Human Oversight in Production AI Systems: A Framework for Evaluating Sufficiency Across Levels of Autonomy

Nikhil Purwaha[1]

[1] Faculty of Business Administration, Thunderbird School of Global Management, Phoenix, USA

Correspondence: Nikhil Purwaha, Faculty of business Administration, Thunderbird School of Global Management, Phoenix, Arizona 85004, USA. E-mail: Nikhil.purwaha@gmail.com

Received: April 16, 2026 Accepted: May 19, 2026 Online Published: June 9, 2026

doi:10.5430/jbar.v15n1p44 URL: https://doi.org/10.5430/jbar.v15n1p44

Abstract

Fast deployment of artificial intelligence (AI) in the production setting has raised concerns over accountability, transparency, and risk. While some research discusses HITL, much of it is conceptual, providing little guidance on how organizations must implement and evaluate human oversight in real-world systems. This paper fills that gap with a practitioner-centered assessment of the suitability of human oversight across AI autonomy levels.

The paper presents a taxonomy of autonomy levels, from assisted intelligence to fully autonomous systems, and their associated oversight requirements. Oversight sufficiency is a multidimensional construct, including observability, intervention capability, timeliness, accountability, and adaptability control. Based on this matrix, the study develops an oversight sufficiency matrix that classifies systems as sufficient, marginal, or insufficient based on the relationship between autonomy, risk, and human control.

The framework includes practical cases from the realms of finance, healthcare, cybersecurity, and digital platforms. It identifies architectural approaches and barriers, such as latency, scalability, and human cognitive limitations. The methodology included measurable measures of intervention success rate, detection latency, and auditability that are easily measured and benchmarked.

The results show that simply having a human overseeing an AI system is not enough to ensure it is being properly managed; it must also be proportionate to its autonomy and risk. By providing a clear and actionable framework, this paper contributes to bridge the gap between theoretical models of governance for AI and its commercial deployment, providing researchers and practitioners with tools to design, evaluate, and regulate responsible AI systems.

Keywords: AI Ethics, AI Governance, AI Risk Management, Explainable AI (XAI), Human Oversight, Decision-Making Systems, Oversight Framework, Socio-Technical Systems

1. Introduction

The adoption of artificial intelligence (AI) into production systems has gained steam in different industries, revolutionizing the way organizations think, act, and deliver services. AI is now trusted with tasks previously performed or supervised by humans—whether it's fraud detection in financial institutions, decision-support in the clinic, or detecting threats autonomously in the cybersecurity field. While AI systems have increased efficiency, scalability, and predictive abilities, increasing autonomy has brought risks in the areas of accountability, transparency, and risk management.

The role of human oversight is also an aspect of this transition. While the concept of Human-in-the-Loop (HITL) has been promoted as a way to ensure that AI systems follow human values and goals, in reality, human oversight in production settings is not what it looks like. In many cases, human involvement is minimal, delayed, or symbolic, raising the question of whether such oversight is effective or simply procedural. As AI systems become more sophisticated and more real-time, humans have to be in the loop, but also to ensure that human participation is meaningful, timely, and can impact outcomes.

Some published research has analyzed human-centered AI, explainability, and algorithmic accountability. But, much of this research is conceptual, offering no practical guidance on how to assess whether their oversight measures are sufficient given the autonomy of their AI systems. For example, there are no structured models to link autonomy of the system, risk, and oversight in a way that can be applied in practice. This gap is becoming increasingly important in the era of high autonomy systems, where inadequate oversight can cause large losses.

This paper tries to fill this gap by providing a practitioner-oriented framework for evaluating the sufficiency of human oversight in production AI systems. We argue that oversight is not an optional attribute, present or absent, but an observable and contextually dependent dimension according to the autonomy and risk of the system. We propose a taxonomy of AI autonomy levels and assess the sufficiency of oversight on several dimensions including observability, intervention, timeliness, accountability, and adaptability control.

In this vein, this paper proposes a framework to allow organisations to distinguish between human-supervised AI and artificial intelligence (AI). The framework is designed to be analytically rigorous but also practical by considering the requirements of AI implementation, such as latency, scaling issues, and human cognition. This paper identifies the need to go beyond the abstract, and consider the operational reality of AI governance to bridge the gap between theoretical models of AI governance and the realities of machine usage.

This study has three contributions: First, it provides a taxonomy for AI autonomy levels that can be used as a context for the supervision needs. Second, it introduces a multi-dimensional model of oversight neediness and provides criteria for assessing it. Third, it provides practical information for practitioners and policymakers working to design and regulate AI systems in a manner that is effective and accountable.

In the end, this research advances the understanding of the role of human oversight in AI by presenting it as dynamic and context-sensitive rather than static design parameter. As organizations continue to embrace more autonomous systems, the evaluation and monitoring of sufficient oversight will become a critical part of implementing AI in a responsible manner.

2. Literature Review

2.1 Human-in-the-Loop (HITL) and Human-on-the-Loop (HOTL)

Human involvement in AI systems has been widely discussed through the concepts of Human-in-the-Loop (HITL) and Human-on-the-Loop (HOTL). HITL refers to systems in which human agents are directly involved in the decision-making process by providing input, validating outputs, or making final decisions (Ottun & Flores, 2025). In these systems, AI primarily functions as a decision-support tool that augments rather than replaces human judgment (Frenette, 2023). HITL models are particularly common in high-risk domains such as healthcare and finance, where ethical obligations and regulatory requirements necessitate direct human control over critical decisions (O'Sullivan et al., 2019).

In contrast, HOTL systems delegate primary decision-making authority to AI systems while maintaining human supervision over operational processes (Sadovski et al., 2025). Humans monitor the system and intervene only when necessary, typically through override mechanisms or exception-handling procedures (Tsamados et al., 2025). This oversight model is frequently applied in environments requiring high-speed processing and scalability, such as cybersecurity systems and large digital platforms (Salgado-Criado, 2025).

The distinction between HITL and HOTL is not merely architectural but functional, reflecting differences in authority allocation, operational responsibility, and the timing of human intervention (Methnani et al., 2021). Humans may assume multiple oversight roles within these systems. As decision-makers, they retain ultimate authority over outcomes, particularly in HITL environments (Ottun & Flores, 2025). As supervisors, they monitor system outputs and intervene when anomalies or risks emerge (Sadovski et al., 2025). As fallback authorities, they serve as safety mechanisms when automated processes fail or encounter uncertainty (Frenette, 2023).

However, the effectiveness of these roles depends heavily on situational awareness, cognitive workload, and system design (Endsley, 2017). Endsley (2017) argues that human operators often struggle to maintain awareness in highly automated environments, particularly when systems operate at speeds beyond human cognitive capacity. Similarly, Narayanan and Feigh (2026) demonstrate that increasing AI dependency and information abstraction can reduce human supervisory effectiveness within collaborative decision-making teams.

Research increasingly suggests that meaningful oversight requires more than nominal human participation. Methnani et al. (2021) argue that variable autonomy models, where humans dynamically adjust automation levels based on operational context and risk, provide stronger foundations for meaningful human control. Tsamados et al. (2025) further emphasize that oversight is evolving from passive supervision toward collaborative human-AI teaming, where humans and AI systems continuously interact rather than relying on isolated intervention points. Holzinger et al. (2025) similarly question whether meaningful oversight remains achievable as AI systems become increasingly autonomous and adaptive.

2.2 AI Autonomy and Decision-Making Systems

The level of autonomy exhibited by an AI system directly influences the type and intensity of human oversight required (Kandikatla & Radeljic, 2025). At lower levels, rule-based systems operate on predefined logic and require substantial

human input during both development and execution (Endsley, 2017). As systems evolve into machine learning and adaptive models, they gain the capacity to learn from data, dynamically adjust behavior, and make increasingly complex decisions with reduced human intervention (Holzinger et al., 2025).

Fully autonomous systems represent the highest level of AI independence, operating in uncertain and rapidly changing environments without real-time human participation (James, 2026). This progression from rule-based systems to autonomous adaptive systems reflects a broader shift from deterministic decision-making toward probabilistic and data-driven reasoning (Frenette, 2023). While such evolution improves scalability and operational efficiency, it also introduces significant challenges related to explainability, predictability, and control (Herrmann, 2025).

Autonomy frameworks in other domains similarly describe automation as a continuum ranging from full human control to complete automation, with intermediate stages of shared authority and collaborative operation (Endsley, 2017). These frameworks highlight the importance of aligning oversight strategies with system autonomy, since higher levels of automation typically require more sophisticated monitoring, auditing, and intervention mechanisms (Methnani et al., 2021).

However, AI systems differ from conventional automated systems because they possess the ability to adapt and evolve over time (Holzinger et al., 2025). Adaptive behavior increases uncertainty and reduces operational transparency, thereby complicating human supervision and governance (Frenette, 2023). Tsamados et al. (2025) note that modern AI oversight models increasingly prioritize cooperative human-AI interaction rather than static supervisory relationships, reflecting the growing complexity of autonomous decision-making systems. Kandikatla and Radeljic (2025) further argue that oversight intensity should scale proportionally with both system autonomy and operational risk exposure.

2.3 AI Governance and Accountability

The increasing deployment of AI systems has accelerated the development of governance frameworks intended to ensure accountability, fairness, transparency, and ethical compliance (Koulu, 2020). Many contemporary governance approaches employ risk-based models that classify AI systems according to their societal and operational impact (Kandikatla & Radeljic, 2025). Applications involving healthcare, finance, critical infrastructure, or public safety are generally categorized as high-risk and therefore require stricter oversight, validation, and documentation procedures (O'Sullivan et al., 2019).

Central to AI governance are the principles of explainability, transparency, and auditability (Herrmann, 2025). Explainability refers to the ability of stakeholders to understand and interpret AI-generated decisions (Ottun & Flores, 2025). Transparency involves clear documentation of system architecture, datasets, operational logic, and governance processes (Salgado-Criado, 2025). Auditability ensures that AI decisions and actions can be traced, reviewed, and evaluated retrospectively (Verdiesen et al., 2021). Together, these principles form the foundation of trustworthy AI governance and support both regulatory compliance and organizational accountability (Holzinger et al., 2025).

Legal and ethical scholarship further emphasizes that governance structures must ensure humans retain meaningful authority over AI systems, particularly in high-risk operational environments (O'Sullivan et al., 2019). Verdiesen et al. (2021) argue that comprehensive human oversight is essential for maintaining accountability in autonomous systems where decisions may produce significant ethical or societal consequences. Laitinen and Sahlgren (2021) additionally emphasize that preserving human autonomy and dignity must remain central in the governance of AI-enabled systems.

Despite these developments, governance implementation in production environments remains inconsistent (Salgado-Criado, 2025). Organizations frequently struggle to balance operational efficiency with compliance obligations, especially in systems requiring real-time decision-making and large-scale automation (Frenette, 2023). Consequently, oversight mechanisms may fail to accurately reflect the true autonomy level or risk profile of deployed systems (James, 2026).

2.4 Gaps in Existing Research

Although existing literature provides important insights into human-AI interaction and governance principles, several critical gaps remain. First, there is a lack of operational frameworks capable of translating theoretical oversight concepts into measurable criteria for evaluating real-world AI systems (Jauhiainen, 2025). Much of the current literature focuses on abstract governance principles without offering practical methodologies for assessing oversight effectiveness in production environments (Ottun & Flores, 2025).

Second, research has given limited attention to operational constraints inherent in production AI systems, including latency requirements, scalability challenges, and the economic cost of maintaining human oversight (Frenette, 2023). In real-world deployments, the need for rapid decision-making and the scale of AI-driven operations often limit the feasibility of continuous human supervision (Salgado-Criado, 2025). These practical considerations significantly affect oversight effectiveness but remain underexplored in academic research (James, 2026).

Third, there is no widely accepted framework for determining when human oversight is sufficient versus merely symbolic (Koulu, 2020). While scholars broadly agree that human involvement is necessary for responsible AI deployment, there is limited consensus regarding how oversight adequacy should be measured or benchmarked (Kandikatla & Radeljic, 2025). This lack of standardization creates challenges for organizations attempting to evaluate governance practices or align them with emerging best practices (Jauhiainen, 2025).

Recent studies also highlight growing tensions between increasing automation and maintaining meaningful human control. Holzinger et al. (2025) question whether effective oversight remains feasible as AI systems become increasingly autonomous and adaptive. Similarly, Narayanan and Feigh (2026) demonstrate that high levels of AI dependency and information abstraction can weaken human supervisory capabilities within collaborative decision-making teams. Sadovski et al. (2025) further identify a persistent "human-oversight dilemma," where humans are expected to supervise systems whose complexity may exceed practical human comprehension.

These limitations demonstrate the need for a structured, practitioner-oriented framework that integrates autonomy levels, operational risk, and production realities to evaluate the sufficiency of human oversight in AI systems (Frenette, 2023; Kandikatla & Radeljic, 2025).

3. Conceptual Foundations

3.1 Defining "Oversight Sufficiency"

Although many think that human oversight in AI systems is either present or absent, this perspective is inappropriate for evaluating effectiveness. When it comes to production environments, oversight must be considered as a multi-dimensional construct, not only when it is present but also how it is performed, when it is performed, and by whom it is exercised. Human oversight is sufficient when it can have a tangible impact on the outcomes of the system.

First, human oversight must be effective in identifying, capturing, and correcting undesirable system behavior. Second, the timeliness of intervention must indicate when it is possible to intervene prior to harm occurring, especially in systems that operate at high speed or scale. Third, the authority of the human operator must mean whether it is possible for the human operator to override, modify, or stop the system decisions.

To operationalize oversight sufficiency, this study defines several key dimensions:

- Intervention Capability: The extent to which humans can intervene in system operations. This includes the availability of override mechanisms, escalation protocols, and manual control options. Effective intervention requires both technical access and procedural clarity.
- Observability: The degree to which system processes and decisions are visible and understandable to human operators. This includes access to logs, explanations, and real-time monitoring tools that enable situational awareness.
- Control Authority: The level of decision-making power assigned to human actors. Oversight is insufficient if humans can observe system behavior but lack the authority to act upon it.
- Response Latency: The time required for humans to detect an issue and execute an intervention. In high-speed systems, even minor delays can render oversight ineffective, transforming human involvement into a purely symbolic function.

Together, these dimensions provide a basis for assessing whether oversight mechanisms work well in practice rather than simply exist in system design.

3.2 Dimensions of AI Autonomy

The autonomy of an AI system influences the need for oversight. In this regard, autonomy refers to the system's ability to make and execute decisions autonomously, without human input. Autonomy is defined in this study in four dimensions.

Decision Complexity reflects the sophistication of the decisions being made.

Simple classifications or rule-based outputs are low complexity, while reasoning in multiple steps, probabilistic inference, and context sensitive decisions are higher complexity. As complexity increases, it becomes harder for the human to understand and validate.

Learning Capability distinguishes between static systems and adaptive systems.

Static systems operate according to fixed rules or pre-trained models, while adaptive systems are continually learning from new information and changing their behavior as it evolves. Adaptive systems may change in unforeseeable ways.

Execution Independence refers to the degree to which a system can act on its decisions without human approval.

The execution independence of the system allows it to initiate, activate or affect outside environments entirely autonomously, providing more efficiencies and less risk.

Environmental Uncertainty captures the variability and unpredictability of the context in which the AI operates.

Systems operating in stable, well-defined contexts require less attention than systems operating in dynamic, open, or dangerous environments where unexpected outcomes are more likely.

In general, these dimensions together describe the autonomy profile of an AI system. Higher values for each of these dimensions typically indicate that the system needs to be supervised in a more aggressive and responsive manner.

3.3 Risk-Based Perspective

One of the main tenets of this study is that the effectiveness and cost of supervision should be analyzed on the basis of system risk. Not all AI systems require the same degree of human oversight; rather, the intensity and cost of supervision should correspond to the consequences of system failure or misuse. The stance of AI governance in general is to deal with issues related to risks by designing rules and controls for AI systems. (Figure 1)

System risk is the impact that the system may have on individuals, organizations, and society. Risks can include financial loss, safety hazards, ethical violations, and damages to reputation. As the risk increases, so must the need for stronger controls, better observation, and less response time.

Healthcare and finance are two of the most risky areas in terms of cost and potential for errors. Human intervention is needed in the decision-making process so that the decision is not taken without human intervention. In the financial sector, automated decisions about credit, trading, or fraud are economically and legally important and will require human intervention.

In contrast, lower risk applications like content recommendation or less critical processes automation may tolerate a greater degree of autonomy with less supervision. But, even in these cases, cumulative effect and scale can pose systemic risks to be considered.

By adopting a risk-based perspective, organizations can allocate oversight resources more effectively, ensuring that human involvement is both proportionate and impactful. This approach also provides a foundation for evaluating oversight sufficiency, as it links the required intensity of human control directly to the autonomy and risk profile of the AI system.

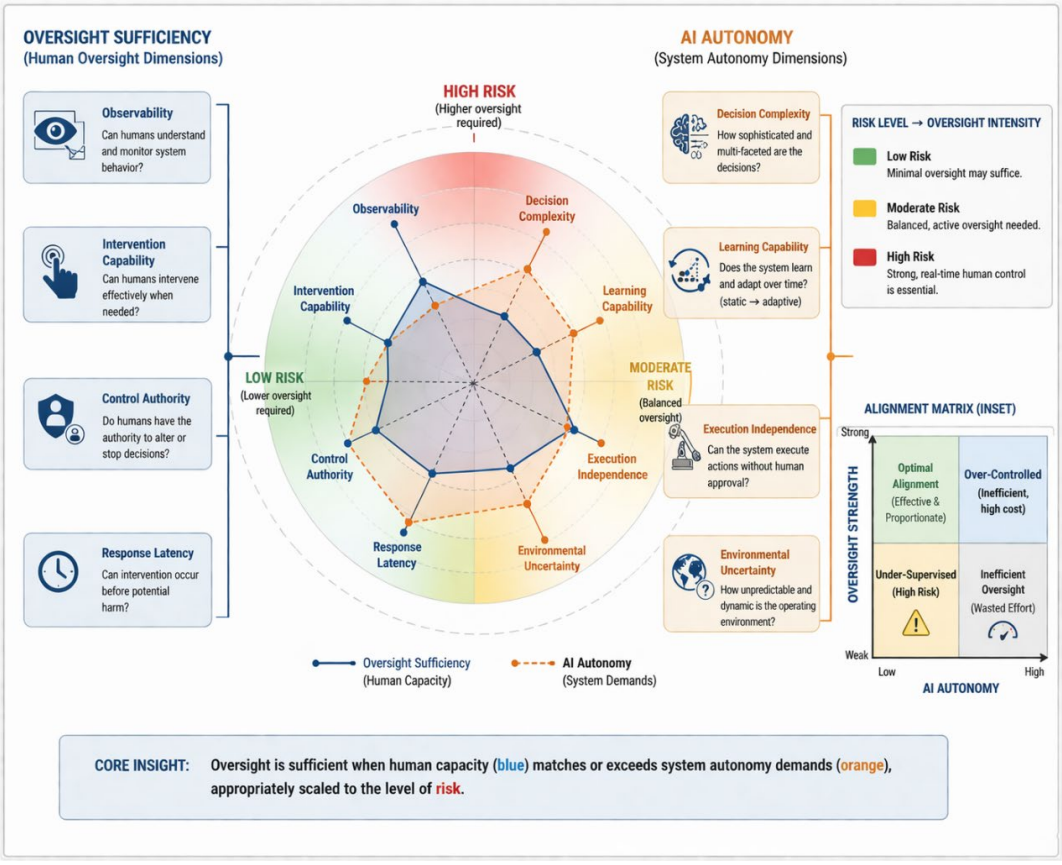

Figure 1. Risk-Based Human Oversight Sufficiency Framework for Autonomous AI Systems

4. Proposed Framework: Oversight Sufficiency Across Autonomy Levels

This section presents a structured framework for evaluating the sufficiency of human oversight in production AI systems. The framework integrates autonomy classification, oversight roles, and multi-dimensional evaluation criteria to provide a practical tool for aligning human control with system behavior and risk exposure.

4.1 Autonomy Level Classification (Example Model)

To systematically assess oversight requirements, AI systems are categorized into six levels of autonomy, reflecting increasing independence in decision-making and execution:

Level 0 – Manual Systems (No AI): All decisions and actions are performed by humans. No automation is involved, and oversight is inherent in human control.

Level 1 – Assisted Intelligence: AI provides recommendations or insights, but humans retain full decision-making authority. The system functions strictly as a support tool.

Level 2 – Partial Automation: AI systems can execute decisions, but only with explicit human approval. Human validation is required before action is taken.

Level 3 – Conditional Automation: AI systems operate independently under normal conditions but allow for human intervention through override mechanisms when necessary.

Level 4 – High Automation: AI systems perform most tasks autonomously, with humans primarily in monitoring roles. Intervention is possible but not routinely exercised.

Level 5 – Full Autonomy: AI systems operate without real-time human involvement. Oversight is limited to pre-deployment validation and post-hoc auditing.

This classification establishes a continuum that links increasing autonomy with decreasing direct human involvement, thereby necessitating more sophisticated and indirect forms of oversight.

4.2 Oversight Sufficiency Matrix

The framework maps autonomy levels to corresponding human roles, oversight types, and sufficiency criteria. Rather than prescribing a single oversight model, it emphasizes alignment between system autonomy and the nature of human involvement. (Table 1)

Table 1. Autonomy Levels and Corresponding Human Oversight Framework

Autonomy Level	Human Role	Oversight Type	Sufficiency Criteria
Low (Levels 0–1)	Decision-maker	HITL	Full visibility and direct control over decisions and execution
Medium (Levels 2–3)	Supervisor	HOTL	Continuous monitoring with reliable and timely override capability
High (Levels 4–5)	Auditor	Post-hoc	Comprehensive logging, explainability, and robust audit mechanisms

At lower levels of autonomy, sufficiency depends on maintaining strong human control and visibility. As autonomy increases, oversight shifts toward monitoring and auditing, requiring enhanced system transparency and traceability. Importantly, the framework recognizes that oversight mechanisms must evolve in sophistication as direct human control diminishes.

4.3 Core Oversight Dimensions (Framework Components)

To evaluate oversight sufficiency across different autonomy levels, the framework incorporates five core dimensions:

- Observability: The extent to which humans can access, interpret, and understand system behavior. This includes real-time dashboards, decision explanations, and system logs that support situational awareness.
- Intervention Capability: The ability of human operators to intervene effectively when necessary. This involves technical mechanisms (e.g., override controls) as well as organizational processes (e.g., escalation protocols).
- Timeliness: The speed at which human intervention can occur relative to system operations. Oversight is only effective if intervention can be executed before undesirable outcomes materialize.
- Accountability Mechanisms: The presence of structures that ensure decisions are traceable and attributable. This includes audit trails, version control, and documentation of system behavior and human actions.

• Adaptability Control: The ability to constrain or guide the learning behavior of adaptive systems. This includes mechanisms for model validation, retraining governance, and safeguards against unintended drift.

These dimensions provide a comprehensive lens for assessing whether oversight mechanisms are functionally capable of managing the risks associated with a given level of autonomy.

4.4 Sufficiency Classification Model

Building on the autonomy classification and oversight dimensions, the framework defines three categories of oversight sufficiency:

• Sufficient Oversight: Oversight mechanisms are well-aligned with the system's autonomy level and risk profile. Humans have adequate visibility, authority, and responsiveness to influence outcomes effectively. Intervention is feasible, timely, and impactful.

• Marginal Oversight: Oversight exists but is limited in effectiveness. This may involve partial visibility, delayed response times, or constrained intervention capabilities. While risks may be mitigated under normal conditions, the system is vulnerable under stress or unexpected scenarios.

• Insufficient Oversight: Human involvement is largely symbolic or ineffective, often described as "human-in-the-loop theater." In such cases, humans lack the necessary authority, information, or timeliness to meaningfully influence system behavior. This misalignment between autonomy and oversight introduces significant operational and ethical risks.

This classification model enables organizations to move beyond compliance-oriented thinking and critically evaluate the real-world effectiveness of their oversight mechanisms. By identifying gaps between autonomy and oversight, the framework supports targeted improvements in system design, governance, and operational practice. (Figure 2)

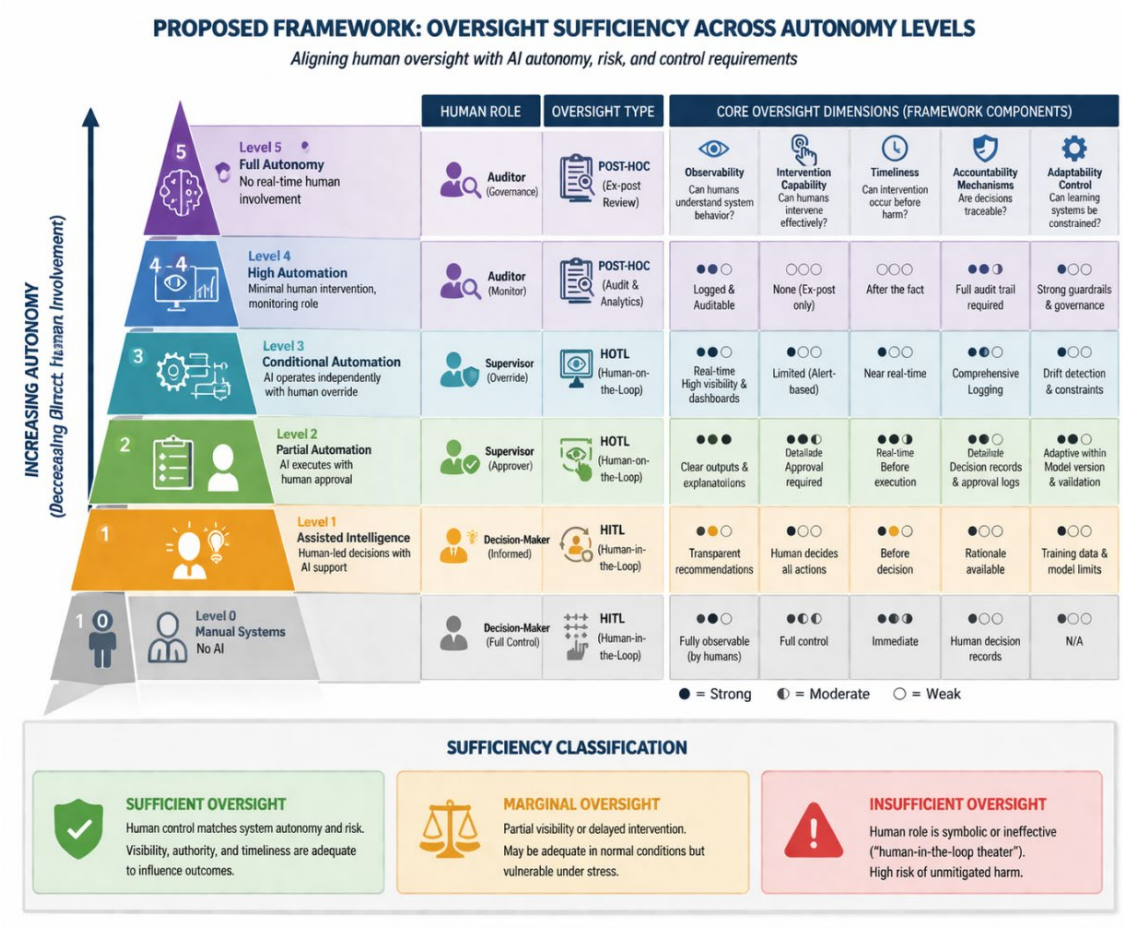

Figure 2. Proposed Framework: Oversight Sufficiency Across Autonomy Levels

5. Real-World Implementation Patterns

While theoretical models of human oversight provide important conceptual guidance, the realities of production AI systems reveal a more complex landscape shaped by domain-specific requirements, architectural constraints, and operational trade-offs. This section examines how organizations implement oversight in practice, highlighting recurring patterns across industries and the limitations that influence their effectiveness.

5.1 Industry Case Examples

Finance (Fraud Detection Systems): In financial institutions, AI-driven fraud detection systems operate at high speed and scale, analyzing transactions on the fly to catch suspicious activity. Oversight is typically a combination of a combination of automated flagging and human review. Low-risk transactions are processed automatically, while high-risk or complex transactions are sent to analysts. Human oversight here serves as a validation and escalation mechanism, thus keeping false positives to a minimum and critical decisions carefully scrutinized. Because of the size of the transaction, human review becomes selective rather than full oversight.

Healthcare (Clinical Decision Support): Clinical decision support systems use AI models to assist physicians in diagnosing, recommending treatment or predicting risk. In this case, the human in-the-loop decision-making process is still the final decision with the clinician determining the outcome; the recommendations from AI are advisory, and context and context interpretation are human decisions. But, relying more heavily on AI's recommendations may be an issue of automation bias, where clinicians over-rely on the recommendations of the AI system, and it would make oversight less effective.

Cybersecurity (AI Threat Detection and Response): AI systems in cybersecurity monitor network traffic, detect suspicious activity, and even respond to attacks by itself. Oversight is by Humans on the Loop. An analyst reviews and controls the actions of the AI system through dashboards and alerts. In extreme cases, humans may intervene to confirm the automated actions taken. But, because of the speed and complexity of cyber threats, the control loop often closes quickly. Analysis of the threat will occur after the event.

E-commerce (Recommendations and Pricing Engines): E-commerce platforms rely heavily on AI for personalization, recommendation, and pricing. Artificial intelligence is in high gear, learning from the user and the market. The human behind these systems is usually in the shadows, closely monitoring the performance metrics and operating checks, and making changes to the system. While AI is seen as not as risky as healthcare or finance, it is widely deployed, and even minor mistakes can have great impact when combined. (Table 2)

Table 2. Industry-Specific AI Oversight Frameworks and Monitoring Indicators

Industry Domain	Typical AI Application	Primary Oversight Model	Typical Monitoring Indicators
Finance	Fraud detection and transaction monitoring	HITL/HOTL hybrid	False positive rate, fraud detection accuracy, transaction escalation rate, override frequency, response latency
Healthcare	Clinical decision support systems	HITL	Diagnostic accuracy, clinician override rate, patient safety incidents, explainability score, intervention success rate
Cybersecurity	Threat detection and automated response	HOTL	Threat detection precision, alert response time, false alarm rate, escalation frequency, recovery time
E-commerce	Recommendation and dynamic pricing systems	Post-hoc oversight	Recommendation relevance, pricing anomaly rate, bias indicators, customer complaint frequency, auditability metrics

5.2 Common Oversight Architectures

Across industries, there are several common architectural patterns that allow the automation of human oversight:

- Approval Workflows: AI-generated decisions are subject to approval from the human before being executed. This is typically the case in moderate risk scenarios where accuracy is critical but real-time processing is not necessary.
- Alert-Based Supervision: Systems notify you when something is going to be amiss. When this happens, the system will set up a threshold or alert you with an anomaly. Then the person handling the system will look at the alert and determine if it is time to intervene. This approach ensures the appropriate mix of scalability and focused control.
- Escalation Systems: Multiple tiers of oversight refer complex cases or high-risk ones to more experienced human

operators, who ensure that decisions are given the attention they need while maintaining efficiency.

- Human Review Queues: AI outputs are batched and queued for human review, often based on the risk level. This model is used in content moderation and fraud detection, where there are large numbers of decisions that require a limited number of human checks.

These systems are a compromise between automation and human oversight, often combining multiple monitoring mechanisms into one device.

5.3 Practical Constraints

The reality, of course, is that these systems are available. But, they are imperfect.

Latency vs. Accuracy Trade-offs: In systems requiring immediate action, such as fraud detection or cybersecurity, human error can make the system ineffective. Organizations must weigh the value of human error against the need to quickly take action. Speed is often preferred over control.

Cost of Human Intervention: Maintaining quality human oversight at scale can be expensive. Highly skilled human operators need training, compensation, and support to maintain quality oversight at scale. This makes maintaining high levels of oversight in large systems economically untenable.

Scalability Challenges: As AI processes increasing amounts of data and decisions, human oversight may not keep up. This can mean more selective or sampled oversight rather than complete coverage, leaving blind spots.

Human Fatigue and Cognitive Overload: Continuous monitoring of AI systems can lead to fatigue, reduced attention, and decision errors by humans. High alert volumes, especially in alert-based systems, can lead to desensitization or missed critical events, thus undermining the purpose of oversight.

Overall, the experience of implementation showed that human oversight is not static but dynamic. Companies have to choose between efficiency, cost, and control while providing imperfect oversight, which implies that the oversight they provide has to be context-specific. Hence, a method such as the one proposed in this study is necessary in order to verify if human oversight is sufficient given the autonomy and risk of production AI systems.

6. Evaluation Methodology

In this section, we describe a method for applying the framework to real-world situations to assess whether human oversight is adequate to monitor the autonomy and risks of AI systems. We use a stepwise process to evaluate and test the system's efficiency and viability using quantitative measures.

6.1 Framework Application Process

In general, the evaluation process involves four step-by-step activities that help practitioners navigate the assessment process from the assessment of the system to the classification of the oversight.

Step 1: Classify the Autonomy Level The first step in categorizing the AI is to identify the autonomy levels in the framework (Level 0–5). In this step, one needs to analyze the AI's decision-making authority, execution independence, and learning capabilities. An accurate classification is crucial as it defines the expectations for oversight.

Step 2: Assess System Risk Following the assessment of the system's risk profile, it is evaluated for potential impact, domain sensitivity and potential consequences of failure. For example, risks may include financial loss, safety risks, regulatory compliance risk, or reputational damage. Risk assessment provides the context for choosing the extent of oversight.

Step 3: Analyze Oversight Dimensions The system is evaluated in the four oversight dimensions – observability, intervention capability, timeliness, accountability and control over adaptability. For each of these dimensions, qualitative and quantitative measures should be used as appropriate. In this step, the focus should be on the functional effectiveness of oversight rather than its mere existence.

Step 4: Classify Sufficiency Level Based on the degree of autonomy, risk, and performance of oversight, the system is categorized as sufficient, marginal, or insufficient oversight. Human involvement is considered to have some influence over outcomes in normal and abnormal circumstances.

6.2 Metrics for Assessment

To help objective evaluation, the framework also incorporates a set of measurable indicators that capture key aspects of oversight performance, including:

- Intervention Success Rate: The percentage of human interventions that successfully prevent or reduce undesirable outcomes. This measures the practicability of oversight measures.
- Latency of Detection: The time between when an issue arises and when human operators or monitoring systems

identify it. Shorter latency indicates faster oversight.

- Error Recovery Rate: A measure of the system's ability to recover from errors by human intervention or corrective processes—not just as a one-time fix but also as a long-term fix.
- Human Override Frequency: The rate at which humans override or modify AI decisions. An override that is sporadic or very low may indicate an over-reliance on automation or an ineffective level of monitoring.
- Scores of explainability: Quantitative or qualitative measures of the degree to which decision-making about the system is observable by the user.

These metrics will allow organizations to go beyond subjective assessments to determine whether oversight is sufficient.

6.3 Validation Approach

As proof that the model is robust and generalizable, it should be validated using multiple approaches representing real-world conditions.

Simulation of Real-World Scenarios: Organizations can simulate operational environments to test how oversight mechanisms perform under typical conditions. This includes replicating decision flows, user interactions, and system responses to assess whether human oversight functions as intended.

Stress-Testing Under Edge Cases: Edge cases and rare events often expose weaknesses in oversight mechanisms. Stress-testing involves deliberately introducing anomalies, extreme conditions, or adversarial inputs to evaluate how effectively humans can detect and respond to unexpected situations.

Cross-Domain Benchmarking: Comparing oversight performance across different domains and system types helps identify best practices and common shortcomings. Benchmarking enables organizations to contextualize their performance and adopt proven strategies from other industries.

Overall, this method of evaluation offers an effective and comprehensive means to examine the role of humans in the oversight of AI systems in production. With structured processes, measurable metrics, and validating tools, this method can be used to identify gaps, improve the oversight approach, and ensure that the role of humans is meaningful as the autonomy of the system grows. (Figure 3)

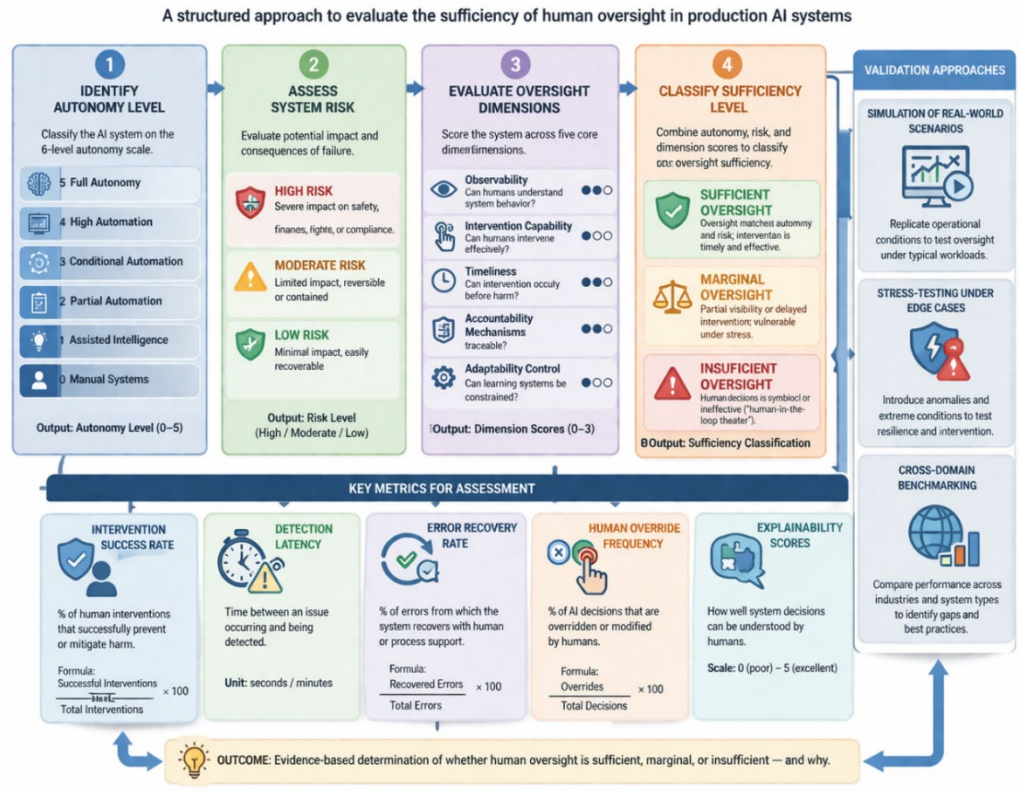

Figure 3. Evaluation Methodology for Oversight Sufficiency in Production AI Systems

6.4 Oversight Score Calculation Model

To quantitatively evaluate the sufficiency of human oversight in production AI systems, this study proposes a weighted scoring model that aggregates performance across the core oversight dimensions. The model enables organizations to systematically compare oversight effectiveness across systems with different autonomy levels and operational risks.

The overall oversight sufficiency score is calculated as:

$$\text{Oversight Score} = \sum(\text{Dimension Score} \times \text{Weight})$$

Where:

- Dimension Score represents the evaluation score assigned to each oversight dimension.
- Weight represents the relative importance of the dimension based on system autonomy and risk profile.

Core Oversight Dimensions

Table 3. The framework evaluates five primary oversight dimensions:

Dimension	Description
Observability (O)	Ability of humans to understand and monitor system behavior
Intervention Capability (I)	Ability of humans to intervene or override AI decisions
Timeliness (T)	Speed at which intervention can occur before harm materializes
Accountability Mechanisms (A)	Traceability and auditability of AI decisions
Adaptability Control (C)	Ability to constrain or govern adaptive system behavior

7. Discussion

The results of this study show that the design of human oversight for production AI systems is not an easy task; it takes competing goals, internal or external resources, and the ability of the system to keep pace with emerging technologies. This section examines the tensions and risks associated with the implementation of human oversight.

A trade-off exists between control and efficiency. While organizations aim for greater automation to increase performance, scalability, and cost reduction, they lose the ability to make an immediate decision. Human intervention reduces the chance of human error. In high-volume systems like financial transaction processing or cybersecurity monitoring, introducing humans into the process can cause a bottleneck and lower the overall performance of the system. Organizations therefore find themselves forced to focus on selective oversight, where humans focus on cases that are particularly high-risk or unusual. When this happens, critical issues can be missed if the problem does not fall within certain thresholds.

Both are related to the problem of over-reliance on automation. As AI systems are known to be highly accurate and consistent, it is possible for human operators to give too much trust in automated outputs (called automation bias). This can reduce human surveillance, since humans are less likely to question or challenge system decisions, even when a problem arises. Over time, this will diminish human expertise as users become passive observers instead of decision-makers. In such a scenario, having human oversight does not necessarily translate into control since cognitive and behavioral characteristics of human-AI interaction can degrade oversight quality.

Another problem is the appearance of oversight, sometimes referred to as “human-in-the-loop theater.” Human involvement in many production processes may be implied, but not effective. For example, humans may be required to approve decisions without the time, information, or authority to evaluate those decisions. At other times, oversight may only occur after the decision has been made, with little ability to prevent harm. So, oversight is more of an artifact of compliance than a safeguard, creating a false sense of security for the organization and people.

Plus, there is a lack of accountability and ethics when AI systems are operated autonomously and with little oversight. It is difficult to determine who is responsible for the outcomes of AI. Is it the designer, the operator, the organization, or the AI itself? This is especially problematic in high-stakes situations where decisions can have a significant impact on individuals and society. Lack of oversight leads to bias, unfair treatment, and unintended harm even when the system behavior is not transparent.

That discussion clearly illustrates that effective human oversight is not obtained through a superficial blend of roles. Rather, it requires alignment between autonomy, risk, and human performance. Organizations need to think about oversight as an active part of their system design, continuously evaluated and adapted to changing circumstances.

Without such an attitude, the benefits of automation delivered by artificial intelligence may come with increasing risks that are difficult to identify, manage, or assign responsibility for.

8. Conclusion

AI deployment in production situations offers unparalleled benefits for scalability, performance, and decision-making driven by data, but also new challenges for accountability, risk, and human oversight. As this study has shown, human involvement in AI systems should not be viewed as a check box for compliance but a multi-faceted and context dependent construct that directly relates to autonomy and risk.

This paper proposes a model of oversight sufficiency to provide clarity and direction to organizations working on ensuring a quality level of human control. The model includes autonomy, oversight aspects, and risk factors to help organizations evaluate whether humans' roles as decision makers, supervisors, or auditors are effective. Using the oversight sufficiency matrix and evaluation procedure, the organizations can define their AI systems' oversight levels as sufficient, marginal, or insufficient to identify gaps and opportunities for improvement.

The analysis of real-world implementation patterns highlights recurring architectures, such as approval workflows, alert-based supervision, and escalation systems, as well as the operational constraints that shape their effectiveness, including latency, scalability, and cognitive limitations. These insights demonstrate that the presence of human oversight alone does not guarantee safe or responsible AI operation. Rather, oversight must be carefully calibrated to system autonomy, risk, and domain-specific requirements.

A final conclusion highlights some of the trade-offs and risks involved in automation, token human involvement, and lack of accountability. Organizations must look at their oversight as an active and risk-based process, taking into account a technology-based, process-based, and human-centered approach.

The idea of regulating an AI system that is increasingly autonomous and inseparable from a decision-making process can bring about more ethical conduct and social trust. The proposed framework could help bridge the gap between research and production for organizations to design, analyze, and continuously improve oversight.

References

Endsley, M. R. (2017). From here to autonomy: Lessons learned from human–automation research. *Human Factors*, *59*(1), 5–27. https://doi.org/10.1177/0018720816681350

Frenette, J. (2023). Ensuring human oversight in high-performance AI systems: A framework for control and accountability. *World Journal of Advanced Research and Reviews*, *20*(2), 1507–1516. https://doi.org/10.30574/wjarr.2023.20.2.2194

Herrmann, T. (2025). Intervenability as a design requirement for autonomy and oversight within human-centered AI. In (Ed.), *The design of human-centered artificial intelligence for the workplace* (pp. 143–166). Springer Nature Switzerland. https://doi.org/10.1007/978-3-031-83512-4_9

Holzinger, A., Zatloukal, K., & Müller, H. (2025). Is human oversight to AI systems still possible? *New Biotechnology*, *85*, 59–62. https://doi.org/10.1016/j.nbt.2024.12.003

James, M. (2026). *Balancing automation and accountability: Human oversight in AI-based planning and forecasting systems*. [Unpublished manuscript].

Jauhiainen, A. (2025). *Effective human oversight in artificial intelligence*. [Unpublished manuscript].

Kandikatla, L., & Radeljic, B. (2025). AI and human oversight: A risk-based framework for alignment. arXiv preprint arXiv:2510.09090. Retrieved from https://arxiv.org/abs/2510.09090

Koulu, R. (2020). Proceduralizing control and discretion: Human oversight in artificial intelligence policy. *Maastricht Journal of European and Comparative Law*, *27*(6), 720–735. https://doi.org/10.1177/1023263X20978649

Laitinen, A., & Sahlgren, O. (2021). AI systems and respect for human autonomy. *Frontiers in Artificial Intelligence*, *4*, 705164. https://doi.org/10.3389/frai.2021.705164

Methnani, L., Aler Tubella, A., Dignum, V., & Theodorou, A. (2021). Let me take over: Variable autonomy for meaningful human control. *Frontiers in Artificial Intelligence*, *4*, 737072. https://doi.org/10.3389/frai.2021.737072

Narayanan, R., & Feigh, K. M. (2026). Designing for oversight: An empirical investigation of the dual impact of AI dependency and information abstraction on human supervision in decision-making teams. *International Journal of Human–Computer Interaction*, 1–30. https://doi.org/10.1080/10447318.2026.2618568

O'Sullivan, S., Nevejans, N., Allen, C., Blyth, A., Leonard, S., Pagallo, U., … Ashrafian, H. (2019). Legal, regulatory, and ethical frameworks for development of standards in artificial intelligence (AI) and autonomous robotic surgery. *The International Journal of Medical Robotics and Computer Assisted Surgery*, *15*(1), e1968. https://doi.org/10.1002/rcs.1968

Ottun, A. R. O., & Flores, H. (2025). Trustworthy AI in practice: A comprehensive review of human oversight and human-in-the-loop approaches. Authorea Preprints. https://doi.org/10.36227/techrxiv.176118749.93102582/v1

Sadovski, E., Aviv, I., & Hadar, I. (2025, September). Navigating the human-oversight dilemma in AI-based systems. In *2025 IEEE 33rd International Requirements Engineering Conference Workshops (REW)* (pp. 454–461). IEEE. https://doi.org/10.1109/REW66121.2025.00069

Salgado-Criado, J. (2025). Human oversight of artificial intelligence: An operations management perspective. *Journal of Industrial Engineering and Management*, *18*(2), 285–304. https://doi.org/10.3926/jiem.8567

Tsamados, A., Floridi, L., & Taddeo, M. (2025). Human control of AI systems: From supervision to teaming. *AI and Ethics*, *5*(2), 1535–1548. https://doi.org/10.1007/s43681-024-00489-4

Verdiesen, I., Santoni de Sio, F., & Dignum, V. (2021). Accountability and control over autonomous weapon systems: A framework for comprehensive human oversight. *Minds and Machines*, *31*(1), 137–163. https://doi.org/10.1007/s11023-020-09532-9