Human Oversight for Autonomous AI Agents: A Governance Framework for Healthcare, Finance, Energy, and Public Infrastructure

Aaron K Montgomery [1*], Hannah E Gallagher [2], Derrick L Mercer [3]
[1-3] Carl H. Lindner College of Business, University of Cincinnati, Ohio, USA

* Corresponding Author: **Aaron K Montgomery**

Article Info

Volume: 01
Issue: 01
January-February 2025
Received: 25-01-2025
Accepted: 17-02-2025
Page No: 18-32

Abstract
Autonomous AI agents can plan, call tools, communicate with other systems, and execute operational actions without continuous human direction. These capabilities create a governance problem that differs from conventional model oversight because harmful effects may arise from sequences of actions, changing environments, and interactions across organizational boundaries. This study develops and empirically evaluates the Human Oversight Governance Framework for Autonomous AI Agents (HOGF-AI). The evidence base is a structured content analysis of 24 public governance instruments available by May 31, 2025, with six documents each from healthcare, finance, energy, and public infrastructure. A 24-item codebook operationalizes eight dimensions: intervention authority, escalation and accountability, monitoring and validation, explainability and traceability, risk and cybersecurity, fairness and contestability, lifecycle control, and organizational learning. Each provision was scored on a five-level operationalization scale. Internal consistency was high across all dimensions, with Cronbach's alpha from 0.842 to 0.972. An exploratory two-component partial least squares model explained 79.0 percent of leave-one-out variation in operational Responsible AI readiness. Monitoring and validation, lifecycle and change control, and escalation and accountability had positive bootstrap-supported coefficients. The equal-weight Human Oversight Governance Index correlated with readiness at Spearman rho = 0.643. Public infrastructure and healthcare scored highest overall, while energy guidance was strongest in safety and cybersecurity but weaker in rights, intervention, and learning. The findings show that meaningful oversight is not a single human approval step. It is a lifecycle capability that combines bounded autonomy, evidence-based escalation, stop authority, continuous validation, audit records, and institutional learning. The framework and index provide a reproducible benchmark for organizations deploying autonomous agents in high-stakes settings.

DOI: https://doi.org/10.54660/.IJECA.2025.1.1.18-32

Keywords: autonomous AI agents, human oversight, Responsible AI, AI governance, critical infrastructure, explainable artificial intelligence, organizational accountability, governance maturity

1. Introduction

Autonomous AI agents are moving the operational boundary of artificial intelligence. A conventional predictive model returns a score, classification, or recommendation. An agent can interpret a goal, decompose it into steps, retrieve information, invoke software tools, communicate with other agents, and execute actions that change an organizational environment. Tool-using language models, planning loops, memory systems, and reflective architectures have made this pattern technically feasible in research and early deployment (Park *et al*., 2023; Schick *et al*., 2023; Shinn *et al*., 2023; Yao *et al*., 2023). The important governance question is not whether an agent appears intelligent. It is whether an institution can constrain what the agent may

High-stakes sectors expose the limits of governance models designed for one-time prediction. In healthcare, an autonomous workflow could summarize clinical records, prioritize patients, prepare orders, or coordinate supply decisions. The same system could amplify an error across several steps before a clinician sees the result. Healthcare governance therefore emphasizes patient safety, professional judgment, secure data handling, and the ability to challenge a machine-supported decision (World Health Organization, 2021, 2023, 2025). In finance, agents may investigate transactions, produce customer communications, adjust limits, or trigger payment controls. Model risk, consumer protection, cybersecurity, and segregation of duties become inseparable (Board of Governors of the Federal Reserve System & Office of the Comptroller of the Currency, 2011; Prudential Regulation Authority, 2023). In energy and public infrastructure, agent actions can affect reliability, dispatch, emergency response, benefits administration, or cyber defense. Small errors may propagate through tightly coupled systems.

Existing governance literature provides essential foundations but leaves an operational gap. Responsible AI principles identify fairness, transparency, accountability, privacy, safety, and human control (Jobin *et al*., 2019; National Institute of Standards and Technology, 2023; OECD, 2024). Human automation research distinguishes levels of automation and shows that human performance depends on workload, situation awareness, trust calibration, and the timing of intervention (Hoff & Bashir, 2015; Lee & See, 2004; Parasuraman *et al*., 2000). Algorithmic accountability research adds documentation, audit, contestability, and institutional answerability (Bovens, 2007; Diakopoulos, 2016; Kroll *et al*., 2017; Raji *et al*., 2020). These strands are often applied separately. Autonomous agents require them to operate as one lifecycle control system.

Human oversight is frequently reduced to a phrase such as human in the loop. That label does not specify who has authority, what evidence reaches the reviewer, how quickly intervention must occur, or whether the human can stop the system. A nominal reviewer may be asked to approve hundreds of outputs, may lack access to primary evidence, or may be unable to reverse an executed action. Explanations can also increase inappropriate reliance when they are plausible but uninformative (Bansal *et al*., 2021; Buçinca *et al*., 2021). Meaningful oversight therefore requires more than exposure to an AI output. It requires decision rights, technical controls, organizational capacity, and an audit trail linking action to accountable owners.

The present study develops the Human Oversight Governance Framework for Autonomous AI Agents, abbreviated HOGF-AI. The framework defines eight measurable governance dimensions and a composite Human Oversight Governance Index, or HOGI. It is empirically examined through a structured analysis of 24 official governance instruments from healthcare, finance, energy, and public infrastructure. This design does not invent organizational survey responses. It treats public governance documents as observable statements of expected control capability and codes the degree to which each document moves from a general principle to an operational requirement. The study addresses six questions. RQ1 asks which mechanisms are required for meaningful oversight. RQ2 asks which capabilities are most strongly associated with operational Responsible AI readiness. RQ3 compares the four sectors. RQ4 examines organizational conditions that make intervention effective. RQ5 tests whether HOGI predicts readiness. RQ6 considers how autonomy should be balanced against human intervention. The contribution is theoretical, empirical, and practical. Theoretically, it integrates human automation, institutional accountability, socio-technical systems, dynamic capabilities, information processing, and organizational learning. Empirically, it supplies a transparent coded corpus, measurement model, index, and sensitivity analysis. Practically, it provides an escalation workflow and maturity roadmap for organizations that need to govern agents rather than merely validate models.

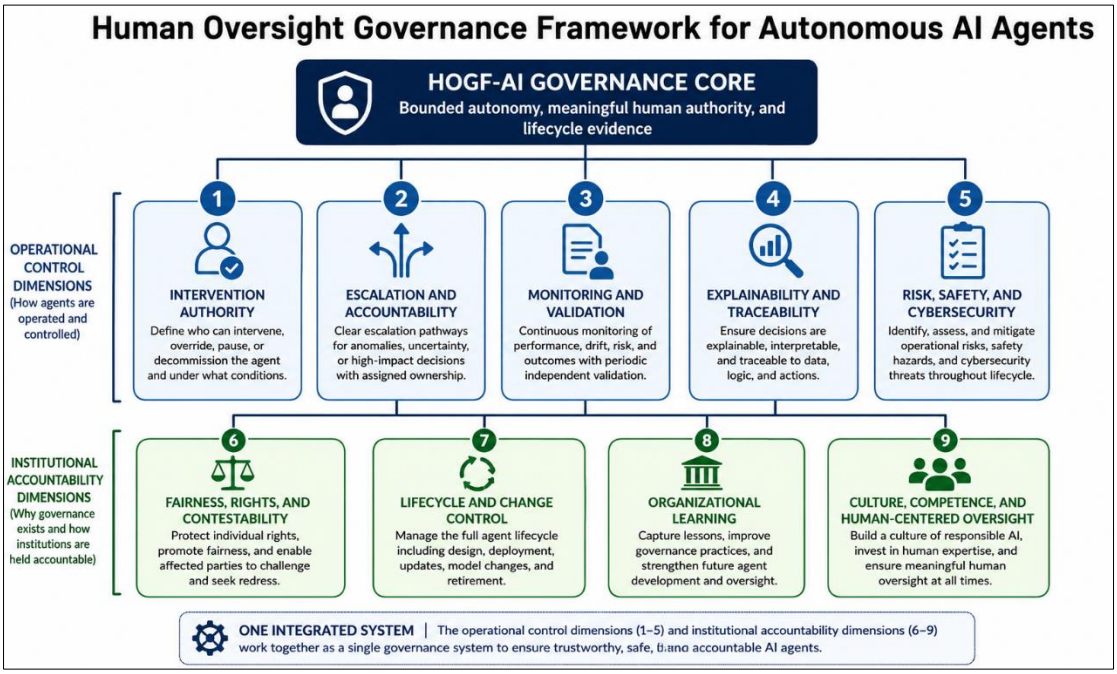

Fig 1: Human Oversight Governance Framework for Autonomous AI Agents.

2. Literature Review

Autonomous agents extend a long tradition of research on intelligent and multi-agent systems. Agents are generally characterized by some degree of autonomy, environmental interaction, goal-directed behavior, and the capacity to select actions (Russell & Norvig, 2021; Wooldridge, 2009). Contemporary systems add natural-language planning and access to external tools. ReAct combines reasoning traces with actions, Toolformer demonstrates learned tool use, Generative Agents introduces memory and reflection, and Reflexion uses feedback to revise subsequent behavior (Park *et al*., 2023; Schick *et al*., 2023; Shinn *et al*., 2023; Yao *et al*., 2023). These designs improve flexibility but enlarge the control surface. An agent may fail through an incorrect plan, an unsafe tool call, a misleading observation, a compromised data source, a cascading action sequence, or coordination with another system.

Human oversight research offers several control models. Human in the loop places a person within the decision process. Human on the loop allows automated operation while a person monitors and can intervene. Human in command reserves authority over system objectives, deployment boundaries, and termination. These forms correspond to different levels and stages of automation (Parasuraman *et al*., 2000). No form is inherently superior. Continuous approval can create bottlenecks and automation bias, while remote monitoring can fail when the human receives weak signals or too little time. Trust should be calibrated to system capability rather than maximized (Hoff & Bashir, 2015; Lee & See, 2004). Human-centered AI therefore combines high automation with high human control rather than treating them as opposites, while established interaction guidelines emphasize feedback, correction, and user control (Amershi *et al*., 2019; Shneiderman, 2020).

Organizational research clarifies why oversight cannot be located solely at the user interface. Human-AI collaboration is shaped by task allocation, professional expertise, incentives, authority, and the surrounding workflow (Jarrahi, 2018; Raisch & Krakowski, 2021). Augmented intelligence is valuable when machine-scale analysis complements human expertise and business-process controls rather than displacing judgment without support (Kamruzzaman *et al*., 2023). In healthcare, secure infrastructure and disciplined human supervision are especially important because data quality, privacy, and patient consequences interact (Hasan *et al*., 2022). Resilient healthcare operations likewise depend on coordinated information and accountable allocation decisions (Rasel *et al*., 2022).

Responsible AI frameworks establish common values but differ in operational detail. The NIST AI Risk Management Framework organizes activity around Govern, Map, Measure, and Manage functions (National Institute of Standards and Technology, 2023). The OECD principles were updated in 2024 to address evolving system capabilities and emphasize robustness, transparency, accountability, and human-centered values (OECD, 2024). The European Union AI Act places human oversight within risk-based legal obligations for high-risk systems (European Union, 2024). ISO/IEC 42001 provides an organizational management-system structure for AI (International Organization for Standardization, 2023). These instruments create broad convergence around lifecycle governance. They do not resolve how an organization should set intervention thresholds for an agent that acts repeatedly and at machine speed.

Accountability fills part of this gap. Institutional accountability requires an actor who must explain and justify conduct to a forum capable of questioning and imposing consequences (Bovens, 2007). In algorithmic settings, this relationship depends on audit records, documentation, named ownership, and review procedures (Diakopoulos, 2016; Kroll *et al*., 2017; Raji *et al*., 2020). Model cards and data documentation make technical artifacts more inspectable (Gebru *et al*., 2021; Mitchell *et al*., 2019). Yet an agent requires action-level provenance in addition to model-level documentation. The organization needs to know which objective was active, which tools were called, what information was retrieved, which constraints were evaluated, and who authorized the autonomy envelope.

Explainability supports oversight only when it is matched to the decision. SHAP and LIME can describe feature contributions or local behavior, while counterfactual explanations can identify changes associated with a different output (Lundberg & Lee, 2017; Ribeiro *et al*., 2016; Wachter *et al*., 2018). For agents, explanation must extend beyond a score to the action sequence and the evidence used at each step. A generated rationale is not necessarily faithful. Oversight should therefore combine explanation with logs, input provenance, policy checks, and independent replay. This caution is consistent with the broader argument that principles alone do not guarantee ethical practice (Mittelstadt, 2019).

Sector literature shows distinct control priorities. Financial governance stresses model inventory, independent validation, senior accountability, change control, and cybersecurity (Financial Stability Board, 2024; U.S. Department of the Treasury, 2024). Algorithmic accountability in consumer FinTech adds fair-lending and financial-stability concerns (Fahim *et al*., 2023), while anomalous-market analytics highlights the need for explainable early-warning systems and human investigation (Jahan *et al*., 2024). Financial reporting and narrative-manipulation research further shows that generative systems can create integrity risks as well as detection capabilities (Pritty *et al*., 2024). Digital-payment transitions increase the importance of transparent and inclusive controls (Islam *et al*., 2025).

Energy guidance emphasizes reliability, cyber resilience, infrastructure dependencies, and controlled deployment. Clean-energy analytics demonstrates the potential of data-intensive forecasting, while sustainable-waste research links machine learning to resource stewardship (Arman, Hasan, *et al*., 2024; Arman, Rasel, *et al*., 2024). Energy-burden-aware underwriting illustrates how AI can connect energy and financial decisions but also introduce distributional and governance risks (Rabbani *et al*., 2024). Public-payment infrastructure raises similar issues of synthetic identity, improper payments, and accountable public-benefit decisions (Rahat *et al*., 2024). Across sectors, complex multimodal models can increase predictive scope but also demand fairness and explanation controls (Rasel *et al*., 2023).

The literature therefore converges on several requirements but lacks a validated cross-sector measurement structure for autonomous agents. The missing link is a framework that treats human intervention, monitoring, explanation, safety, rights, change control, and learning as complementary

capabilities. Supplementary Table S1 summarizes the literature streams and the governance gap addressed by HOGF-AI.

3. Theoretical Framework and Hypotheses

HOGF-AI integrates six theoretical lenses. Human automation theory explains allocation of functions between people and machines. Socio-technical reasoning treats performance as an outcome of technology, work design, roles, and institutional context. Information-processing theory predicts that greater uncertainty and interdependence require stronger sensing and interpretation capacity. Dynamic capabilities theory adds sensing, seizing, and reconfiguration when conditions change. Organizational learning explains how incidents and overrides become revised routines. Institutional accountability establishes answerability, review, and consequences.

The framework begins with bounded autonomy. An agent should operate only within an approved envelope that defines objectives, permitted tools, data sources, transaction limits, affected populations, and prohibited actions. Intervention authority is the first dimension because a human cannot provide meaningful oversight without practical ability to modify, pause, or terminate behavior. This authority must be technically executable and institutionally protected. H1 therefore predicts that intervention authority is positively associated with operational Responsible AI readiness.

Escalation and accountability form the second dimension. Agents encounter uncertainty, novel states, conflicts among rules, and conditions outside their authorization. Escalation converts those states into a named human decision. A useful protocol identifies trigger conditions, responsible roles, response time, evidence requirements, and senior ownership. Accountability is strengthened when responsibility remains with an institution rather than being displaced onto an automated system. H2 predicts a positive association between escalation and accountability and readiness.

Monitoring and validation form the third dimension. Predeployment testing cannot exhaust the states an autonomous agent may encounter. Organizations need continuous monitoring of performance, constraint violations, tool behavior, data drift, security events, and human overrides. Periodic validation should examine whether the autonomy envelope remains appropriate. H3 predicts that monitoring and validation have a positive association with readiness.

Explainability and traceability form the fourth dimension. Supervisors need enough information to understand the basis of an action and to reconstruct the sequence after the event. This includes user-facing reasons where appropriate, internal provenance, model and prompt versions, data sources, tool calls, policy checks, and human decisions. H4 predicts a positive association with readiness. The theory also anticipates that explanation without authority may have limited effect, so the empirical analysis treats dimensions jointly rather than assuming independent sufficiency.

Risk, safety, and cybersecurity form the fifth dimension. Autonomous agents can create conventional model error, unsafe action, cyber compromise, data leakage, and system-level propagation. Risk assessment should address intended use, foreseeable misuse, failure containment, adversarial manipulation, third-party tools, and critical dependencies. H5 predicts a positive association between this dimension and readiness.

Fairness, rights, and contestability form the sixth dimension. High-stakes agents may allocate attention, interrupt services, affect credit, or influence access to public resources. Oversight must preserve nondiscrimination, meaningful notice, and routes for challenge and correction. H6 predicts that stronger fairness and contestability provisions are associated with readiness and that their implementation differs by sector.

Lifecycle and change control form the seventh dimension. Agents change through model updates, prompt revisions, new tools, memory accumulation, policy changes, and shifts in operating context. Each change can alter behavior even when the base model remains fixed. Predeployment review, version control, rollback, and retirement criteria are therefore central. H7 predicts a positive association between lifecycle control and readiness.

Organizational learning forms the eighth dimension. Overrides and incidents have limited value when they remain isolated events. Mature organizations capture feedback, conduct post-incident review, update training and controls, and transfer lessons across agent portfolios. H8 predicts a positive association between organizational learning and readiness.

Finally, H9 predicts that the composite Human Oversight Governance Index is positively associated with operational Responsible AI readiness. H10 predicts sector differences because healthcare and public administration place greater emphasis on rights and direct human review, finance on validation and accountability, and energy on resilience and cybersecurity. Figure 1 presents the integrated framework.

4. HOGF-AI Framework Development

Framework development followed a theory-to-measurement sequence. First, the literature was mapped to governance functions that remain meaningful across sectors. Second, overlapping concepts were consolidated. Transparency was placed within explainability and traceability; auditability was distributed across traceability, monitoring, and lifecycle control; cybersecurity was joined with risk and safety because operational containment is shared. Third, each dimension was represented by three observable indicators, producing a 24-item codebook. The indicators were designed to distinguish a general statement of principle from an operational control.

The coding scale has five levels. A score of zero means that the provision is absent. One indicates acknowledgment or aspiration. Two indicates a defined procedure or control. Three requires assigned responsibility, trigger, or documented evidence. Four indicates a measurable, testable, or enforceable requirement. This scale avoids equating frequent use of governance language with operational maturity. A document that endorses human oversight but provides no stop authority receives a lower score than one that specifies who can intervene and under what conditions.

The framework is deliberately lifecycle based. Before deployment, an organization defines the purpose, risk tier, prohibited actions, human owner, and evidence required for authorization. During operation, the agent is monitored against an autonomy envelope. Uncertainty, potential harm, policy conflict, or cyber anomaly triggers escalation. A responsible person can approve, modify, pause, or terminate action. After the event, logs support audit and feedback updates controls, training, or retirement decisions. Figure 2 illustrates this cycle.

Sector adaptation occurs through the meaning of risk rather than through removal of core dimensions. In healthcare, intervention may require a licensed professional and contestability may involve clinical review. In finance, threshold changes and adverse actions are subject to model-risk and consumer-protection controls. In energy, intervention timing and fail-safe operation must reflect reliability constraints. Public infrastructure adds due process, public transparency, and mission continuity. Figure 3 shows a shared governance core surrounded by sector-specific safeguards.

The HOGI score is the arithmetic mean of the eight normalized dimension scores. Equal weighting is the primary specification because the framework treats every dimension as a necessary part of meaningful oversight and no empirical basis exists for allowing strength in one dimension to compensate fully for absence in another. Entropy weighting is reported as a sensitivity analysis. A maturity interpretation maps scores below 40 to ad hoc, 40 to 59.9 to controlled, 60 to 74.9 to managed, 75 to 89.9 to assured, and 90 or above to adaptive governance. These labels describe document-level operationalization, not verified organizational practice.

HOGF-AI also distinguishes three oversight modes. Direct human control is appropriate when consequences are immediate, difficult to reverse, or legally reserved for a person. Supervisory control is appropriate when the agent may execute within narrow limits but must surface uncertainty and exceptions before material action. Strategic human command applies to lower-frequency decisions about purpose, autonomy level, resource access, and continued deployment. These modes can coexist within one agent. A clinical scheduling agent may autonomously resolve routine conflicts, require supervisory approval for high-risk patients, and remain subject to executive restrictions on data access and system integration. Treating oversight as a layered design avoids the false choice between full automation and manual review.

The framework further distinguishes action control from outcome review. Action control constrains what an agent can do before or during execution. Outcome review evaluates consequences after an action and can trigger correction, compensation, or system change. Both are necessary. Strong ex ante restrictions may block useful performance when environments are uncertain, while exclusive reliance on ex post review may allow irreversible harm. The appropriate combination depends on impact, reversibility, latency, and the ability to observe consequences. HOGF-AI therefore requires each use case to identify the point at which human authority is most effective rather than assuming that a visible approval button provides adequate control.

A minimum-control principle accompanies the composite index. Some capabilities are noncompensatory for high-impact uses. An organization should not authorize an agent that can affect health, finance, energy reliability, or public rights when there is no named owner, no stop mechanism, or no reproducible action log, even if the remaining HOGI dimensions are strong. The total score supports benchmarking, but authorization decisions should also apply critical-control floors. This distinction prevents index optimization from replacing professional and legal judgment.

Table 1: Governance constructs and measurement items

Item	Indicator	Operational definition
IA1	Human approval	High-impact actions require prior or concurrent human authorization.
IA2	Manual override	A responsible person can modify or reverse an agent action.
IA3	Stop authority	The agent can be paused, isolated, or terminated through an operational control.
EA1	Named owner	A role is accountable for the agent and its outcomes.
EA2	Escalation path	Triggers, recipients, evidence, and response times are defined.
EA3	Senior review	Material risks and changes receive executive or board-level review.
MV1	Continuous monitoring	Performance, actions, constraints, and anomalies are monitored during use.
MV2	Periodic validation	Independent or scheduled validation tests ongoing suitability.
MV3	Incident detection	Incidents and near misses are identified, classified, and reported.
ET1	Decision explanation	Affected users or reviewers receive a reason appropriate to the decision.
ET2	Disclosure	System purpose, limits, and human role are communicated.
ET3	Provenance and logs	Inputs, model or prompt versions, tools, actions, and approvals are reconstructable.
RS1	Risk assessment	Intended use, foreseeable misuse, failure modes, and dependencies are assessed.
RS2	Testing and assurance	Safety, robustness, and stress tests are defined and evidenced.
RS3	Cybersecurity and resilience	Access, data, tool, supply-chain, and recovery risks are controlled.
FR1	Bias and fairness testing	Group impacts and discriminatory pathways are evaluated.
FR2	Rights protection	Privacy, nondiscrimination, and human rights constraints are explicit.
FR3	Appeal and contestability	Affected parties can challenge facts, reasons, or outcomes.
LC1	Predeployment review	Purpose, risk tier, data, tools, and autonomy are approved before use.
LC2	Change control	Model, prompt, tool, permission, and policy changes are versioned and validated.
LC3	Rollback and retirement	Rollback, suspension, renewal, and retirement criteria are defined.
OL1	Feedback capture	Overrides, complaints, incidents, and user observations are recorded.
OL2	Staff competence	Supervisors and operators receive role-specific training and exercises.
OL3	Post-incident improvement	Lessons update controls, data, models, and operating procedures.

Note: Items were scored from 0, absent, to 4, measurable or enforceable operational requirement.

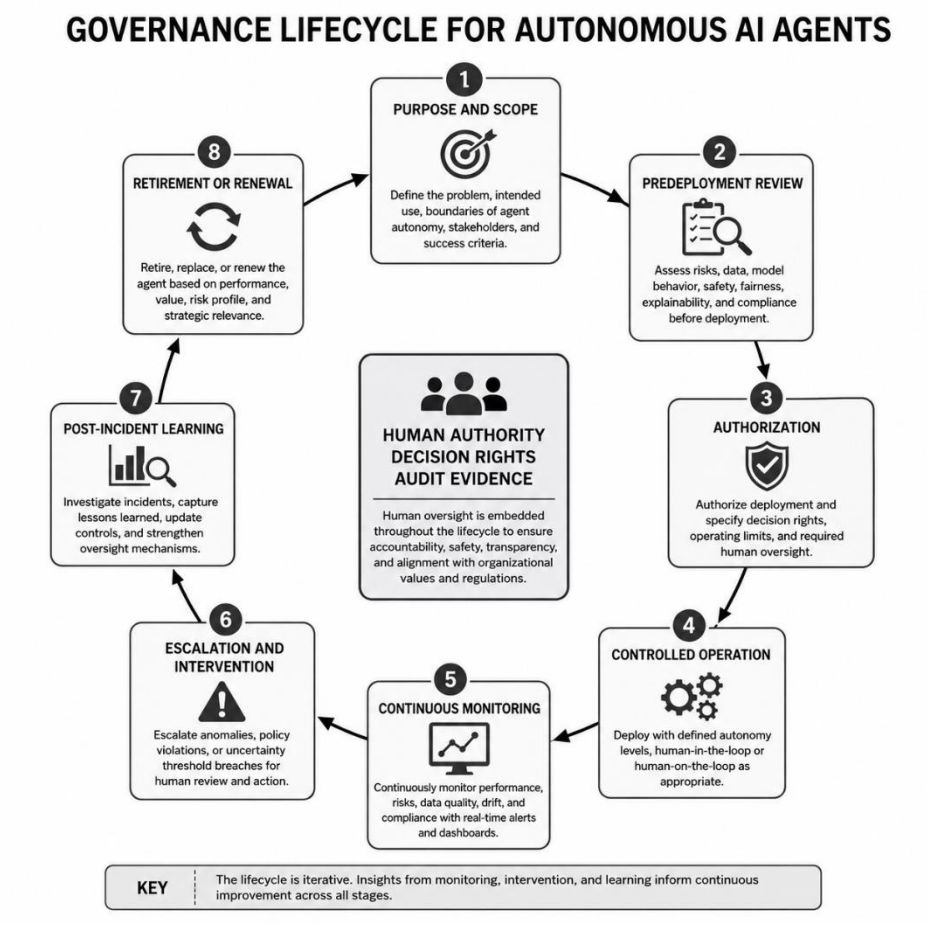

Fig 2: Governance lifecycle for autonomous AI agents.

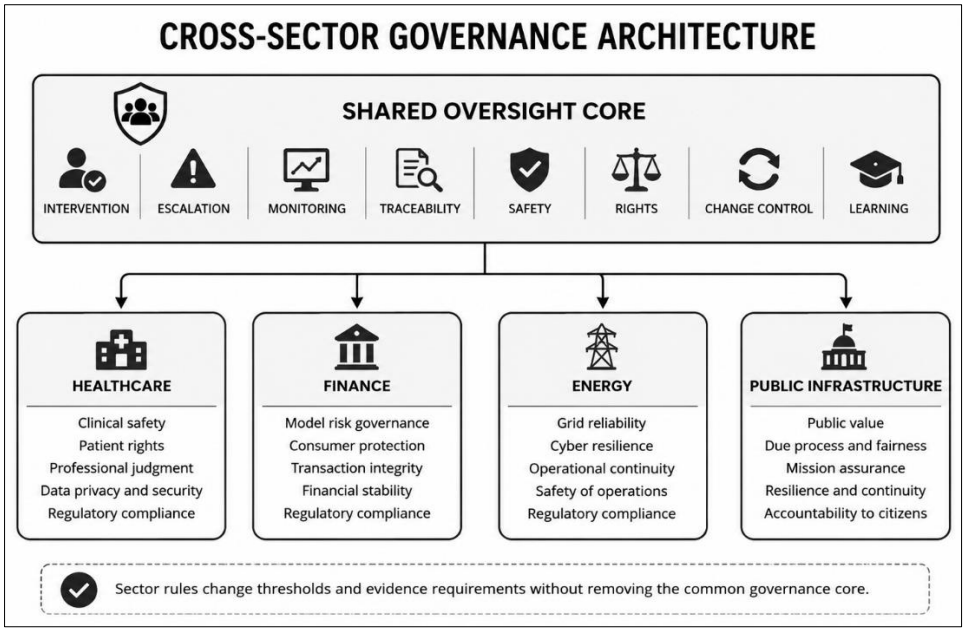

Fig 3: Cross-sector governance architecture.

5. Research Methodology

The empirical design is a structured document analysis with an evidence cutoff of May 31, 2025. Official governance instruments were selected because the research objective concerns institutional expectations for oversight, not public opinion about AI. The corpus includes 24 documents, six per sector. Eligibility required a public release before the cutoff, direct relevance to AI or model governance, sufficient operational detail for coding, and an identifiable issuing organization. Later web updates were not used to revise historical documents.

The healthcare corpus includes World Health Organization guidance, U.S. Food and Drug Administration lifecycle materials, and the Coalition for Health AI blueprint. The finance corpus includes U.S. and U.K. model-risk guidance, FINRA, the U.S. Treasury, the Financial Stability Board, and the Monetary Authority of Singapore. The energy corpus includes U.S. Department of Energy, North American

Electric Reliability Corporation, and International Energy Agency materials. Public infrastructure includes U.S. Office of Management and Budget memoranda, the Government Accountability Office framework, DHS and CISA roadmaps, and the Australian voluntary AI safety standard. Table 2 records dates and corpus roles.

Each document was reviewed in two passes. The first pass identified provisions corresponding to the 24 items. The second pass applied the operationalization scale consistently across the corpus and checked that comparable language received comparable scores. A keyword concordance audit searched terms such as override, escalation, monitoring, audit, appeal, incident, rollback, and training. Keyword presence did not determine a score; it served as a completeness check. One coder conducted the final scoring, so intercoder reliability is not claimed. The complete matrix is supplied as supplementary data and this limitation is reported directly.

Eight construct scores were calculated as the mean of their three items and normalized to a 0 to 100 scale. Operational Responsible AI readiness was assessed separately with three criterion indicators: mandatory implementation or an implementation timeline, a designated governance body or accountable owner, and a measurable reporting or assurance mechanism. Criterion items used the same 0 to 4 scale. This separation reduces, but does not eliminate, conceptual overlap between governance capability and readiness.

Reliability analysis included Cronbach's alpha, composite reliability, item-total loadings, and average variance extracted. Because the sample is small and ordinal, these statistics are interpreted as internal consistency of the coded provisions rather than psychometric validation of a population survey. Discriminant validity was assessed with the heterotrait-monotrait ratio. A threshold of 0.90 was used for the exploratory analysis. Construct correlations were also inspected for redundancy.

The structural analysis used two-component partial least squares regression. PLS was selected because the sample contains 24 documents, the objective is prediction of an operational criterion, and the governance dimensions are correlated. Predictors and the criterion were standardized. The number of components was selected by leave-one-out cross-validation. Bootstrap confidence intervals used 5,000 document-level resamples. Predictive performance was evaluated with leave-one-out R-squared and root mean squared error. The coefficients are exploratory associations, not causal effects.

Sector comparisons used means and Kruskal-Wallis tests because each sector contains six documents and normality cannot be assumed. The HOGI-readiness relationship was evaluated with Spearman correlation. Importance-performance analysis combined the absolute standardized PLS coefficient with the mean sector-neutral construct score. Robustness checks included equal versus entropy weighting, leave-one-document-out HOGI rankings, alternative readiness aggregation, and exclusion of superseded OMB M-24-10 from the public-infrastructure sector. The code and machine-readable tables reproduce all reported results.

The design has three safeguards against temporal leakage. First, every source has a dated release at or before the cutoff. Second, current webpages were used only to locate the historically dated document. Third, post-cutoff publications were excluded even when they would have been relevant. For example, later 2025 research by several recurring authors was not used. Eligible works by Rasel, Kamrul Islam, and Fahim were retained only when their publication date fell within the boundary.

Document selection sought functional comparability rather than identical legal status. The corpus contains binding regulation, supervisory guidance, voluntary standards, roadmaps, and technical frameworks. These instruments differ in enforceability, but all influence organizational expectations and can be coded for operational content. Legal force was not added to the governance score because the research question concerns control design. It was captured indirectly in the separate readiness criterion through mandatory implementation, designated ownership, and assurance obligations. This choice prevents a detailed voluntary standard from being treated as legally binding while still recognizing its technical specificity.

Content validity was strengthened through three crosswalks. The first compared measurement items with NIST AI RMF functions. The second mapped items to accountability practices in the GAO framework. The third checked human-control provisions against automation and trust research. Items that could not be located in at least two theoretical or policy traditions were removed during framework consolidation. The final dimensions therefore reflect both deductive theory and recurring governance practice. The codebook also avoids technology-specific terms where possible. A tool permission, model update, or memory change can all be evaluated under lifecycle control even when agent architectures evolve.

Common-method risk remains possible because constructs and readiness are coded from the same documents. Several design choices reduce the problem. Readiness uses criterion indicators that are not included in the 24 governance items. The structural model is validated out of sample at the document level through leave-one-out prediction. Results are also compared with a conservative readiness rule. These steps do not create independent outcome data, so the study describes the model as exploratory. Future organizational surveys and audits should measure control implementation and outcomes from different respondents or data systems.

Table 2: Governance-document corpus and timeline

ID	Sector	Issuer	Document	Release
H1	Healthcare	World Health Organization (2021)	Ethics and governance of AI for health	June 2021
H2	Healthcare	World Health Organization (2023)	Regulatory considerations on AI for health	October 2023
H3	Healthcare	World Health Organization (2025)	Guidance on large multi-modal models	March 2025
H4	Healthcare	U.S. Food and Drug Administration (2021)	AI/ML-Based SaMD Action Plan	January 2021
H5	Healthcare	U.S. Food and Drug Administration (2024)	Predetermined Change Control Plans	December 2024
H6	Healthcare	Coalition for Health AI (2023)	Blueprint for Trustworthy AI	April 2023
F1	Finance	Federal Reserve and OCC (2011)	SR 11-7 Model Risk Management	April 2011
F2	Finance	FINRA (2020)	Artificial Intelligence in the Securities Industry	June 2020
F3	Finance	Prudential Regulation Authority (2023)	SS1/23 Model Risk Management Principles	May 2023
F4	Finance	U.S. Department of the Treasury (2024)	AI-Specific Cybersecurity Risks in Financial Services	March 2024
F5	Finance	Financial Stability Board (2024)	AI and Financial Stability	November 2024
F6	Finance	Monetary Authority of Singapore (2022)	Veritas FEAT Assessment Methodology	2022
E1	Energy	U.S. Department of Energy (2020)	AI and Machine Learning for Energy Systems	August 2020
E2	Energy	U.S. Department of Energy (2023)	FECM Responsible AI Implementation Pathway	December 2023
E3	Energy	U.S. Department of Energy (2024a)	AI Compliance Plan	September 2024
E4	Energy	U.S. Department of Energy (2024b)	AI for Energy	April 2024
E5	Energy	North American Electric Reliability Corporation (2024)	AI/ML in Real-Time System Operations	November 2024
E6	Energy	International Energy Agency (2025)	Energy and AI	April 2025
P1	Public infrastructure	Office of Management and Budget (2024)	M-24-10 Federal Agency Use of AI	March 2024
P2	Public infrastructure	Office of Management and Budget (2025)	M-25-21 Federal Use of AI	April 2025
P3	Public infrastructure	Government Accountability Office (2021)	AI Accountability Framework	June 2021
P4	Public infrastructure	Department of Homeland Security (2024)	Artificial Intelligence Roadmap	March 2024
P5	Public infrastructure	Cybersecurity and Infrastructure Security Agency (2023)	Roadmap for Artificial Intelligence	November 2023
P6	Public infrastructure	Australian Government (2024)	Voluntary AI Safety Standard	September 2024

Note: All documents were publicly available on or before May 31, 2025. M-24-10 was retained as a historical governance instrument and analyzed separately from its April 2025 replacement.

6. Empirical Results

The 24-document corpus contains 576 substantive governance-item observations and 72 readiness-item observations. Public infrastructure instruments are the most operationalized on average, followed by healthcare, finance, and energy. This order should not be interpreted as a ranking of sector performance. It reflects the content of selected public guidance, which may be more detailed where formal government obligations or safety regulation already exist.

Internal consistency was high. Cronbach's alpha ranged from 0.842 for risk, safety, and cybersecurity to 0.972 for fairness, rights, and contestability. Composite reliability ranged from 0.904 to 0.981, and average variance extracted ranged from 0.761 to 0.946. Item loadings were all above 0.75. The results support aggregation of the three indicators within each dimension. HTMT values were below 0.90. The largest values occurred between explainability and fairness and between monitoring and lifecycle control, which is theoretically plausible because traceable reasons often support contestability and monitoring often triggers controlled change.

The two-component PLS model produced a leave-one-out R-squared of 0.790 and root mean squared standardized prediction error of 0.458. Monitoring and validation had the largest positive standardized coefficient, beta = 0.460, with a 95 percent bootstrap interval from 0.319 to 0.593. Lifecycle and change control followed, beta = 0.309, interval 0.173 to 0.365. Escalation and accountability was also positive, beta = 0.229, interval 0.076 to 0.330. Organizational learning had beta = 0.118 and an interval that narrowly included zero. Intervention authority, explainability, safety, and fairness did not show independent positive coefficients after the correlated dimensions were entered together.

These null coefficients should not be read as evidence that the dimensions are unnecessary. The corpus exhibits ceiling effects in safety and strong covariance among rights, explanation, and intervention. A framework can require a capability even when cross-document variation in that capability does not independently predict the criterion. The result is consistent with a formative governance model: the index is a profile of necessary controls, not a set of interchangeable causal predictors.

The primary equal-weight HOGI correlated with operational readiness at Spearman rho = 0.643, $p < 0.001$. Entropy weighting produced rho = 0.624, $p = 0.001$. The two weighting schemes had rank correlation of 0.988, indicating that the overall ordering is not driven by a single weighting choice. Entropy weighting placed greater weight on fairness and organizational learning because those dimensions varied more across documents. Equal weighting was retained as

primary because greater dispersion does not necessarily imply greater normative importance.

Sector results show distinct profiles. Public infrastructure had the highest equal-weight HOGI at 89.8 and readiness at 87.5. Healthcare scored 86.8 and 81.9. Finance scored 76.2 and 75.0. Energy scored 68.4 and 65.3. The Kruskal-Wallis test indicated sector differences in intervention authority, explainability and traceability, fairness and rights, and the composite HOGI. Sector differences in readiness were not statistically conclusive with six documents per group.

Healthcare guidance was strongest in intervention authority, fairness, and explainability. The pattern reflects professional accountability, patient rights, and lifecycle documentation. Finance was strongest in escalation, monitoring, and change control, consistent with model-risk governance. Energy documents scored strongly in risk and cybersecurity but lower in contestability and organizational learning. Public-infrastructure documents showed the broadest coverage, particularly after federal guidance required designated AI governance roles, impact assessment, monitoring, and public accountability.

Hypothesis decisions reflect both coefficients and sector tests. H2, H3, H7, and H9 were supported. H8 received qualified support because organizational learning was positive but its bootstrap interval narrowly crossed zero. H1, H4, and H5 were not supported as independent predictors in the multivariate model, although their reliability and framework role remained strong. H6 was supported at the sector-comparison level because fairness and rights differed significantly across sectors, but its independent structural coefficient was not positive. H10 was supported for HOGI profile differences, not for a definitive readiness ranking.

Importance-performance analysis placed monitoring and validation in the highest-priority quadrant because it combined the largest coefficient with an average performance score below the strongest dimensions. Lifecycle control and escalation were the next priorities. Organizational learning had moderate importance and the lowest mean among the four positive coefficients. The practical implication is that many documents state values and safety principles, but readiness is more clearly distinguished by continuous evidence, controlled change, and the capacity to learn after deployment.

Robustness checks preserved the main interpretation. Entropy weighting changed no sector's relative position. Leave-one-document-out HOGI estimates changed sector means by less than 2.3 points. Removing OMB M-24-10, which was superseded by M-25-21 before the cutoff, reduced the public-infrastructure sample to five documents but did not change its leading position. Replacing the readiness mean with a conservative minimum-item score reduced overall levels but preserved a positive HOGI association. These tests do not remove selection uncertainty, but they show that findings are not driven by one source or one aggregation rule.

Table 3: Reliability and convergent validity

Construct	Alpha	CR	AVE	Item loading range
Intervention authority	0.930	0.957	0.880	0.927 to 0.955
Escalation and accountability	0.915	0.946	0.855	0.895 to 0.947
Monitoring and validation	0.958	0.973	0.923	0.934 to 0.977
Explainability and traceability	0.966	0.978	0.937	0.936 to 0.984
Risk, safety, and cybersecurity	0.842	0.904	0.761	0.751 to 0.940
Fairness, rights, and contestability	0.972	0.981	0.946	0.962 to 0.985
Lifecycle and change control	0.917	0.948	0.859	0.888 to 0.955
Organizational learning	0.957	0.973	0.923	0.936 to 0.975

Note: CR is composite reliability. AVE is average variance extracted. Values are exploratory document-coding diagnostics.

Table 4: Exploratory PLS structural results

Code	Construct	Beta	95% bootstrap CI	p	Assessment
IA	Intervention authority	-0.028	[-0.132, 0.089]	0.788	Not supported
EA	Escalation and accountability	0.229	[0.076, 0.330]	0.004	Supported
MV	Monitoring and validation	0.460	[0.319, 0.593]	<.001	Supported
ET	Explainability and traceability	-0.040	[-0.110, 0.045]	0.413	Not supported
RS	Risk, safety, and cybersecurity	-0.073	[-0.268, 0.166]	0.558	Not supported
FR	Fairness, rights, and contestability	-0.077	[-0.146, 0.059]	0.235	Not supported
LC	Lifecycle and change control	0.309	[0.173, 0.365]	<.001	Supported
OL	Organizational learning	0.118	[-0.001, 0.224]	0.053	Qualified

Note: Two PLS components; 5,000 document-level bootstrap resamples. LOOCV R2 = 0.790; standardized RMSE = 0.458.

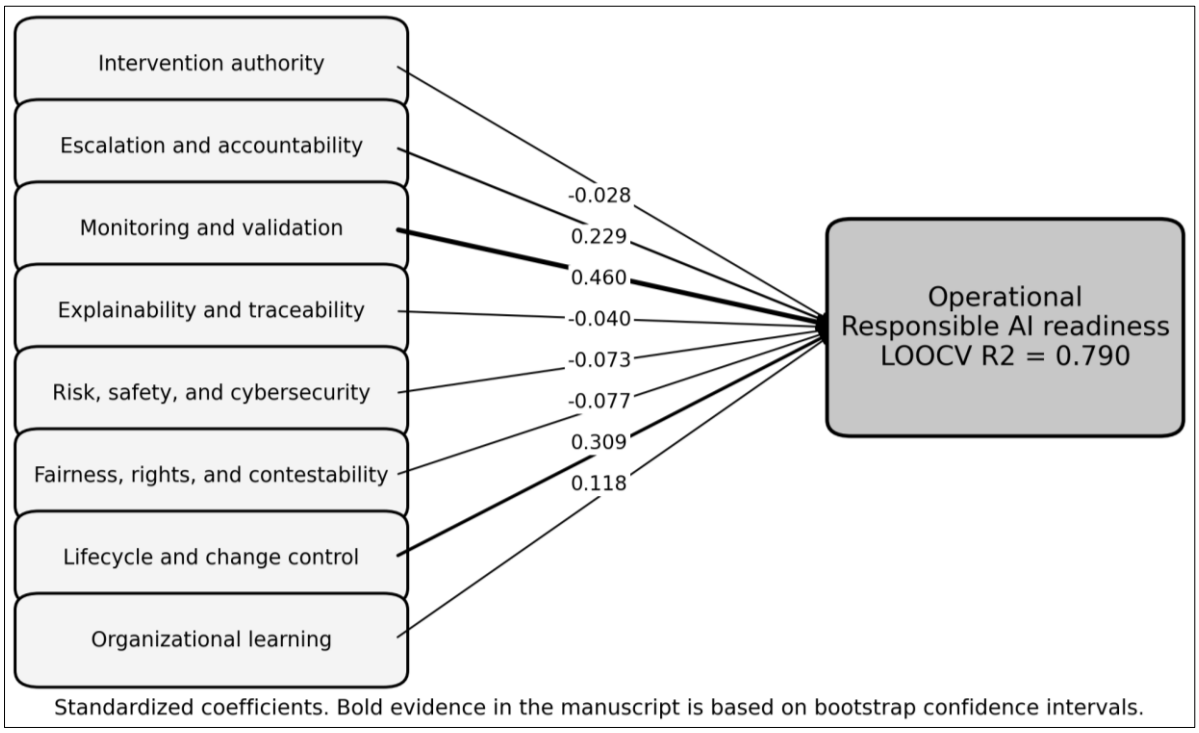

Fig 4: Exploratory structural model.

Table 5: Cross-sector governance comparison

Sector	IA	EA	MV	ET	RS	FR	LC	OL	HOGI	Ready
Healthcare	88.9	88.9	87.5	90.3	91.7	80.6	90.3	76.4	86.8	81.9
Finance	70.8	84.7	79.2	77.8	90.3	63.9	84.7	58.3	76.2	75.0
Energy	65.3	70.8	70.8	66.7	87.5	58.3	73.6	54.2	68.4	65.3
Public infrastructure	88.9	93.1	86.1	90.3	97.2	90.3	91.7	80.6	89.8	87.5

Note: Scores range from 0 to 100. IA intervention authority; EA escalation and accountability; MV monitoring and validation; ET explainability and traceability; RS risk and safety; FR fairness and rights; LC lifecycle control; OL organizational learning.

7. Discussion

The results answer the central research problem by showing that meaningful human oversight is an operational system rather than a checkpoint. Monitoring, controlled change, escalation, and learning distinguish governance instruments that are closer to implementation. This pattern is important for autonomous agents because their risk arises over sequences of actions. A one-time approval cannot determine whether later tool calls remain appropriate, whether the environment has changed, or whether accumulated memory has altered behavior.

Monitoring and validation had the strongest relationship with readiness. This result aligns with the NIST and GAO lifecycle view and with sector guidance that requires ongoing performance review (Government Accountability Office, 2021; National Institute of Standards and Technology, 2023). For agents, monitoring should include more than accuracy. Organizations need action success, constraint violations, tool-call anomalies, uncertainty, escalation frequency, override outcomes, security events, and impact on affected groups. A monitoring dashboard without an intervention route is incomplete, but an intervention route without monitoring is reactive.

Lifecycle control was the second strongest predictor. Traditional model governance often treats a new model version as the unit of change. Agents change through prompts, tools, permissions, retrieval sources, memory, and external services. Change control must therefore cover the entire agent configuration, including dependencies that otherwise accumulate as hidden technical debt (Sculley *et al*., 2015). The FDA predetermined change-control approach illustrates the value of specifying anticipated modifications and validation procedures before deployment. Financial model-risk guidance similarly requires inventories, validation, and controlled changes. The cross-sector lesson is that autonomy should be an authorized configuration, not a permanent characteristic of a model.

Escalation and accountability also mattered. A system can detect risk but still fail when no person owns the response. Effective escalation needs thresholds that reflect consequence and reversibility. A low-impact, reversible task may permit post-action review. A clinical order, payment block, grid-control action, or public-benefit determination may require approval before execution. The decision-escalation workflow in Figure 5 separates the risk gate from the human decision and records the outcome for learning.

The lack of independent positive coefficients for intervention, explainability, safety, and fairness is instructive. These dimensions had high scores and strong correlations with other controls. Their effects may operate through monitoring, escalation, or lifecycle practice. Explainability is useful when it supplies evidence to a person who has authority and time to act. Safety statements become operational when linked to tests and stop criteria. Fairness provisions become effective when they trigger review and correction. The result therefore rejects a checklist interpretation in which each principle independently causes readiness.

Sector profiles reveal different institutional histories. Finance has mature model-risk practices but public documents devote less attention to direct contestability than healthcare or public-sector guidance. Energy governance emphasizes reliability and cyber risk, yet affected-person rights are less developed because many applications are infrastructure-facing. Public infrastructure has broad accountability requirements because government action must be justified to the public, although implementation across agencies may vary. Healthcare guidance combines professional responsibility with patient-centered rights. HOGF-AI does not erase these differences. It provides a common architecture within which sector-specific thresholds can be defined.

The findings also qualify the idea that human oversight should minimize autonomy. High reliability may require automated reaction when a human cannot respond quickly enough. The governance objective is bounded autonomy with human command over objectives, limits, and exceptional states. A mature organization can authorize greater autonomy when controls are tested, logs are complete, rollback is reliable, and incident learning is demonstrated. Supplementary Figure S6 captures this progression.

Finally, the study adds to the work of Rasel, Kamrul Islam, and Fahim by connecting sector applications to a common oversight structure. Healthcare security and supply-chain resilience show why institutional continuity and human authority matter (Hasan *et al*., 2022; Rasel *et al*., 2022). Financial cybersecurity and FinTech accountability show the need for governance alongside prediction (Fahim *et al*., 2023; Hasan *et al*., 2023). Multimodal lending and market-surveillance research demonstrate that richer models create explanation and fairness obligations (Jahan *et al*., 2024; Rasel *et al*., 2023). Clean-energy, waste, and energy-finance applications extend the same problem to sustainability and infrastructure (Arman, Hasan, *et al*., 2024; Arman, Rasel, *et al*., 2024; Rabbani *et al*., 2024). HOGF-AI organizes these concerns around action authority and lifecycle evidence.

The distinction between human oversight and human burden deserves emphasis. Poorly designed governance can transfer operational risk to reviewers by requiring frequent approval without improving evidence. Review queues can become ceremonial when staff face time pressure, repeated low-value alerts, or unclear authority. HOGF-AI addresses this problem through risk-tiered escalation. Routine actions remain automated when they are reversible and monitored. Human attention is reserved for uncertainty, policy conflict, novel states, high consequence, or evidence of control failure. This design treats scarce human judgment as a governance resource.

The framework also clarifies the relationship between oversight and cybersecurity. An agent may be technically accurate yet unsafe because an attacker manipulates its prompt, tool, data source, or memory. Conversely, an aggressive security control can create service denial or harmful interruption. Cyber monitoring must therefore feed the same escalation and intervention system as performance monitoring. Security operations, model owners, domain professionals, and compliance teams need preassigned roles. A critical alert should not depend on informal coordination after the event.

Organizational learning is especially important for multi-agent environments. One agent's failure may expose a weak tool, ambiguous policy, or shared data problem that affects other agents. Portfolio governance should aggregate incidents and overrides across systems, identify common dependencies, and revise reusable controls. This is where the framework moves beyond isolated model cards. Documentation remains necessary, but learning requires a governance forum with authority to change standards, budgets, staffing, and permitted autonomy across the organization.

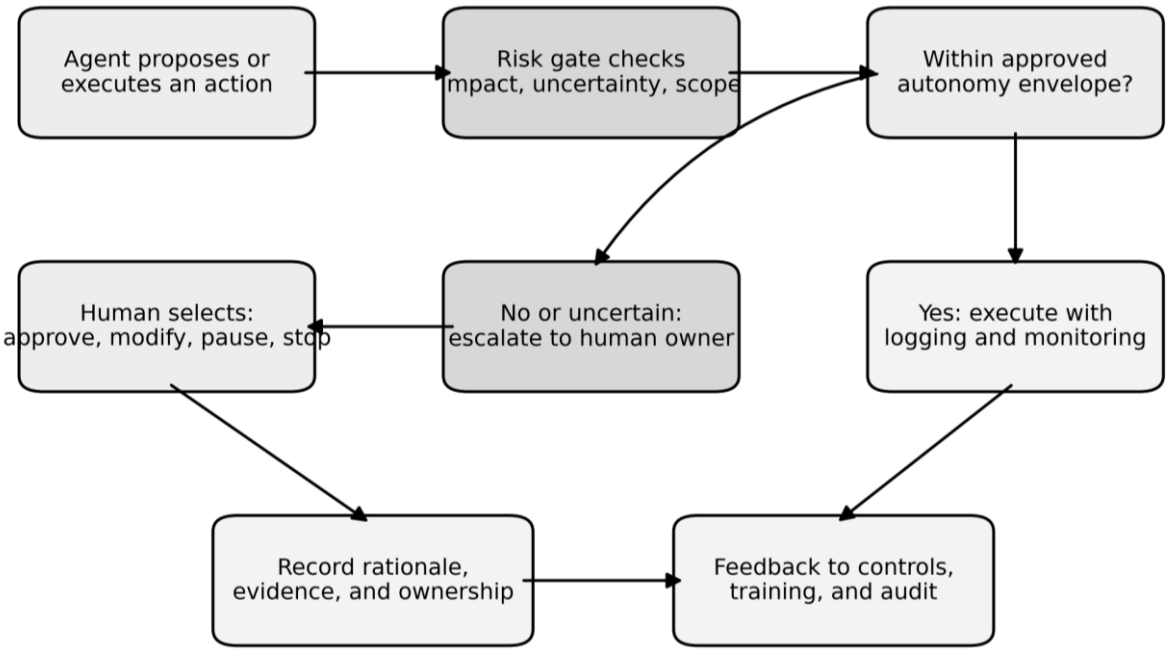

Fig 5: AI decision-escalation workflow.

8. Theoretical, Practical, and Policy Implications

The theoretical contribution is a shift from human presence to oversight capability. Human-in-the-loop labels describe location in a workflow but not institutional effectiveness. HOGF-AI defines oversight as a multidimensional capability: authority, escalation, monitoring, traceability, safety, rights, change control, and learning. This formulation reconciles human automation theory with accountability. It also supports a formative interpretation. A system may be highly explainable but still unaccountable if nobody can stop it, or safe in routine operation but fragile under change.

The framework also extends dynamic capabilities theory. Sensing appears in monitoring, seizing in escalation and intervention selection, and reconfiguration in change control and organizational learning. Information-processing theory explains why high interdependence requires richer evidence and faster escalation. Organizational learning becomes a measurable control rather than a general aspiration. These connections provide a basis for future organization-level testing.

For managers, the first task is an agent inventory. Each entry should identify purpose, owner, model, tools, data, users, affected parties, autonomy level, prohibited actions, escalation route, and retirement criteria. The second task is to define an autonomy envelope. The envelope should specify which actions are automatic, which require approval, and which are prohibited. Risk gates should combine impact, uncertainty, reversibility, novelty, and security signals. A single confidence score is insufficient.

The third task is to engineer intervention. Pause and stop functions must work under realistic load and cyber conditions. A human reviewer needs primary evidence, not only a generated explanation. Temporary overrides should expire, and emergency autonomy should have a time limit. The fourth task is continuous assurance. Independent validation should test behavior across scenarios, tool failures, prompt injection, data drift, group impacts, and escalation latency. Audit logs should support replay without exposing unnecessary sensitive data.

The implementation roadmap begins with ownership and scope, then adds logs and risk gates, independent assurance, and portfolio learning. The HOGI can be used as a baseline and repeated after controls are implemented. It should not be used as a certification score without evidence of operating effectiveness. Organizations should report dimension profiles because the total can conceal a weak right of appeal or missing stop authority.

Policy makers should avoid a generic requirement for human oversight. Rules should specify expected capability in proportion to consequence. High-impact agents should have named accountable owners, documented autonomy limits, predeployment testing, continuous monitoring, stop authority, incident reporting, and contestability where individual interests are affected. Procurement rules should require access to logs, change notifications, and exit plans. Vendor claims that a system is explainable or human supervised should be testable.

Sector policy should preserve context. Healthcare should link agent action to licensed professional responsibility and patient safety. Finance should integrate agents into model-risk, consumer-protection, and cybersecurity governance. Energy should define fail-safe states and manual fallback consistent with reliability obligations. Public infrastructure should emphasize public notice, due process, records management, and continuity of essential services. Cross-sector standards can define the control architecture, while sector authorities define the thresholds.

Internationally, the framework can support crosswalks among NIST, OECD, ISO, the EU AI Act, and sector guidance. The purpose is not to replace legal requirements. It is to translate them into a common operational vocabulary that can be audited across an agent portfolio.

9. Limitations and Future Research

The study has several limitations. The unit of analysis is a governance document, not an organization or deployed agent. A high document score does not demonstrate implementation. The sample contains 24 purposively selected instruments and is not a census of global governance. Sector means are sensitive to document selection, and six observations per sector limit inferential power.

One coder completed the final scoring. The two-pass protocol and full coding matrix improve transparency, but they do not substitute for intercoder reliability. Future work should use independent coders, expert panels, and organizations from multiple jurisdictions. The readiness criterion is separate from the eight dimensions but remains conceptually related to governance operationalization. The PLS coefficients therefore describe association within the corpus, not causal effects.

Autonomous-agent governance is evolving quickly. The May 31, 2025 cutoff preserves historical integrity but excludes later standards, enforcement actions, and agent-specific practices. Future research should validate HOGF-AI through Delphi studies, organizational surveys, field audits, and incident data. Longitudinal designs could examine whether higher baseline scores predict fewer harmful events, faster intervention, better appeal outcomes, or more reliable recovery. Human-subject experiments should test whether escalation interfaces improve calibrated intervention rather than simply increasing trust.

The framework should also be tested at different autonomy levels. A document-summarization agent, a fraud-control agent, and a grid-control agent require different latency and authority. Future versions of HOGI could include consequence-weighted scoring and minimum-control floors so that high performance in one dimension cannot compensate for absence of a critical safeguard.

10. Conclusion

Autonomous AI agents require governance that remains active after deployment. The cross-sector evidence shows that operational readiness is most clearly associated with continuous monitoring, controlled change, escalation, accountability, and organizational learning. Human oversight is meaningful only when people have timely evidence, practical authority, and institutional support to intervene.

HOGF-AI provides an eight-dimension governance architecture and HOGI provides a transparent benchmark. The framework does not require humans to approve every low-risk action. It requires organizations to define autonomy, detect exceptions, preserve stop authority, document decisions, protect rights, and learn from outcomes. Greater autonomy becomes defensible when control evidence matures with it. In high-stakes environments, responsibility cannot be delegated to an agent. It remains with the institution that authorizes, monitors, and benefits from the system.

References

1. Amershi S, Weld D, Vorvoreanu M, Fourney A, Nushi B, Collisson P, Suh J, Iqbal S, Bennett PN, Inkpen K, Teevan J, Kikin-Gil R, Horvitz E. Guidelines for human-AI interaction. In: Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems; 2019 May 4-9; Glasgow, UK. New York: ACM; 2019. p. 1-13. doi: 10.1145/3290605.3300233.
2. Arman M, Hasan MN, Rasel IH. Clean energy transition in USA: Big data analytics for renewable energy forecasting and carbon reduction. J Manag World. 2024;2024(3):192-206. doi: 10.53935/jomw.v2024i4.1196.
3. Arman M, Rasel IH, Razib MNH, Fahim ASM. Big data and machine learning for sustainable waste reduction. J Posthumanism. 2024;4(2):448-467. doi: 10.63332/joph.v4i2.3361.
4. Australian Government, Department of Industry, Science and Resources. Voluntary AI safety standard. Canberra: Australian Government; 2024.
5. Bansal G, Wu T, Zhou J, Fok R, Nushi B, Kamar E, Ribeiro MT, Weld DS. Does the whole exceed its parts? The effect of AI explanations on complementary team performance. In: Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems; 2021 May 8-13; Yokohama, Japan. New York: ACM; 2021. Article 81. doi: 10.1145/3411764.3445717.
6. Board of Governors of the Federal Reserve System, Office of the Comptroller of the Currency. Supervisory guidance on model risk management. Washington, D.C.: Federal Reserve; 2011. SR 11-7/OCC Bulletin 2011-12.
7. Bovens M. Analysing and assessing accountability: A conceptual framework. Eur Law J. 2007;13(4):447-468. doi: 10.1111/j.1468-0386.2007.00378.x.
8. Buçinca Z, Malaya MB, Gajos KZ. To trust or to think: Cognitive forcing functions can reduce overreliance on AI in AI-assisted decision-making. Proc ACM Hum-Comput Interact. 2021;5(CSCW1):Article 188. doi: 10.1145/3449287.
9. Coalition for Health AI. Blueprint for trustworthy AI implementation guidance and assurance for healthcare. CHAI; 2023.
10. Cybersecurity and Infrastructure Security Agency. Roadmap for artificial intelligence. Washington, D.C.: CISA; 2023.
11. Department of Homeland Security. Artificial intelligence roadmap 2024. Washington, D.C.: DHS; 2024.
12. Diakopoulos N. Accountability in algorithmic decision making. Commun ACM. 2016;59(2):56-62. doi: 10.1145/2844110.
13. European Union. Regulation (EU) 2024/1689 laying down harmonised rules on artificial intelligence. Off J Eur Union. 2024;L 1689.
14. Fahim ASM, Ibrahim M, Pritty AA, Tania TA. Algorithmic accountability in U.S. consumer FinTech: Governance mechanisms for credit risk, fair lending, and financial stability. J Econ Finance Account Stud. 2023;5(4):80-93. doi: 10.32996/jefas.2023.5.4.8.
15. Financial Industry Regulatory Authority. Artificial intelligence in the securities industry. Washington, D.C.: FINRA; 2020.
16. Financial Stability Board. The financial stability implications of artificial intelligence. Basel: FSB; 2024.
17. Gebru T, Morgenstern J, Vecchione B, Vaughan JW, Wallach H, Daumé III H, Crawford K. Datasheets for datasets. Commun ACM. 2021;64(12):86-92. doi: 10.1145/3458723.
18. Government Accountability Office. Artificial intelligence: An accountability framework for federal agencies and other entities. Washington, D.C.: GAO; 2021. GAO-21-519SP.
19. Hasan MN, Rasel IH, Arman M, Ibrahim M, Jahan N. Strengthening U.S. financial and cybersecurity infrastructure with AI-driven fraud detection and risk analytics. J Comput Anal Appl. 2023;31(2):15-32.
20. Hasan MN, Rasel IH, Rahman M, Islam K, Arman M, Jahan N. Securing U.S. healthcare infrastructure with machine learning: Protecting patient data as a national security priority. Int J Comput Exp Sci Eng. 2022;8(3):85-93. doi: 10.22399/ijcesen.3987.
21. Hoff KA, Bashir M. Trust in automation: Integrating empirical evidence on factors that influence trust. Hum Factors. 2015;57(3):407-434. doi: 10.1177/0018720814547570.
22. International Energy Agency. Energy and AI. Paris: IEA; 2025.
23. International Organization for Standardization. ISO/IEC 42001:2023 information technology, artificial intelligence, management system. Geneva: ISO; 2023.
24. Islam MK, Rabbani SF, Rahat YO. Modeling the consumer transition from cash to digital payments and its implications for financial system efficiency, transparency, and inclusion. J Econ Finance Account Stud. 2025;7(3):118-134. doi: 10.32996/jefas.2025.7.3.11.
25. Jahan N, Pritty AA, Ibrahim M, Zadid MU, Fahim ASM, Mahmud S. Machine learning-driven early warning analytics for identifying market manipulation, irregular trading activity, and suspicious market signals in U.S. stock markets. J Comput Sci Technol Stud. 2024;6(2):257-283. doi: 10.32996/jcsts.2024.6.2.26.
26. Jarrahi MH. Artificial intelligence and the future of work: Human-AI symbiosis in organizational decision making. Bus Horiz. 2018;61(4):577-586. doi: 10.1016/j.bushor.2018.03.007.
27. Jobin A, Ienca M, Vayena E. The global landscape of AI ethics guidelines. Nat Mach Intell. 2019;1:389-399. doi: 10.1038/s42256-019-0088-2.
28. Kamruzzaman M, Saha S, Islam MK. Augmented intelligence in healthcare: Bridging machine learning, human expertise, and business process automation. Int J Eng Technol Res Manag. 2023;7(6):250-264.
29. Kroll JA, Huey J, Barocas S, Felten EW, Reidenberg JR, Robinson DG, Yu H. Accountable algorithms. Univ Pa Law Rev. 2017;165(3):633-705.
30. Lee JD, See KA. Trust in automation: Designing for appropriate reliance. Hum Factors. 2004;46(1):50-80. doi: 10.1518/hfes.46.1.50_30392.
31. Lundberg SM, Lee SI. A unified approach to interpreting model predictions. In: Advances in Neural Information Processing Systems 30 (NeurIPS 2017); 2017 Dec 4-9; Long Beach, CA. Red Hook: Curran; 2017. p. 4765-4774.
32. Mitchell M, Wu S, Zaldivar A, Barnes P, Vasserman L, Hutchinson B, Spitzer E, Raji ID, Gebru T. Model cards for model reporting. In: Proceedings of the Conference on Fairness, Accountability, and Transparency; 2019 Jan

29-31; Atlanta, GA. New York: ACM; 2019. p. 220-229. doi: 10.1145/3287560.3287596.
33. Mittelstadt B. Principles alone cannot guarantee ethical AI. Nat Mach Intell. 2019;1:501-507. doi: 10.1038/s42256-019-0114-4.
34. Monetary Authority of Singapore. Veritas initiative: FEAT assessment methodology and case studies. Singapore: MAS; 2022.
35. National Institute of Standards and Technology. Artificial intelligence risk management framework (AI RMF 1.0). Gaithersburg: NIST; 2023. NIST AI 100-1. doi: 10.6028/NIST.AI.100-1.
36. North American Electric Reliability Corporation. Artificial intelligence and machine learning in real-time system operations. Atlanta: NERC; 2024.
37. OECD. Recommendation of the Council on artificial intelligence. Paris: OECD; 2024.
38. Office of Management and Budget. Advancing governance, innovation, and risk management for agency use of artificial intelligence. Washington, D.C.: OMB; 2024. M-24-10.
39. Office of Management and Budget. Accelerating federal use of AI through innovation, governance, and public trust. Washington, D.C.: OMB; 2025. M-25-21.
40. Parasuraman R, Sheridan TB, Wickens CD. A model for types and levels of human interaction with automation. IEEE Trans Syst Man Cybern A. 2000;30(3):286-297. doi: 10.1109/3468.844354.
41. Park JS, O'Brien JC, Cai CJ, Morris MR, Liang P, Bernstein MS. Generative agents: Interactive simulacra of human behavior. In: Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology; 2023 Oct 29-Nov 1; San Francisco, CA. New York: ACM; 2023. Article 2. doi: 10.1145/3586183.3606763.
42. Pritty AA, Ibrahim M, Fahim ASM, Zadid MU. Generative AI and U.S. financial reporting integrity: Detecting narrative manipulation, risk disclosure gaming, and fraud signals in 10-K filings. J Econ Finance Account Stud. 2024;6(4):113-129. doi: 10.32996/jefas.2024.6.4.11.
43. Prudential Regulation Authority. Model risk management principles for banks. London: Bank of England; 2023. SS1/23.
44. Rabbani SF, Rahat YO, Islam MK. From utility bills to retrofit finance: An AI framework for energy-burden-aware underwriting in residential decarbonization. Front Comput Sci Artif Intell. 2024;3(2):72-88. doi: 10.32996/fcsai.2024.3.2.10.
45. Rahat YO, Islam MK, Rabbani SF. Safeguarding federal payment infrastructure: Trustworthy AI for improper-payment prevention, synthetic identity detection, and public-benefit disbursement integrity in the United States. J Bus Manag Stud. 2024;6(5):238-251. doi: 10.32996/jbms.2024.6.5.25.
46. Raisch S, Krakowski S. Artificial intelligence and management: The automation-augmentation paradox. Acad Manag Rev. 2021;46(1):192-210. doi: 10.5465/amr.2018.0072.
47. Raji ID, Smart A, White RN, Mitchell M, Gebru T, Hutchinson B, Smith-Loud J, Theron D, Barnes P. Closing the AI accountability gap: Defining an end-to-end framework for internal algorithmic auditing. In: Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency; 2020 Jan 27-30; Barcelona, Spain. New York: ACM; 2020. p. 33-44. doi: 10.1145/3351095.3372873.
48. Rasel IH, Arman M, Hasan MN, Bhuyain MMH. Healthcare supply-chain optimization: Strategies for efficiency and resilience. J Med Health Stud. 2022;3(4):171-182. doi: 10.32996/jmhs.2022.3.4.26.
49. Rasel IH, Ibrahim M, Pritty AA, Fahim ASM, Jahan N. Beyond FICO: Enhancing mortgage default forecasting and inclusive lending via multimodal graph neural networks and urban mobility analytics. Front Comput Sci Artif Intell. 2023;2(2):62-81. doi: 10.32996/fcsai.2023.2.2.5.
50. Ribeiro MT, Singh S, Guestrin C. "Why should I trust you?" Explaining the predictions of any classifier. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining; 2016 Aug 13-17; San Francisco, CA. New York: ACM; 2016. p. 1135-1144. doi: 10.1145/2939672.2939778.
51. Russell S, Norvig P. Artificial intelligence: A modern approach. 4th ed. Hoboken: Pearson; 2021.
52. Schick T, Dwivedi-Yu J, Dessì R, Raileanu R, Lomeli M, Zettlemoyer L, Cancedda N, Scialom T. Toolformer: Language models can teach themselves to use tools. In: Advances in Neural Information Processing Systems 36 (NeurIPS 2023); 2023 Dec 10-16; New Orleans, LA. 2023.
53. Sculley D, Holt G, Golovin D, Davydov E, Phillips T, Ebner D, Chaudhary V, Young M, Crespo JF, Dennison D. Hidden technical debt in machine learning systems. In: Advances in Neural Information Processing Systems 28 (NeurIPS 2015); 2015 Dec 7-12; Montreal, Canada. 2015.
54. Shinn N, Cassano F, Gopinath A, Narasimhan K, Yao S. Reflexion: Language agents with verbal reinforcement learning. In: Advances in Neural Information Processing Systems 36 (NeurIPS 2023); 2023 Dec 10-16; New Orleans, LA. 2023.
55. Shneiderman B. Human-centered artificial intelligence: Reliable, safe and trustworthy. Int J Hum-Comput Interact. 2020;36(6):495-504. doi: 10.1080/10447318.2020.1741118.
56. U.S. Department of Energy. Artificial intelligence and machine learning for energy systems: Report to the Secretary of Energy. Washington, D.C.: DOE; 2020.
57. U.S. Department of Energy. Implementation pathway for responsible artificial intelligence in fossil energy and carbon management. Washington, D.C.: DOE; 2023.
58. U.S. Department of Energy. Department of Energy artificial intelligence compliance plan. Washington, D.C.: DOE; 2024.
59. U.S. Department of Energy. AI for energy: Opportunities for a modern grid and clean energy economy. Washington, D.C.: DOE; 2024.
60. U.S. Department of the Treasury. Managing artificial intelligence-specific cybersecurity risks in the financial services sector. Washington, D.C.: Treasury; 2024.
61. U.S. Food and Drug Administration. Artificial intelligence/machine learning-based software as a medical device action plan. Silver Spring: FDA; 2021.
62. U.S. Food and Drug Administration. Marketing submission recommendations for a predetermined change control plan for artificial intelligence-enabled device software functions: Guidance for industry and

Food and Drug Administration staff. Silver Spring: FDA; 2024.
63. Wachter S, Mittelstadt B, Russell C. Counterfactual explanations without opening the black box: Automated decisions and the GDPR. Harv J Law Technol. 2018;31(2):841-887.
64. Wooldridge M. An introduction to multiagent systems. 2nd ed. Chichester: Wiley; 2009.
65. World Health Organization. Ethics and governance of artificial intelligence for health: WHO guidance. Geneva: WHO; 2021.
66. World Health Organization. Regulatory considerations on artificial intelligence for health. Geneva: WHO; 2023.
67. World Health Organization. Ethics and governance of artificial intelligence for health: Guidance on large multi-modal models. Geneva: WHO; 2025.
68. Yao S, Zhao J, Yu D, Du N, Shafran I, Narasimhan K, Cao Y. ReAct: Synergizing reasoning and acting in language models. In: International Conference on Learning Representations (ICLR 2023); 2023 May 1-5; Kigali, Rwanda. 2023.