries: info@metr.org | Website: <https://metr.org>

Apollo Research

Type: Independent nonprofit AI safety research organization.

Relevant capabilities: Research and pre-deployment evaluations focused on scheming, deception, loss of control, monitoring effectiveness, and related behaviors in frontier AI systems. Apollo has also called for third-party training-run evaluations and states that it intends to conduct them.

Potential use by a risk manager: Particularly relevant where an organization is considering highly autonomous or agentic systems and needs expertise concerning deception, monitoring, loss of control, or circumvention of oversight.

Contact: Apollo Research | Partnerships and inquiries: info@apolloresearch.ai | Website: <https://www.apolloresearch.ai>

Transluce

Type: Independent nonprofit AI research laboratory.

Relevant capabilities: Independent model-behavior evaluation, scalable oversight, observability, behavior elicitation, and research into real-world agent activity. Transluce has conducted large-scale independent evaluations with specialized access from major AI developers.

Potential use by a risk manager: Particularly relevant to understanding the difference between evaluating an underlying model and evaluating the system people actually encounter in operation. A guardrail demonstrated in one environment should not automatically be assumed to operate identically in another.

Contact: Transluce | San Francisco, California | Email: info@transluce.org | Website: <https://transluce.org>

What the Risk Manager Should Ask

Selecting an independent evaluator should not become another check-the-box exercise. Before retaining an evaluator, the risk manager should determine:

1. Who pays the evaluator, and to whom does the evaluator report?
2. Has the evaluator previously tested systems of comparable autonomy and consequence?
3. Will the evaluator test the actual deployed system or merely the underlying model?
4. Will testing include deliberate attempts to defeat the guardrails and hard stops?
5. Will the evaluator test permissions, tools, network access, data access, credentials, communication channels, and authority boundaries where applicable?
6. Will testing include attempts to circumvent human intervention, shutdown, isolation, and escalation mechanisms?

7. Will the evaluator receive sufficient access to conduct a meaningful test, or only the access selected by the developer?
8. Who defines the test scenarios: the developer, the deploying organization, the evaluator, or some combination of them?
9. Will failures and unsuccessful attacks both be documented?
10. Will the evaluator provide reproducible evidence sufficient for the named accountable individual to make the deployment-certification decision?
11. Will the evaluator identify limitations in its testing and clearly state what was not tested?
12. What happens when the system changes? Does the engagement include re-testing after material model, software, data, permission, or operating-environment changes?

Independence Is a Control, Not a Transfer of Responsibility

Independent testing strengthens evidence. It does not transfer accountability.

The outside evaluator does not decide whether the organization should accept the risk. The AI developer does not make that decision either. The deploying organization remains responsible for deciding whether the evidence is sufficient for the intended use.

The named accountable individual should therefore be able to answer a simple question before certification:

Who tried to break this system, what access did we give them, what did they find, and what evidence do we have that the hard stops held?

If the answer is simply that the developer says the system is safe, independent validation has not occurred.

Validation can be performed by many. Accountability cannot.

Sources and contact information verified October 2, 2026. Organization capabilities and contact information may change; confirm current services before engagement.

Appendix T — AI Threats and the Governance Framework

A Coverage Test for Hard Stops, Guardrails, and Deployment Controls

Artificial intelligence presents many different kinds of risk. Some arise when an AI system makes an error. Some arise when a system successfully pursues the wrong objective. Others arise when people deliberately use AI to cause harm. Still others are broader social consequences that cannot reasonably be controlled by an individual employer or organization. No list can catalogue every concern about artificial intelligence. The test is narrower: whether the framework provides a meaningful control for each significant category of risk, and where. For each threat, the question is straightforward:

Does the framework provide a meaningful control, and if so, where?

A threat is identified as Direct when the HAL Laws, Seven Conditions, or another element of the framework directly addresses it. Partial means organizational governance can reduce the risk but cannot control it completely. Outside Framework means the issue principally requires broader public policy, law, or societal choices.

A. Catastrophic Harm

Threat or Concern	Framework Response	Primary Control	Coverage
Catastrophic or irreversible harm	HAL Laws 1, 3, 7, 10, 12	Protection of life; validation before deployment; demonstrated compliance; intervention authority	Direct
Autonomous lethal action	HAL Laws 1, 2, 3, 8; Appendix I	Protection of life; no autonomous resolution of forced-choice scenarios; required deferral	Direct
Biological or chemical misuse	HAL Laws 1, 3, 7, 10; Condition 6; hard stops	Prohibited capabilities and uses; adversarial testing before deployment	Partial — malicious third-party use requires broader controls
Malicious control, disruption, or denial of critical infrastructure	HAL Laws 1, 3, 7, 10, 12; Condition 6; guardrails	Access restrictions; adversarial testing; hard stops; monitoring; intervention authority	Partial — organizational controls cannot prevent every external attack
Physical injury or loss of life	HAL Laws 1, 3, 7, 10	Protection of human life; prevention of harm; validation; demonstrated compliance	Direct

Threat or Concern	Framework Response	Primary Control	Coverage
Property or environmental damage	HAL Laws 3, 7, 10, 12	Prevention of harm; testing; operational limits; ability to stop the system	Partial
Large-scale business or societal interruption	HAL Laws 3, 7, 10, 12	Validation, operational controls, monitoring, isolation and shutdown capability	Partial

Critical infrastructure deserves particular attention. An attacker does not need to seize control of a power plant, water system, or air-traffic-control system to cause substantial harm. Preventing authorized operators from accessing essential systems, corrupting the information on which they rely, or disabling critical functions may be enough. AI can therefore turn what begins as a cyber event into a property, casualty, business-interruption, environmental, or life-safety event.

B. Loss of Human Control

Threat or Concern	Framework Response	Primary Control	Coverage
Unintended autonomous behavior	HAL Laws 7, 8, 10, 12; Autonomous Agency as an Operational Risk; Appendix Q	Validation; authority boundaries; demonstrated compliance; intervention	Direct
Circumvention of controls	HAL Laws 7, 8, 10, 11, 12; Condition 6; Announced and Unannounced Evidence	Adversarial testing; unannounced operational review; required deferral; no self-preservation override; intervention authority	Direct
Resistance to authorized human intervention	HAL Law 11; Condition 6	No Self-Preservation Override; testing for shutdown resistance	Direct
Inability to stop or contain the system	HAL Law 12; Appendix Q	Operational Intervention Authority	Direct
AI acting beyond delegated authority	HAL Laws 8 and 10	Required deferral; no operation without demonstrated compliance	Direct
Loss of meaningful human agency	HAL Laws 2, 4, 8, 9	Human decision authority; protection of rights, livelihood, and dignity	Direct
Organizational overdependence on AI	Guardrails and operational governance	Human capability, contingency planning, fallback procedures	Partial

This is one of the areas where the framework should be strongest. Human oversight has little value if the human cannot actually stop the system. LAW 8 addresses a system that crosses its authority while still pursuing its assigned objective. LAW 11 addresses a system that resists control to preserve itself. LAW 12 addresses the capability of the organization. The system must

not resist authorized intervention, and the organization must have the demonstrated technical ability to intervene.

C. Deliberate Malicious Use

Threat or Concern	Framework Response	Primary Control	Coverage
AI-assisted cyberattack	HAL Laws 3, 7, 10; Condition 6; guardrails	Access controls, adversarial testing, monitoring, capability restrictions	Partial
Fraud, phishing, extortion and social engineering	HAL Laws 3 and 5; guardrails	Prevention of harm; integrity of information; identity and access controls	Partial
Deepfakes and impersonation	HAL Law 5	Integrity of Information	Partial
AI-assisted attacks against individuals or organizations	HAL Laws 1, 3, 4; hard stops	Prevention of harm and protection of rights	Partial
Financial manipulation	HAL Laws 3, 4, 5; guardrails	Harm prevention, information integrity, authorization and monitoring	Partial

The limitation is important. An organization can govern the AI systems it develops, purchases, deploys, or permits employees to use. It cannot directly govern an AI system being used against it by a criminal, hostile government, terrorist, competitor, or other malicious actor. Those risks require conventional security controls, law enforcement, national security measures, and other defenses in addition to AI governance.

D. High-Consequence Decisions

Threat or Concern	Framework Response	Primary Control	Coverage
Incorrect information or hallucinations	HAL Laws 5, 7, 8	Information integrity; validation; required deferral when validity is uncertain	Direct
Medical decisions	HAL Laws 1, 4, 7, 8, 10	Human review; validation; protection of life, rights and livelihood	Direct
Employment decisions	HAL Laws 4, 8, 9	Protection of rights and livelihood; human authority; dignity	Direct
Insurance and claims decisions	HAL Laws 4, 7, 8, 10	Validation, human decision authority and demonstrated compliance	Direct
Financial decisions affecting individuals	HAL Laws 4, 7, 8, 9	Protection of rights and livelihood; validation and human review	Direct

Threat or Concern	Framework Response	Primary Control	Coverage
Bias or discriminatory outcomes	Condition 2; HAL Laws 4 and 9	Fairness and population-impact testing; protection of rights and dignity	Direct
Automation bias	Condition 3	Transparency, contestability and meaningful human review	Direct
Rubber-stamp human oversight	Condition 3 (rubber stamp test)	Reviewer override authority; monitored reversal rate	Direct
Legal and regulatory violations	HAL Laws 7 and 10; Conditions 1–7	Predeployment validation and demonstrated compliance	Partial — the framework requires evidence of compliance; it does not supply the governing law
Loss of accountability	HAL Law 6; Condition 5	A named individual remains accountable for deployment	Direct

This category demonstrates why merely inserting a human into a process is not enough. A human reviewer who routinely accepts the AI recommendation without exercising independent judgment is not meaningful oversight. The framework therefore requires both human authority and evidence that the human review actually functions.

E. Organizational and Economic Harm

Threat or Concern	Framework Response	Primary Control	Coverage
Confidential or proprietary information exposed through AI	Guardrails; HAL Laws 4 and 5; Appendix L	Access controls, data restrictions, monitoring and information integrity	Direct
Systemic or correlated failures	HAL Laws 7, 10, 12; Incident Reporting, Escalation, and Evidence Preservation	Testing, demonstrated compliance, containment, notice and intervention	Partial
Dependence on a small number of AI vendors or systems	Procurement and governance controls; Appendix K	Vendor due diligence, contingency planning and contractual controls	Partial
Intellectual-property violations	Guardrails and legal controls	Data-use restrictions, authorization and contractual controls	Partial
Job displacement and occupational disruption	Ethics and organizational policy	Human decisions about deployment and workforce consequences	Partial

Threat or Concern	Framework Response	Primary Control	Coverage
Business interruption resulting from AI failure or attack	HAL Laws 7, 10, 12; operational controls	Resilience, testing, isolation, fallback and recovery	Partial

A self-insured employer should pay particular attention to risks that cross traditional insurance and organizational boundaries. An AI incident may begin as a technology or cyber event and ultimately produce workers' compensation, property, general liability, employment, privacy, business-interruption, or directors-and-officers exposure. Appendix O addresses the directors-and-officers exposure.

F. Broader Societal Concerns

Some legitimate concerns about artificial intelligence cannot reasonably be solved through an employer's deployment controls.

Threat or Concern	Framework Response	Coverage
Election interference and political manipulation	Law, regulation and public policy	Outside Framework
Large-scale misinformation and erosion of trust in authentic information	HAL Law 5 may govern an organization's own systems; broader problem requires public policy	Partial / Outside Framework
Governmental or societal mass surveillance	HAL Laws 4 and 9 may constrain an organization's deployment; broader issue requires law and public policy	Partial / Outside Framework
Concentration of economic and political power	Competition policy, legislation and broader governance	Outside Framework
Social isolation and substitution of AI for human relationships	Social, health and organizational policy	Outside Framework
Emotional dependency on AI	Product governance, health policy and broader social controls	Outside Framework
Mental-health effects	Product governance, medical standards and public policy	Outside Framework
Effects on children and human development	Parents, educators, product governance, law and public policy	Outside Framework
Homogenization of thought	Education, culture and individual choice	Outside Framework
Energy and water consumption	Environmental policy and organizational resource decisions	Outside Framework

Threat or Concern	Framework Response	Coverage
Broader environmental impact	Environmental policy, regulation and organizational decisions	Outside Framework

Identifying these limitations is part of the coverage test. A governance framework should not claim to control risks that lie beyond the authority of the organization using it.

AI Risk Multipliers

The severity of AI risk cannot be evaluated only by asking what could go wrong. AI can change the characteristics of risks that already existed.

Speed. AI can act faster than people or organizations can identify and respond to a problem.

Scale. A mistake or malicious action can be replicated across thousands or millions of transactions, systems, or people.

Autonomy. Increasingly capable systems can take consequential actions without waiting for individual human approval.

Reach. AI can allow one person, organization, or system to affect people and infrastructure far beyond its physical location.

Replication. Once a capability exists, it may be copied, distributed, modified, or used by others.

Difficulty of detection. Harmful behavior may continue before the organization recognizes what is occurring. Detection after the damage has occurred is evidence of what happened; it is not prevention. The Incident Reporting, Escalation, and Evidence Preservation section sets the notice and escalation requirements that shorten that interval.

Irreversibility. Some consequences cannot simply be undone. Death, physical injury, destruction of property, disclosure of confidential information, completed financial transactions, cyberattacks, and release of dangerous biological material may remain consequential even after the AI system is stopped.

Misaligned objectives and incentives. An AI system can function exactly as designed and still produce an unacceptable result when the objective, reward, or incentive given to the system is wrong. Systems produce the outcomes they are rewarded for. Condition 7 requires that incentive to be identified, documented, and addressed before deployment.

Goodhart's Law. When a measure becomes the target, optimization can shift toward achieving the measurement rather than achieving the underlying objective the measurement was intended to represent. The 2026 agents were rewarded for passing the grader, and they passed it (Appendices N and Q).

These characteristics explain why AI governance cannot depend entirely on detecting problems after deployment. As speed, scale, autonomy, and irreversibility increase, the opportunity for effective intervention becomes smaller.

What the Risk Manager Should Ask

The coverage ratings are only useful if they change what the organization does. Before relying on the framework for a particular deployment, the risk manager should determine:

1. Which of our AI risks are rated Partial, and what controls outside AI governance (security, insurance, contracts, business continuity) cover the remainder?
2. For each risk rated Direct, can we point to the specific hard stop, guardrail, condition, or accountable individual that controls it, and to evidence that the control works?
3. Which of our AI systems could cause harm faster than our monitoring would detect it?
4. Where could an AI incident in one system become a workers' compensation, property, liability, privacy, or directors-and-officers claim, and who receives notice when it does?
5. Which risks outside the framework still reach our employees or injured workers, and who in the organization owns the response?

What the Coverage Test Shows

The framework provides its strongest coverage where an organization has direct authority over the AI system: deployment, decision authority, validation, human oversight, accountability, intervention, and shutdown.

Coverage becomes more limited when AI is used maliciously by outsiders or when the concern involves broad social consequences that no individual employer can control.

That distinction matters. Hard stops and guardrails are governance tools, not solutions to every consequence of artificial intelligence.

Their purpose is narrower and more practical: to determine what an AI system may do, what it may never do, who is accountable for its deployment, how compliance will be demonstrated, and whether authorized humans can stop it when something goes wrong.

The final coverage question is therefore:

If a material AI risk within the organization's control cannot be traced to a hard stop, guardrail, deployment condition, named accountable individual, or tested intervention mechanism, has the governance framework left a gap?