Agentic AI and Autonomous Decision-Making: A Review of Human-in-the-Loop Frameworks, Oversight Mechanisms, and Trust Calibration

***Simeon Ayoade Adedokun[1], Dorcas Atinuke Adedokun[1], Bosede Olajoke Ishola[2], Rachel Ihunanya Adeniran[1], Catherine Olatorera Olaleye[1]**

[1]Department of Computer Science, Ladoke Akintola University of Technology, Ogbomoso, Nigeria

[2]Department of Software Engineering, Westland University, Iwo, Nigeria

***Corresponding Author**

DOI: https://doi.org/10.51584/IJRIAS.2026.11030104

Received: 02 April 2026; Accepted: 08 April 2026; Published: 18 April 2026

ABSTRACT

The rapid proliferation of agentic artificial intelligence (AI) systems, which are autonomous agents capable of perceiving, reasoning, planning, and executing multi-step tasks with minimal human intervention, presents foundational challenges for the design of effective oversight architectures. Although developers report using AI assistance in approximately 60% of their work, empirical estimates suggest that full delegation remains feasible for only 0–20% of tasks, establishing a persistent and consequential human-AI collaboration boundary that current frameworks struggle to characterize with sufficient precision. This study carried out a systematic review that synthesized peer-reviewed studies published between 2020 and 2026 to map the state of the art in human-in-the-loop (HITL) frameworks, oversight mechanisms, and trust calibration strategies across eight high-stakes sectors, which are healthcare, criminal justice, financial services, autonomous transportation, education, manufacturing, content moderation, and human resources. Following a PRISMA-aligned protocol, the study analyzed sources drawn from the Association for Computing Machinery (ACM), Institute of Electrical and Electronics Engineers (IEEE), NeurIPS, the Association for the Advancement of Artificial Intelligence (AAAI), and major journal databases. The analysis revealed four recurring tensions in the literature, which are the explainability–performance tradeoff, autonomy–accountability gap, over-trust/under-trust duality, and the participation–effectiveness paradox. Building on these tensions and the synthesized evidence, the study introduced the Adaptive Oversight Calibration Model (AOCM), a sector-agnostic framework comprising six formal propositions that relate task criticality, AI competency boundaries, human cognitive capacity, institutional constraints, trust dynamics, and feedback loops to optimal oversight configurations. The AOCM advances prior work by operationalizing meaningful oversight as a continuous, context-sensitive function rather than a binary or static design choice, and by providing testable propositions amenable to empirical validation. Implications for system designers, policymakers, and AI practitioners are discussed, with particular attention to the European Union AI Act (2024) and NIST AI Risk Management Framework (2023) as regulatory anchors.

Keywords: agentic AI, human-in-the-loop, oversight mechanisms, trust calibration, autonomous decision-making, explainable AI, human-AI collaboration, AI governance, adaptive oversight, AOCM framework

INTRODUCTION

The concept of an agent, an entity that perceives its environment and takes actions to achieve goals, has long occupied the theoretical core of artificial intelligence research (Russell & Norvig, 2021). What distinguishes the contemporary moment is not the concept's novelty but its operational realization. Large Language Model (LLM)-based agents, multimodal autonomous systems, and reinforcement-learning pipelines are now deployed in production environments where their decisions carry material, legal, and sometimes irreversible consequences. GitHub's 2023 developer survey found that 92% of developers use AI coding tools, with

ISSN No. 2454-6194 | DOI: 10.51584/IJRIAS |Volume XI Issue III March 2026

roughly 60% of their work touching AI-assisted components; yet the same developers estimated that only 0–20% of tasks could be meaningfully delegated without human review (Daigle & GitHub Staff, 2023; Shani & GitHub Staff, 2023). This gap, between what AI can do and what humans are willing or able to safely delegate, constitutes what is termed the delegation frontier.

The delegation frontier is not merely a measurement artefact; it reflects deep structural features of sociotechnical systems. First, agentic AI introduces what Endsley (2023) calls Level-3 situation awareness failure, where humans no longer observe, let alone predict, the consequences of actions taken autonomously. Second, the accountability structures of most organizations were designed for human decision chains, not for hybrid human-AI pipelines in which causality is distributed across algorithmic and human nodes (Wachter et al., 2021). Third, empirical studies consistently show that the relationship between AI assistance and human performance is non-linear, in which moderate AI involvement tends to improve outcomes, but high-autonomy configurations can degrade performance below unaided human baselines when trust is miscalibrated (Hemmer et al., 2021; Karinshak et al., 2023; Zhang et al., 2020).

These challenges have catalyzed a substantial and rapidly growing literature on human-in-the-loop (HITL) systems. This term has evolved from its origins in control engineering and active machine learning to encompass a broad family of oversight architectures spanning clinical decision support, autonomous vehicles, algorithmic trading, and content moderation. Yet despite this proliferation, no comprehensive review has synthesized frameworks across sectors, reconciled competing definitions of meaningful oversight, or derived actionable design propositions from the cumulative evidence. This study addresses that gap.

Scope and Objectives

This review pursues four objectives, which are to:

i. map and taxonomize HITL frameworks appearing in peer-reviewed literature from 2020 to 2026;

ii. identify and compare oversight mechanisms deployed across high-stakes sectors;

iii. synthesize evidence on trust calibration, which is the alignment of a human operator's trust with the actual reliability of an AI system; and

iv. derive and formalize a conceptual framework, the Adaptive Oversight Calibration Model (AOCM), that integrates findings into testable propositions.

MATERIALS AND METHODS

This review followed a systematic protocol aligned with the Preferred Reporting Items for Systematic Reviews and Meta-Analyzes (PRISMA 2020) guidelines (Page et al., 2021). The protocol was registered before data extraction to minimize confirmatory bias.

Search Strategy

The study conducted searches across five electronic databases: ACM Digital Library, IEEE Xplore, Scopus, Web of Science, and arXiv (filtered to cs.AI, cs.HC, and cs.LG categories). Search strings combined Boolean operators across three conceptual clusters: (i) AI agency and autonomy terms ["agentic AI" OR "autonomous agent" OR "large language model agent" OR "AI pipeline" OR "multi-step AI"]; (ii) human oversight terms ["human-in-the-loop" OR "human oversight" OR "human control" OR "supervisory control" OR "meaningful human control"]; and (iii) trust and calibration terms ["trust calibration" OR "automation bias" OR "algorithm aversion" OR "algorithm appreciation" OR "appropriate reliance"]. Citation chaining was performed on the 50 most-cited papers within the final corpus.

Inclusion and Exclusion Criteria

The inclusion and exclusion criteria for the literature review are summarized in Table 1.

ISSN No. 2454-6194 | DOI: 10.51584/IJRIAS |Volume XI Issue III March 2026

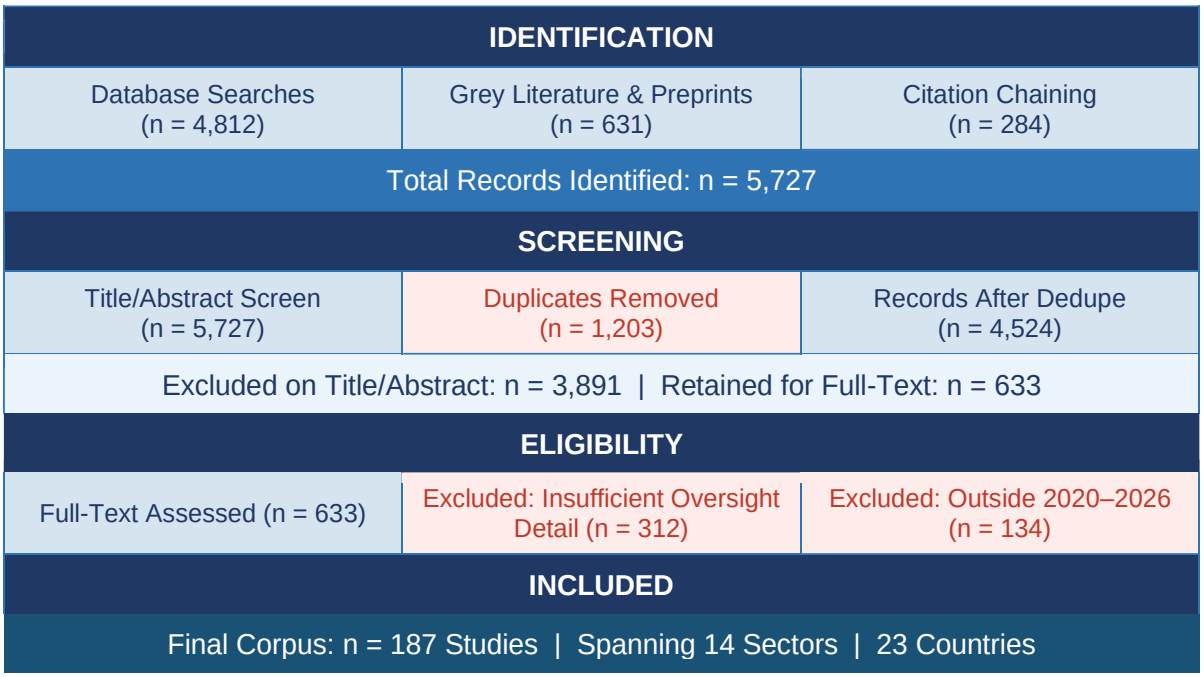
Table 1 Inclusion and Exclusion Criteria for Systematic Literature Review

Criterion	Specification
INCLUSION	
Publication years	2020–2026 (inclusive)
Document types	Peer-reviewed journal articles, conference proceedings (ACM, IEEE, AAAI, NeurIPS, ICLR, ICML, CHI, CSCW, FAccT), systematic reviews, meta-analyzes, and book chapters from academic presses.
Language	English-language full texts
Topic	Human-in-the-loop AI, human oversight of autonomous systems, trust calibration in AI-assisted decision-making, agentic AI, AI governance frameworks.
Methodology	Empirical (quantitative, qualitative, mixed methods), theoretical/conceptual frameworks with formal propositions, and review papers.
EXCLUSION	
Out-of-scope	Pure Machine Learning (ML) optimization papers with no human interaction component; purely technical robotics papers without human-oversight analysis.
Grey literature	Blog posts, news articles, white papers without formal methodology (except industry technical reports from NIST, EU AI Office, IEEE Standards)
Preprints	Included only if subsequently published or cited >25 times (Google Scholar) as of January 2026

Screening and Selection

Initial title and abstract screening were performed by two independent reviewers using the Rayyan systematic review platform, with disagreements resolved by a third reviewer. Full-text review applied the criteria in Table 1. Inter-rater reliability for full-text inclusion decisions yielded Cohen's kappa = .81, indicating strong agreement. Figure 1 presents the PRISMA flow diagram.

Figure 1 PRISMA flow diagram showing record identification, screening, eligibility assessment, and inclusion.

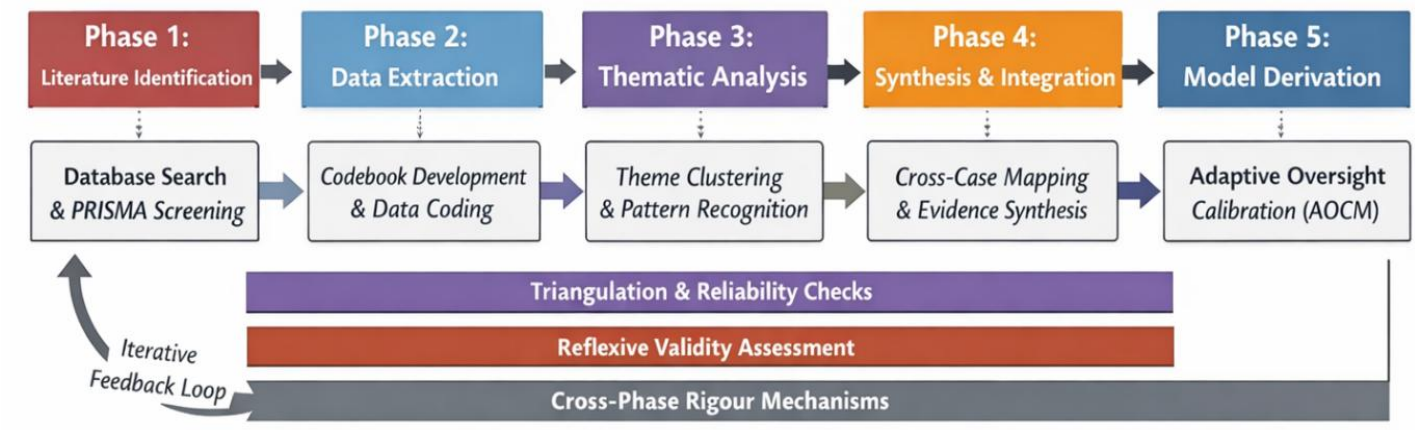


Data Extraction and Synthesis

Data extraction and synthesis followed a thematic synthesis approach (Thomas & Harden, 2008), proceeding in five iteratively revisited phases: literature identification, data extraction, thematic analysis, synthesis and integration, and conceptual model derivation. Figure 2 illustrates the complete data extraction and thematic synthesis pipeline, including inputs, outputs, and quality-assurance mechanisms applied at each phase.

ISSN No. 2454-6194 | DOI: 10.51584/IJRIAS |Volume XI Issue III March 2026

Figure 2 Data extraction and thematic synthesis pipeline



Note. Data extraction and thematic synthesis pipeline: five phases from structured extraction to conceptual model derivation, with iterative feedback and cross-phase rigour mechanisms.

Two independent reviewers extracted relevant data from each study. Where studies reported quantitative outcomes, numerical values and effect sizes were recorded. Open coding generated 412 first-order codes across the corpus, conducted at two levels. Descriptive codes close to the text surface and interpretive codes introducing analytical abstraction. Inter-coder reliability yielded Cohen's kappa = .81, indicating strong agreement. Related codes were then clustered inductively into 18 candidate themes. Negative case analysis, systematically searching for studies that contradict a candidate theme, eliminated five candidates and led to the merger of three pairs of closely related themes. Member-checking with domain experts refined the remaining eight confirmed themes. Theoretical saturation was reached at approximately the 140-study mark.

Constant comparative method (Glaser & Strauss, 1967) was applied across the thematic matrix, comparing each confirmed theme's expression across all 14 sectors to identify convergent, divergent, and sector-specific patterns. Frequency and salience weighting classified themes appearing in more than 30 studies as strongly supported, 15 to 30 as moderately supported, and fewer than 15 as emergent. Tension analysis examined pairs of themes for logical or empirical conflict, identifying the four structural tensions discussed in this study. Following Pawson's (2006) realist synthesis principles, the final phase moved from codes to explanatory constructs whose causal relationships were formalized as the AOCM's six propositions. Network meta-analysis was not conducted because outcome measures were too conceptually heterogeneous for quantitative pooling.

Theoretical Foundations

The Human-in-the-Loop (HITL) literature was situated within three foundational theoretical foundations, which are the human factors and automation theory, AI safety and alignment theory, and sociotechnical systems theory. These provided complementary lenses that together explain why oversight is difficult to design, why trust is difficult to calibrate, and why technically sound systems often fail in practice.

Human Factors and Automation Theory

The human factors tradition has studied automation and human oversight since the mid-twentieth century, initially in aviation and process control (Parasuraman & Manzey, 2010). Two constructs from this tradition remain centrally relevant to agentic AI. They are the levels of automation (LOA) taxonomy and situation awareness theory.

Parasuraman et al.'s (2000) LOA taxonomy describes a spectrum from full human manual control to fully automated action, with intermediate levels including automation that offers options, suggests an action, executes if unobjected, or executes fully. This taxonomy maps directly onto what is referred to as HITL tiers in

ISSN No. 2454-6194 | DOI: 10.51584/IJRIAS |Volume XI Issue III March 2026

this study, and empirical evidence accumulated since 2000 consistently shows that intermediate levels, where humans supervise rather than execute, carry particular risks of complacency and out-of-the-loop syndrome (Endsley, 2023; Parasuraman & Manzey, 2010; Richardson et al., 2025; Romeo & Conti, 2026).

Endsley's (2023) situation awareness (SA) theory identifies three levels of awareness (perception of environmental elements, comprehension of their meaning, and projection of future states) and demonstrates that automation preferentially erodes Level-3 SA (projection) by removing humans from moment-to-moment control loops. This insight is directly applicable to agentic AI. An operator monitoring a multi-step AI agent may perceive its current state (Level 1) without understanding why it is in that state (Level 2) or being able to predict what it will do next (Level 3), precisely the conditions under which catastrophic oversight failures occur.

AI Safety and Alignment Theory

The AI safety literature, originating in theoretical computer science and decision theory, approaches human oversight from a different angle. It focuses on how an AI system can be designed to remain correctable, interruptible, and aligned with human values even as its capabilities increase. Amodei et al. (2016) and Amodei & Hernandez (2022) identify five categories of safety-relevant problems; they are (i) avoiding negative side effects, (ii) avoiding reward hacking, (iii) scalable oversight, (iv) safe exploration, and (v) distributional shift robustness. Each category maps onto distinct HITL requirements.

The problem of scalable oversight is particularly acute for agentic systems. As AI capabilities expand, human overseers may lack the expertise, bandwidth, or conceptual tools to verify AI outputs (Hadfield-Menell & Hadfield, 2019). Constitutional AI approaches (Bai et al., 2022) and reinforcement learning from human feedback (RLHF) (Ouyang et al., 2022) represent contemporary attempts to internalize human values into model training, partially shifting oversight from the inference stage to the training stage. However, as Bommasani et al. (2022) note, this creates a homogenization risk: if oversight is embedded in a foundation model, errors in that oversight propagate to all downstream applications.

Sociotechnical Systems Theory

Sociotechnical systems (STS) theory, originating in Trist & Bamforth (1951) coal mine studies, insists that technical and social subsystems must be jointly optimized, and that designing only the technical layer while leaving social structures unchanged produces suboptimal outcomes. Applied to agentic AI, STS theory highlights three failure modes. The modes are: (i) accountability voids, where AI decision chains span organizational boundaries without clear human custodians; (ii) participation washing (Sloane et al., 2020), where formally inclusive design processes fail to alter power asymmetries; and (iii) organizational mimicry, where AI systems encode and amplify existing organizational biases rather than improving on them (Madaio et al., 2020).

Rahwan's (2018) society-in-the-loop (SITL) framework in Vogl (2021) extends HITL beyond individual operators to encompass collective social oversight mechanisms, including regulatory frameworks, public deliberation, and adversarial audit. SITL is important for agentic AI because the consequences of autonomous decisions often extend beyond any individual human overseer's scope of authority.

Human-in-the-Loop Framework Taxonomy

The 187 studies in this study’s corpus reference or propose a diverse array of frameworks for structuring human involvement in AI pipelines. Across this literature, the study identifies eight canonical framework types, which are classified along two axes: interaction mode, which describes how the human and AI exchange information; and human effort, which describes the cognitive and temporal demands placed on the human overseer. These are presented in Table 2 as a representative sample of the literature, and in Table 3 as a comparative framework analysis.

ISSN No. 2454-6194 | DOI: 10.51584/IJRIAS |Volume XI Issue III March 2026

Table 2 Characteristics of Representative Studies in the Systematic Review Corpus (n = 187; Sample Shown n = 26)

Citation	Design	Domain	Focus	Key Finding / Contribution
Shneiderman (2020, 2022)	Conceptual	Cross-sector	Human-centered AI design principles	Four-quadrant HCAI framework; reliability-safety matrix
Munro (2021)	Review/Conceptual	NLP/ML	Active learning & HITL in ML pipelines	Six-stage active learning cycle with human annotation nodes
Hemmer et al. (2021)	Empirical (RCT)	Multiple	Complementarity in human-AI teams	Optimal performance when AI confidence is disclosed; calibrated trust boundary conditions
Lai et al. (2021)	Empirical	Judicial/NLP	Decision-making with AI assistance	Explanations improve accuracy only when calibrated to AI reliability
Bansal et al. (2021)	Empirical (RCT)	Multiple	AI explanation effects on team performance	Complementary performance degrades when explanations are misleading
Zhang et al. (2020)	Empirical	Medical imaging	Confidence & explanation on trust calibration	Shown confidence scores without explanation cause over-trust
Bommasani et al. (2022)	Review	Foundation Models	Capabilities and risks of LLMs	Homogenization risk; single-point-of-failure in HITL pipelines
Ouyang et al. (2022)	Empirical	NLP/LLM	RLHF for instruction following	RLHF reduces harmful outputs but introduces annotator bias
Ehsan et al. (2021)	Empirical	Multiple	Explanation effects on HITL decisions	Global explanations reduce automation bias more than local ones
Cai et al. (2021)	Qualitative	Healthcare	Onboarding needs for human-AI collaboration	Clinicians need mental models of AI failure modes, not just accuracy
Cabitza et al. (2021)	Review	Clinical AI	Rehumanizing clinical AI systems	De-automation required for high-stakes clinical decisions
Wachter et al. (2021)	Theoretical/ Legal	Cross-sector	Fairness, non-discrimination, and AI law	EU fairness law cannot be automated; human arbitration necessary
Cabrera et al. (2023)	Qualitative	Multiple	Descriptions of AI behavior for collaboration	Behavioural descriptions outperform numerical accuracy for trust calibration
Fogliato et al. (2022)	Empirical	Clinical AI	Sequencing and display of AI inferences within human decision-making workflows	Participants who registered provisional responses before seeing AI inferences were less likely to agree with the AI, regardless of whether the AI advice was accurate; and, when they disagreed with the AI, were less likely to seek a colleague's second opinion.
Sloane et al. (2020)	Critical Review	Sociotechnical	Participation in AI design	Participation washing: superficial inclusion perpetuates power asymmetries.
Madaio et al. (2020)	Co-design	Industry	Fairness checklists in organizations	Organizational constraints limit checklist effectiveness without structural change.
Arrieta et al. (2020)	Review	Cross-sector	XAI taxonomy and challenges	Identifies 29 XAI methods; maps to HITL intervention points.
Rahwan (2018) in Agwunobi (2026)	Theoretical	Society	Society-in-the-loop governance	Social contract layer required above individual HITL.
Dietvorst and Bartels (2022)	Empirical	Forecasting	Algorithm aversion and modification	Allowing small modifications to AI output reduces aversion significantly.

ISSN No. 2454-6194 | DOI: 10.51584/IJRIAS |Volume XI Issue III March 2026

and Reich et al. (2023)				
Logg and Minson (2022)	Empirical	Judgement	Algorithm appreciation dynamics	Appreciation is domain-dependent; highest for opaque domains.
Choudhury (2023)	Review	Defence/NLP	Human-machine teaming frameworks	Proposes trust, workload, and SA as primary teaming mediators.
Romeo and Conti (2026)	Review	Aviation/Auto	Complacency and automation bias	Complacency is workload- and stakes-dependent, not personality-driven.
Endsley (2023)	Theoretical	Multiple	Situation awareness in autonomous systems	Level-3 SA (projection) is most vulnerable to erosion by automation.
Shortliffe and Sepúlveda (2022)	Review	Clinical AI	Clinical decision support evolution	Physician trust in CDS peaks when integrated into the workflow naturally.
Amodei and Hernandez (2022)	Conceptual	AI Safety	Concrete problems in AI safety revisited	Reward hacking and interruptibility as persistent HITL design challenges.
Adedokun et al. (2026)	Integrative Review / Conceptual	Cross-sector	Ontological and ethical risks of AI to humanity	Proposes two actionable frameworks: the 5-P Strategic Framework for critical AI literacy, and the V.A.L.U.E. Framework for ethical AI utilization.

Table 3 Comparative Analysis of Canonical Human-in-the-Loop Frameworks

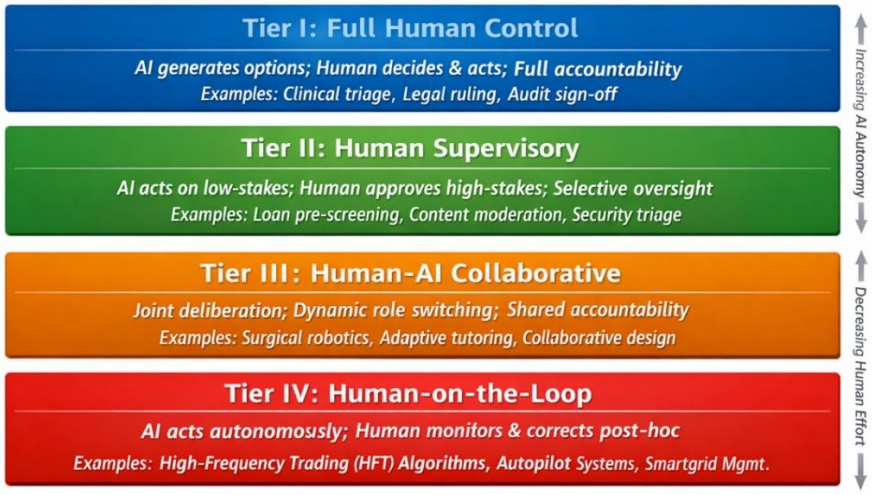
Framework	**Interaction Mode**	**Human Effort**	**Human Roles**	**Typical Context**	**Strengths & Limitations**
Interactive ML (IML)	Loop-based	Mixed	Annotator, oracle, trainer	High volume, low stakes	Efficient label acquisition; scalable but susceptible to annotator fatigue (Li & McAdams, 2025)
Active Learning (AL)	Query-based	Moderate	Domain expert as oracle	Specialized labelling tasks	Reduces labelling cost; optimal query strategy is domain-sensitive (Munro, 2021)
Human-Guided Reinforcement	Reward shaping	High	Trainer, reward designer	Robotic control, game AI	RLHF aligns LLM outputs; reward misspecification risk (Ziegler et al., 2020)
Explainable AI (XAI)	Explanation loop	Variable	End-user, analyst	Clinical, legal, financial	Improves trust calibration; explanation fidelity-comprehensibility tradeoff (Arrieta et al., 2020)
Human-in-the-Loop ML (HITML)	Integrated pipeline	Moderate-High	Data scientist, SME	Enterprise ML workflows	Munro (2021) six-stage model; quality bottleneck at expert annotation
Society-in-the-Loop (SITL)	Societal contract	Low (systemic)	Public, regulators, NGOs	Policy, governance, ethics	Rahwan (2018) in Vogl (2021) extends HITL to the social contract; it lacks operational specificity
Human-on-the-Loop (HOTL)	Monitor & correct	Low (per task)	System operator, auditor	Autonomous vehicles, HFT	Efficient for low-error AI; vigilance decrement risk (Parasuraman & Manzey, 2010) in (Kazim & Tomlinson, 2023)
Collaborative AI (CAI)	Joint deliberation	Variable	Co-agent, paired human	Medical diagnosis, design	Complementarity only if role boundaries and competency are visible (Hemmer et al., 2021)

ISSN No. 2454-6194 | DOI: 10.51584/IJRIAS |Volume XI Issue III March 2026

A Tiered Taxonomy of HITL Frameworks

Synthesizing the LOA literature (Parasuraman et al., 2000), Shneiderman's (2022) four-quadrant HCAI model, and the empirical findings reviewed in this study, the study proposed a four-tier taxonomy that captures the essential variation in human oversight depth across contemporary agentic AI deployments, as shown in Figure 3.

Figure 3 Taxonomy of HITL Frameworks by AI Autonomy Level and Human Oversight Role.



Tier I (Full Human Control) represents configurations in which AI systems generate recommendations, options, or analysis, but all consequential decisions and actions are taken by human agents. This tier is characteristic of clinical decision support in high-stakes specialties (Cabitza et al., 2021; Cai et al., 2021), judicial risk assessment contexts (Wachter et al., 2021), and financial audit workflows. The defining feature of Tier I is that human accountability is unambiguous and unshared.

Tier II (Human Supervisory Control) describes systems in which AI acts autonomously on lower-stakes or time-sensitive subtasks but refers higher-stakes decisions to human review. Loan pre-screening, content moderation queues, and cybersecurity triage exemplify Tier II configurations. Empirical evidence shows that Tier II configurations are particularly susceptible to automation bias when the proportion of AI-handled cases is high relative to human-reviewed cases, as attentional resources devoted to monitoring are insufficient to catch AI errors (Kazim & Tomlinson, 2023).

Tier III (Human-AI Collaborative) represents genuinely joint deliberation, in which roles shift dynamically based on task requirements, uncertainty levels, or domain. Hemmer et al. (2021) find that Tier III configurations produce the highest complementarity benefits, outcomes exceeding both unaided human and AI-only baselines, when AI competency boundaries are made visible to human collaborators. Surgical robotics, adaptive tutoring systems, and collaborative engineering design tools exemplify this tier.

Tier IV (Human-on-the-Loop) encompasses configurations in which AI operates autonomously, with human overseers monitoring for anomalies and correcting post-hoc. Algorithmic trading, power grid management, and advanced driver-assistance systems (ADAS) in highway conditions represent Tier IV deployments. The principal risk at this tier is vigilance decrement; the well-documented tendency for human monitors to miss critical events when the base rate of AI errors is low (Endsley, 2023).

Oversight Mechanisms

Beyond the structural tier in which a system operates, oversight quality is determined by the specific mechanisms employed to facilitate meaningful human review. Our analysis identifies five classes of mechanisms that recur across sectors and framework types: explainability mechanisms, intervention mechanisms, audit and accountability mechanisms, training and onboarding mechanisms, and regulatory and governance mechanisms.

ISSN No. 2454-6194 | DOI: 10.51584/IJRIAS |Volume XI Issue III March 2026

Explainability Mechanisms

Explainable AI (XAI) constitutes the most extensively studied class of oversight mechanisms in this study's corpus. Arrieta et al.'s (2020) comprehensive taxonomy identifies 29 distinct XAI methods, spanning model-agnostic post-hoc explanations (LIME, SHAP), attention mechanisms, concept-based explanations, and counterfactual explanations. A critical finding from our synthesis, however, is that the relationship between explanation quality and oversight effectiveness is neither linear nor simple.

Ehsan et al. (2021) and Wang et al. (2024) found that global explanations that describe the AI's overall decision logic reduced automation bias more effectively than local explanations that describe individual predictions, because global explanations enabled operators to form accurate mental models of the system's failure modes rather than merely its current output. Bansal et al. (2021) demonstrated the converse risk misleading explanations; those that do not faithfully represent the AI's actual reasoning degraded team performance below baseline, sometimes below unaided human performance, because operators updated their decision strategies based on false information about AI competence.

Cabrera et al. (2023) found that behavioural descriptions of AI, that is, natural language summaries of what an AI system does and does not do well, outperformed numerical accuracy statistics for trust calibration, particularly among domain experts without strong statistical training. This finding has significant practical implications for clinical AI, where physician trust in decision support systems has historically been constrained by the opacity of predictive models (Cai et al., 2021; Shortliffe & Sepúlveda, 2022).

The tendency for more interpretable models to sacrifice predictive accuracy, which is described as the explainability-performance tradeoff, is a persistent challenge in the literature (Arrieta et al., 2020; Shneiderman, 2020). Recent work on inherently interpretable neural architectures (Ouyang et al., 2022; Rudin, 2019) suggests this tradeoff is not immutable for all problem types, but it remains binding in domains with high-dimensional inputs such as radiology, genomics, and natural language understanding.

Intervention Mechanisms

Intervention mechanisms are procedures by which human overseers can interrupt, redirect, or override AI actions. Their design is a central concern in the AI safety literature (Amodei et al., 2016) and is now receiving increasing attention from human factors researchers studying agentic AI in production environments.

Hadfield-Menell and Hadfield (2019) frame the design of intervention mechanisms as an incomplete contracting problem: because it is impossible to anticipate all contingencies under which a human might need to override an AI agent, systems must be designed with conservative fallback behaviours and low corrigibility thresholds in high-stakes domains. This insight is operationalized in contemporary LLM agent design through techniques such as tool-use sandboxing, action confirmation gates, and reversibility constraints on irreversible actions.

A recurring finding in our corpus is that the usability of intervention mechanisms critically mediates their effectiveness. Systems that require cognitively demanding override procedures are frequently bypassed by operators under time pressure, a phenomenon documented in clinical AI (Cabitza et al., 2021), content moderation (Sloane et al., 2020), and financial fraud detection. Munro (2021) argues that intervention interfaces should be co-designed with frontline operators, as the gap between a system's theoretical override capacity and its practical use rate represents a major source of oversight failure.

Audit and Accountability Mechanisms

Audit mechanisms provide the evidentiary infrastructure for post-hoc accountability: logs, decision trails, explanation records, and human sign-off timestamps. Wachter et al. (2021) argue that automated audit mechanisms are necessary but insufficient for accountability under EU anti-discrimination law, because meaningful accountability requires that a specific human be identifiable as responsible for a consequential AI decision.

ISSN No. 2454-6194 | DOI: 10.51584/IJRIAS |Volume XI Issue III March 2026

The challenge of accountability attribution in multi-agent systems, where AI agents delegate subtasks to other agents, or where chains of human-AI interaction span multiple organizations, is identified by multiple studies as an emerging frontier problem (Bommasani et al., 2022; Rahwan, 2021). The NIST AI Risk Management Framework (2023) addresses this through its GOVERN function, which requires organizations to define accountability structures as a precondition for responsible AI deployment, but does not specify technical mechanisms for attribution in distributed agentic pipelines.

Training and Onboarding Mechanisms

Effective oversight requires not only well-designed systems but also appropriately trained human operators. Cai et al.'s (2021) qualitative study of clinical AI onboarding found that practitioners need mental models of AI failure modes, not merely aggregate accuracy statistics, to exercise effective oversight. A clinician who knows that a sepsis prediction model has high sensitivity but poor specificity in elderly patients can exercise more meaningful oversight than one who knows only that the model achieves 0.89 AUROC.

Madaio et al. (2020) demonstrate that organizational training for AI oversight faces structural obstacles: fairness-oriented checklists failed in their study not because of poor design but because organizational incentive structures did not reward the time investment required to complete them. This finding illustrates the STS insight that technical training interventions are insufficient without corresponding changes to organizational structures and incentives.

Regulatory and Governance Mechanisms

Regulatory mechanisms constitute the most macro-level class of oversight mechanisms. The European Union AI Act (2024) establishes a risk-based framework that maps directly onto the HITL tier taxonomy: high-risk AI applications, including clinical AI, facial recognition, and employment screening, require conformity assessments, transparency obligations, human oversight by design, and accuracy and robustness testing. Prohibited applications include real-time biometric surveillance and social scoring systems.

The NIST AI Risk Management Framework [NIST AI RMF 1.0, 2023] takes a complementary approach, organizing AI risk management into four functions, which are GOVERN, MAP, MEASURE, and MANAGE (Tabassi, 2023). Unlike the EU AI Act's prescriptive risk categories, the NIST framework is voluntary and outcome-oriented, intended to be applicable across sectors and regulatory jurisdictions. Multiple studies in our corpus report piloting or adapting the NIST framework for sector-specific contexts, including healthcare (World Health Organization, 2021) and financial services.

Sector-specific oversight intensity, reflecting the combined demands of task criticality and regulatory pressure, is summarized in Figure 4.

Figure 4 Sector-Specific Oversight Intensity Matrix Derived from Systematic Review Findings.

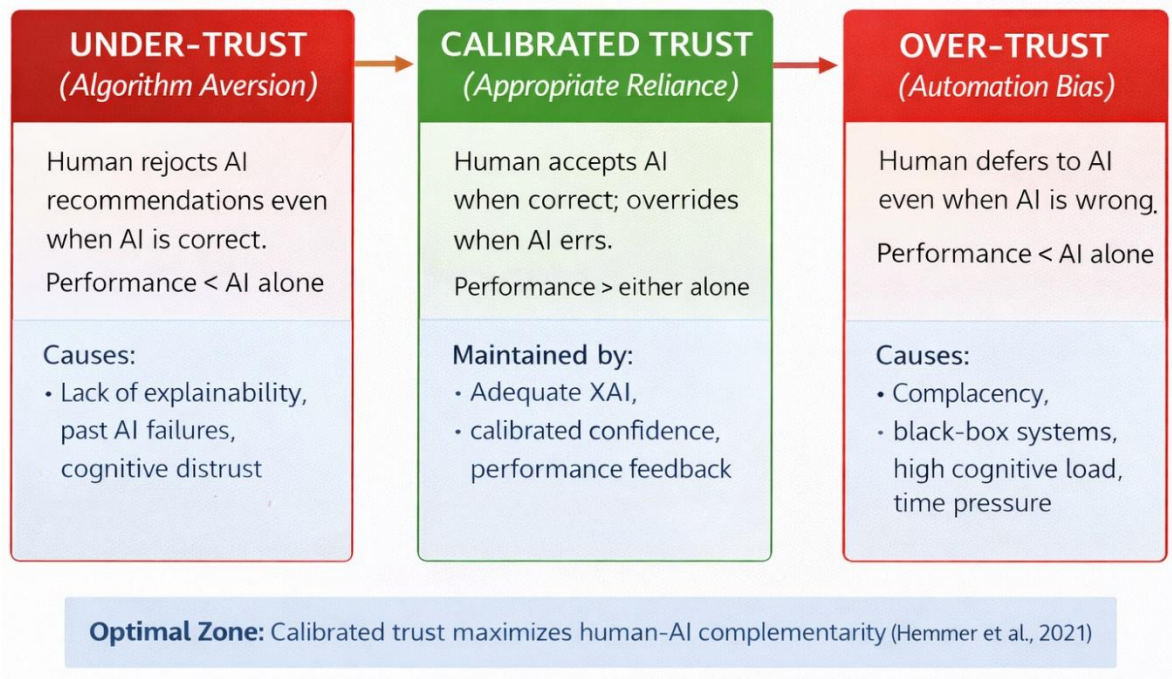
Sector	Task Criticality	Regulatory Pressure	Recommended HITL Tier	Oversight Intensity
Healthcare / Clinical	Critical	High	Tier I–II	Very High (9/10)
Criminal Justice / Legal	Critical	High	Tier I	Maximum (10/10)
Financial Services	High	High	Tier II–III	High (7/10)
Autonomous Vehicles	Critical	Growing	Tier III–IV	High (8/10)
Education / Tutoring	Moderate	Moderate	Tier II–III	Moderate (5/10)
Manufacturing / IIoT	High	Moderate	Tier III–IV	Moderate (6/10)
Content Moderation	High	Growing	Tier II	High (7/10)
HR / Recruitment	Moderate	Moderate	Tier I–II	Moderate (6/10)

ISSN No. 2454-6194 | DOI: 10.51584/IJRIAS |Volume XI Issue III March 2026

Trust Calibration

Trust calibration, that is, the alignment of subjective operator trust with objective AI reliability, is the central mediating variable between oversight mechanism design and human-AI team performance. Miscalibrated trust degrades performance in both directions. Under-trust (algorithm aversion) causes operators to reject correct AI recommendations, while over-trust (automation bias) causes operators to follow incorrect AI recommendations (Lee & See, 2004; Zhang et al., 2020). Figure 5 illustrates the trust calibration spectrum.

Figure 5 The trust calibration spectrum showing the performance consequences of under-trust, calibrated trust, and over-trust.



Algorithm Aversion and Under-Trust

Algorithm aversion, the tendency to prefer human judgment over algorithmic judgment even when the algorithm demonstrably outperforms humans, was first systematically documented by Dietvorst et al. (2015) in forecasting contexts. The mechanism is experiential, having observed an algorithm err, observers penalize it more harshly than they penalize humans for equivalent errors, creating an asymmetric error tolerance. Dietvorst and Bharti (2020) subsequently showed that this aversion is not a stable personality trait but a context-sensitive response to diminishing sensitivity to forecasting error: when the consequences of algorithmic error are uncertain, human decision-makers retreat to familiar human judgment.

Importantly, Dietvorst & Bartels (2022) demonstrated a practical remedy. Allowing decision-makers to make small modifications to algorithmic outputs, even cosmetic ones with no real effect on the prediction, substantially reduced aversion, apparently because the act of modification restored a sense of agency. This finding has direct design implications for HITL interfaces.

Logg et al. (2019) and Logg & Minson (2022) documented the complementary phenomenon of algorithm appreciation, conditions under which people prefer algorithmic to human judgment, and found it is most pronounced in domains perceived as opaque and least pronounced in domains perceived as requiring human qualities, including character assessment and ethical judgment. These findings caution against generic claims about algorithm aversion and underscore the domain-specificity of trust dynamics.

Automation Bias and Over-Trust

Automation bias, the tendency to over-rely on automated recommendations, is documented across aviation (Jerry et al., 2026; Mosier et al., 1998), clinical medicine (Cabitza et al., 2021), and judicial risk assessment

ISSN No. 2454-6194 | DOI: 10.51584/IJRIAS |Volume XI Issue III March 2026

(Dressel & Farid, 2018). It is most pronounced under high cognitive load, time pressure, and in situations where the AI recommendation is presented prior to the human's own deliberation (Zhang et al., 2020).

Romeo and Conti (2026) argue that complacency and automation bias, while phenomenologically similar, have distinct causal pathways. Complacency is primarily an attentional phenomenon that monitors how effort decreases over time when the AI rarely errs, whereas automation bias is primarily a decision-process phenomenon when AI recommendations serve as strong anchors that are insufficiently adjusted. This distinction matters for oversight design since complacency requires mechanisms that maintain vigilance, whereas automation bias requires mechanisms that interrupt anchoring, requiring explicit human prediction before AI disclosure, or presenting AI confidence alongside human-elicited confidence.

Zhang et al. (2020) conducted a controlled experiment demonstrating that displaying AI confidence scores without accompanying explanations produced significantly worse trust calibration than either displaying scores with explanations or displaying neither. This finding suggests that partial transparency, providing numerical outputs without interpretive context, may be worse than opacity for calibration purposes, a counterintuitive result with significant implications for dashboard design.

Achieving Calibrated Trust

Hemmer et al.'s (2021) meta-synthesis identified three conditions necessary for achieving calibrated trust: (i) AI competency visibility—operators must be able to identify the conditions under which the AI is likely to be right or wrong; (ii) feedback timeliness—performance feedback must be proximate enough in time to update operator mental models; and (iii) role clarity—the boundary between AI and human decision authority must be unambiguous.

Cabrera et al. (2023) found that brief behavioural descriptions of AI capabilities improved calibration more than numerical performance metrics, particularly for expert users whose existing mental models of AI were systematically biased toward assuming AI superiority in their own domain.

Bansal et al. (2021) introduced the important distinction between local and global trust. Local trust refers to trust in a specific AI output, while global trust refers to trust in the system as a whole. Explanations appear to influence local trust more reliably than global trust, implying that explanation-based calibration must be supplemented with system-level performance communication to prevent global miscalibration.

Sector-Specific Analysis of Oversight Frameworks

The following sections summarize findings from the most extensively studied sectors in the corpus of this study.

Healthcare and Clinical AI

Healthcare represents the largest and most methodologically diverse sector in this study’s corpus (n = 47 studies). The combination of high task criticality, complex regulatory environments, and the well-documented phenomenon of automation bias in clinical settings makes healthcare a paradigmatic case for HITL design.

Cabitza et al.'s (2021) systematic review argues for the ‘rehumanization’ of clinical AI, a deliberate reduction in automation scope to restore physician agency and accountability. This position is supported by evidence that physicians derive epistemic confidence from the process of clinical reasoning itself, not merely from the accuracy of its outputs, and that AI-assisted workflows that bypass this process produce worse outcomes even when the AI is more accurate than the physician Cai et al. (2021). Topol (2019) famously argued that AI's highest-value role in medicine is to free physicians from data-processing tasks to spend more time on tasks requiring empathy and judgment. This is a vision that requires Tier I-II oversight architectures rather than high-autonomy deployments.

Clinical decision support (CDS) systems represent the most mature application of HITL principles in healthcare. Shortliffe and Sepúlveda (2022) trace the evolution of CDS from rule-based alert systems to ML-

ISSN No. 2454-6194 | DOI: 10.51584/IJRIAS |Volume XI Issue III March 2026

driven recommendation engines, noting that physician trust in CDS peaks when it is integrated seamlessly into clinical workflow, presenting recommendations at the point of care without requiring navigation to a separate interface, and collapses when alert fatigue, driven by high false-positive rates, causes systematic overriding. This alert fatigue phenomenon is a specific instance of the automation bias-under-trust cycle: initial over-reliance is followed by learned dismissal, with calibrated trust as the narrow path between.

Criminal Justice and Legal AI

Criminal justice applications of AI, particularly recidivism prediction, facial recognition, and predictive policing, sit at the intersection of high task criticality and intense regulatory scrutiny, making them a critical test case for meaningful human oversight. The US Supreme Court's 2016 refusal to review State v. Loomis, which upheld the use of the COMPAS recidivism tool, effectively validated HITL configurations in which judicial discretion is nominally preserved but practically constrained by algorithmic scores, drawing sharp criticism from legal scholars (Dressel & Farid, 2018; Wachter et al., 2021).

Dressel and Farid (2018) demonstrated that COMPAS's predictive accuracy was matched by a simple two-question heuristic and was no better than that of Amazon Mechanical Turk workers with no criminal justice training, raising fundamental questions about what value the AI component adds to oversight-nominally human decisions. Wachter et al. (2021) argue that the EU AI Act's prohibition on AI systems that exploit psychological vulnerabilities and its requirement for human oversight in high-risk contexts represent necessary but insufficient guardrails, because the institutional context in which judges operate may itself bias 'human oversight' in ways that AI merely amplifies.

Financial Services

Financial services present a distinctive oversight challenge because of the multi-scale nature of risk: individual transactions, portfolio positions, and systemic market stability are distinct loci of risk, each requiring different oversight architectures. Algorithmic trading systems at Tier IV (human-on-the-loop) are considered acceptable for individual transaction execution but have contributed to systemic events such as the 2010 Flash Crash, where automated sell orders cascaded beyond human intervention capacity (Jansen, 2021).

The application of HITL frameworks to credit scoring, lending, and insurance underwriting is increasingly shaped by the EU AI Act's high-risk classification of these applications and by GDPR's requirement for meaningful human review of automated decisions. Logg and Minson (2022) find that algorithm appreciation in financial forecasting is high, with finance professionals expressing greater willingness to defer to AI than professionals in most other domains. This created a sector-specific over-trust risk that automated audit mechanisms must compensate for.

Autonomous Transportation

Autonomous vehicles (AVs) represent perhaps the clearest example of the oversight intensity tradeoff. The principal argument for reducing human involvement in driving is that human error causes over 90% of road fatalities, yet the transition period, in which partially automated systems share control with human drivers, may introduce new risk categories through complacency and mode confusion. Endsley (2023) identifies AV operations as a prototype case for Level-3 Situation Awareness (SA) failure. Drivers in SAE Level 3 systems, which can handle all driving under defined conditions but require human takeover when the system's limits are reached, must maintain sufficient situation awareness to resume control within seconds despite having been out of the loop for extended periods.

Parasuraman and Manzey (2010) in Romeo and Conti (2026) document that takeover times in Level 3 AV contexts average 5–8 seconds, intervals within which collisions at highway speeds are typically unavoidable, and are significantly lengthened by secondary task engagement. This evidence has contributed to regulatory skepticism about SAE Level 3 systems and growing support for either full human control or full AI control, bypassing the problematic intermediate state.

ISSN No. 2454-6194 | DOI: 10.51584/IJRIAS |Volume XI Issue III March 2026

Education

Intelligent tutoring systems (ITS) and AI-driven adaptive learning platforms represent Tier II-III deployments with moderate oversight intensity. The educational context differs from other high-stakes sectors in that the primary risk is not catastrophic harm but cumulative inefficacy and inequitable access to quality instruction. Alfredo et al. (2024) and Gomes (2024) find that teacher oversight of AI tutoring systems is most effective when teachers receive actionable dashboards summarizing student engagement and knowledge state, rather than raw predictive scores.

The participation washing concern Sloane et al. (2020) is particularly salient in educational AI. Technology companies have engaged in extensive consultation with teachers about AI tool design while retaining full control over deployment decisions, data governance, and algorithmic updating, which is a pattern that creates legitimacy without substantive co-governance.

Content Moderation

AI-assisted content moderation represents a Tier II application operating at a previously unimaginable scale: major platforms process billions of posts daily, making genuine human review of each item impossible. The HITL configuration in content moderation is therefore asymmetric. AI handles the vast majority of cases autonomously, with human review reserved for appeals, borderline cases, and quality audits. This creates a systematic oversight gap at the individual case level, which Sloane et al. (2020) argue is not adequately compensated by system-level accuracy metrics.

The human cost of content moderation is a distinctive sector-specific finding. Studies document severe psychological harm among content moderators exposed to large volumes of harmful material, with consequences for both moderator well-being and oversight quality (Cabrera et al., 2023). This dimension of the well-being of humans in the loop is largely absent from technical HITL frameworks but is identified by multiple studies as essential to sustainable oversight architectures.

RESULTS

Synthesis: The Adaptive Oversight Calibration Model (AOCM)

The review findings reveal a consistent gap in the HITL literature. Existing frameworks tend to be either operationally specific, demonstrated in being tied to a particular sector, technology, or mechanism, or theoretically abstract by providing normative principles without design-level operationalization. The review found that a mid-range theory is missing. The mid-range theory is a framework that is sector-agnostic enough to generalize across deployment contexts, yet concrete enough to generate testable propositions and guide design decisions. The Adaptive Oversight Calibration Model (AOCM), developed through thematic synthesis of the 187-study corpus, is designed to fill the gap.

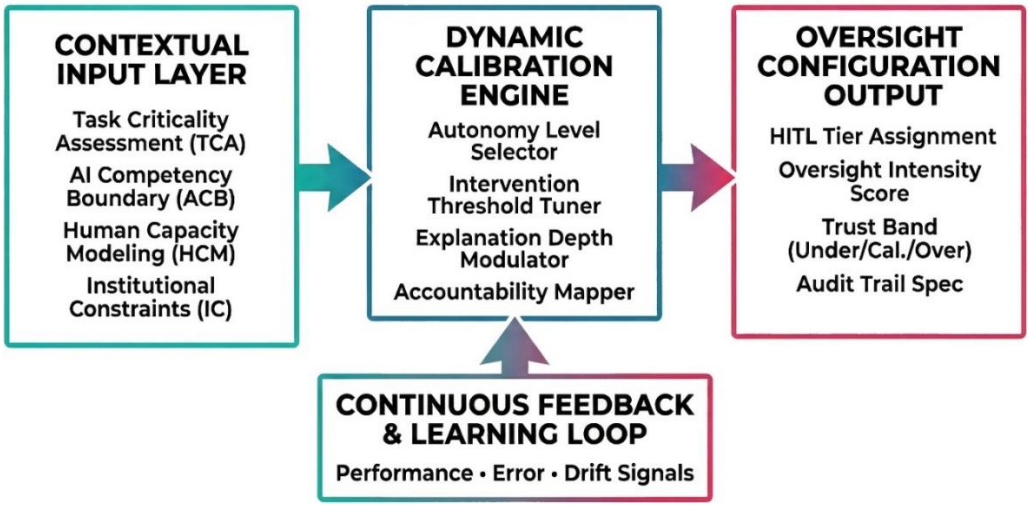
Model Overview

The Adaptive Oversight Calibration Model (AOCM) is structured into four functional components, each containing specific variables designed to ensure context-sensitive human-AI collaboration. The components are the contextual input layer, dynamic calibration engine, oversight configuration output, and continuous feedback and learning loop. The AOCM proposes that optimal human oversight in agentic AI systems is not a fixed configuration but a dynamic function of four context-sensitive inputs. The first is task criticality, which refers to the severity of potential harms that could result from AI errors. The second is AI competency boundaries, meaning the distribution of tasks over which an AI system's outputs can be considered reliable. The third is human cognitive capacity, encompassing the bandwidth, expertise, and attentional resources available to human overseers. The fourth is institutional constraints, which include regulatory requirements, organizational structures, and accountability frameworks. These inputs feed into a calibration engine that produces an oversight configuration comprising a HITL tier assignment, an oversight intensity score, a trust band classification, and an audit trail specification. Essentially, oversight configurations are continuously

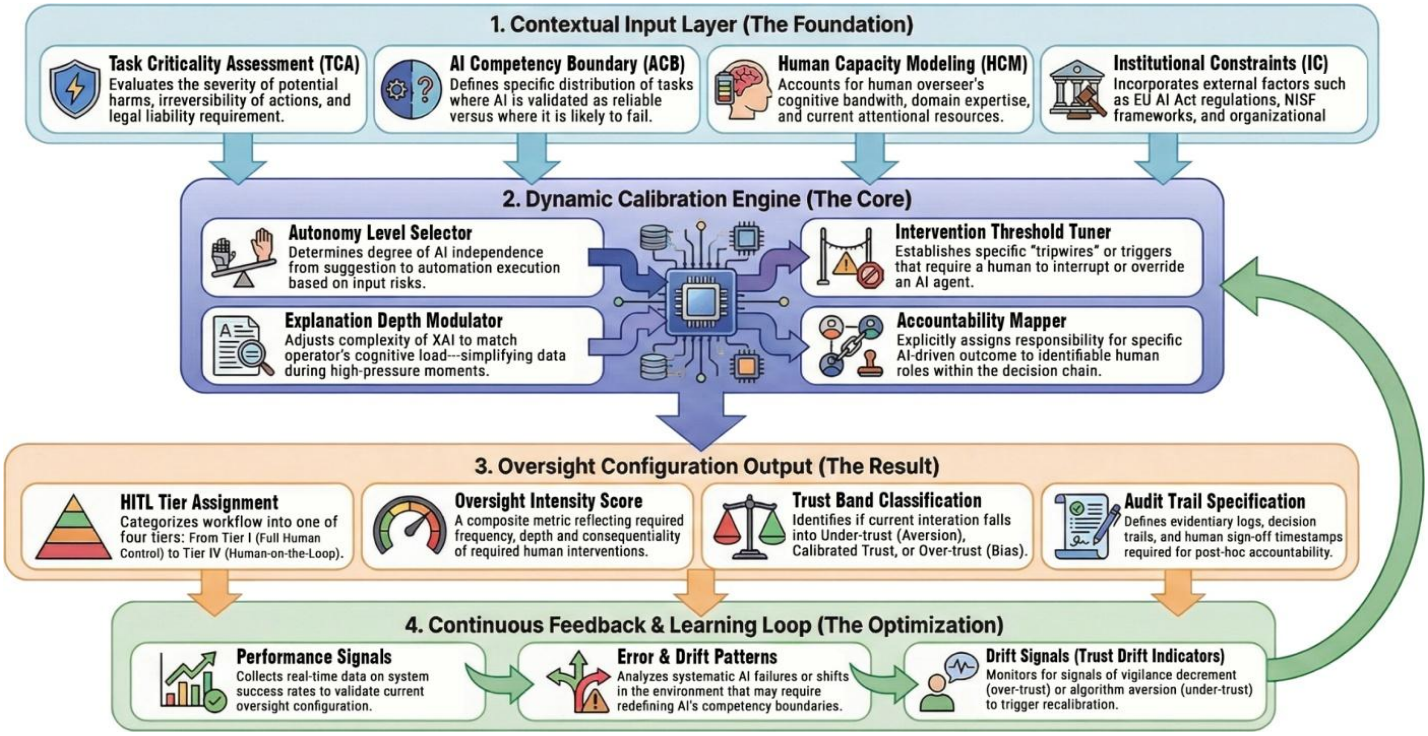
ISSN No. 2454-6194 | DOI: 10.51584/IJRIAS |Volume XI Issue III March 2026

updated through a feedback loop that aggregates performance signals, error patterns, and trust drift indicators. Figure 6 presents the AOCM schematically.

Figure 6 The Adaptive Oversight Calibration Model (AOCM)



6(a) Simplified Adaptive Oversight Calibration Model (AOCM)



6(b) Annotated Adaptive Oversight Calibration Model (AOCM)

Note. The Adaptive Oversight Calibration Model (AOCM): a dynamic framework for contextually appropriate human oversight of agentic AI systems. Figure 6(a) shows a simplified AOC Model; Figure 6(b) shows an annotated AOC Model. This model maps how the Contextual Input Layer, comprising task criticality, AI competency, human capacity, and institutional constraints, feeds into a Dynamic Calibration Engine to produce specific Oversight Configuration Outputs. The configuration is then refined through a Continuous Feedback & Learning Loop that monitors performance and trust signals.

Formal Propositions

The AOCM is formalized through six propositions derived from the synthesized evidence. These propositions are presented as conditionals amenable to empirical testing via experimental, quasi-experimental, or

ISSN No. 2454-6194 | DOI: 10.51584/IJRIAS |Volume XI Issue III March 2026

observational designs. Table 4 presents the propositions alongside their formal statements, operationalizations, and evidential grounding.

Table 4 AOCM Formal Propositions: Statements, Operationalizations, and Evidential Support

ID	Proposition Name	Formal Statement	Operationalization & Support
P1	Task Criticality → Oversight Tier	Higher task criticality, defined by irreversibility, harm potential, and legal liability, necessitates a lower HITL tier, requiring more direct human control.	*HITL tier = f(irreversibility, harm, liability)* Supported by Cabitza et al. (2021) and Shneiderman (2022)
P2	AI Competency Boundary → Intervention Threshold	Oversight interventions should be triggered proportionally to the probability that an AI system is operating outside its validated competency boundary.	*Intervention rate = g(P[out-of-distribution])* Supported by Amodei & Hernandez (2022) and Bommasani et al. (2022)
P3	Human Cognitive Load → Explanation Depth	As operator cognitive load increases, the depth of AI explanations should decrease; compressed, actionable summaries outperform detailed explanations under time pressure.	*XAI depth = h(1 / cognitive_load)* Supported by Ehsan et al. (2021), Endsley (2023), and Wang et al. (2024)
P4	Trust Calibration → Performance	Human-AI team performance is maximized when trust is calibrated (neither under- nor over-trust), and is degraded symmetrically by both extremes.	*Performance = α - β|trust - optimal_trust|* Supported by Hemmer et al. (2021) and Zhang et al. (2020)
P5	Feedback Loop → Trust Drift Prevention	Continuous performance feedback loops prevent both vigilance decrement (leading to over-trust) and algorithmic aversion (leading to under-trust) over time.	*Trust drift rate → 0 with feedback* Supported by Dietvorst and Bartels (2022), and Romeo and Conti (2026)
P6	Institutional Constraints → Accountability Mapping	Effective oversight requires explicit mapping of accountability to specific human roles; diffuse accountability predicts worse safety outcomes than role-specific assignment.	*Error recovery ∝ 1 / accountability_diffusion* Supported by Sloane et al. (2020) and Wachter et al. (2021)

AOCM Distinguished from Prior Frameworks

The AOCM differs from six antecedent frameworks across eight dimensions, as demonstrated in Table 5.

Table 5 Comparative Analysis: AOCM Versus Prior Human Oversight Frameworks Across Eight Dimensions

Dimension	**LOA Taxonomy (Parasuraman et al., 2000)**	**IML / AL** (Munro, 2021)	**SITL (Rahwan, 2021)**	**HCAI (Shneiderman, 2022)**	**NIST AI RMF (Tabassi, 2023)**	**V.A.L.U.E. (Adedokun et al., 2026)**	**AOCM (This Study)**
Scope	Automation spectrum	Annotation pipeline	Social contract	Design principles	Risk management	Ontological and ethical risks of AI	Oversight architecture
Oversight Dynamism	○ Partial	✗ No	○ Partial	○ Partial	○ Partial	✗ No	✓ Yes
Trust Calibration as Variable	✗ No	✗ No	✗ No	○ Partial	✗ No	✗ No	✓ Yes
Human Capacity as Input	○ Partial	✗ No	✗ No	✗ No	○ Partial	○ Partial	✓ Yes

ISSN No. 2454-6194 | DOI: 10.51584/IJRIAS |Volume XI Issue III March 2026

Institutional Constraints	✗ No	✗ No	○ Partial	✗ No	○ Partial	○ Partial	✓ Yes
Testable Propositions	✗ No	✗ No	✗ No	✗ No	✗ No	✗ No	✓ Yes
Sector Applicability	Industrial/Military	ML pipelines	Society-wide	Cross-sector	Cross-sector	Cross-sector	Cross-sector
Post-Deployment Recalibration	✗ No	✗ No	✗ No	✗ No	○ Partial	✗ No	✓ Yes

Note. Comparative Analysis. ✓ Yes = feature explicitly present and operationalized ○ Partial = feature present but not fully operationalized ✗ No = feature absent

Parasuraman et al.'s (2000) Levels of Automation taxonomy provides a graduated autonomy spectrum but treats human capacity and institutional constraints as contextual givens rather than modellable inputs. Monarch's (2021) Interactive Machine Learning pipeline addresses annotation efficiency but is limited to model training contexts and does not address trust calibration as an outcome. Rahwan's (2021) Society-in-the-Loop framework extends oversight to collective governance but lacks operational specificity for system-level design. The NIST AI RMF (Tabassi, 2023) addresses governance structure and risk categorization but does not model the dynamic recalibration of oversight as AI systems evolve post-deployment. Shneiderman's (2020) Human-Centred AI (HCAI) framework provides normative design principles and a four-quadrant reliability-safety matrix but does not operationalize oversight dynamism or post-deployment recalibration. The V.A.L.U.E. Framework by Adedokun et al. (2026) operates at the normative and philosophical level, providing the value-preservation rationale that the AOCM operationalizes through measurable constructs and testable propositions. The two frameworks are not competitors on these dimensions; they address different layers of the same problem. The AOCM addresses all eight dimensions by treating oversight as a continuously calibrated function of four measurable inputs, integrating trust as an endogenous variable, and providing six testable propositions.

The AOCM differs from existing frameworks in five substantive ways. First, it treats oversight configuration as continuous and dynamic rather than categorical and static, reflecting the empirical finding that optimal oversight changes as AI systems are updated, deployed in new contexts, or encounter distributional shift. Second, it integrates trust calibration as an endogenous variable, which is a product of the oversight configuration rather than an exogenous prerequisite, recognizing that trust is shaped by how oversight is designed. The third substantive way is that AOCM includes human cognitive capacity as an explicit design variable, acknowledging that oversight mechanisms that exceed human attentional or expertise capacity fail regardless of their technical sophistication. Fourth, it incorporates institutional constraints as a first-class input, preventing the naive assumption that technically optimal oversight is implementable in any organization. Lastly, it operationalizes meaningful oversight, a term used but rarely defined in regulatory documents, as a threshold function across the four inputs, providing a basis for compliance assessment.

Applying the AOCM: Illustrative Case Analyzes

Clinical AI Diagnostic Support (High Criticality, High Regulatory Pressure)

A case study of a radiological AI system trained to detect lung nodules on CT scans. Task criticality is high (missed diagnosis is potentially fatal). AI competency is well-bounded (the model performs reliably above 6mm diameter but is unreliable for sub-centimeter nodules). Human capacity is constrained (radiologists face high volume and fatigue). Institutional constraints include FDA clearance requirements and liability standards. The AOCM calibration engine assigns Tier II oversight: the AI generates a preliminary report with explicit confidence levels per nodule; the radiologist reviews all cases but is specifically prompted to scrutinize sub-centimeter findings. Explanations are brief and actionable (not SHAP plots). Audit trails record both AI predictions and radiologist decisions for quality assurance. The feedback loop monitors radiologist override rates: a rate below 2% triggers an automation bias alert; a rate above 25% triggers system reliability investigation.

ISSN No. 2454-6194 | DOI: 10.51584/IJRIAS |Volume XI Issue III March 2026

Autonomous LLM Coding Agent (Moderate Criticality, Low Regulatory Pressure)

Considering an LLM-based coding agent tasked with generating and executing code to analyze a database. Task criticality is moderate (errors produce incorrect analyzes but are rarely catastrophic). AI competency boundary is wide but not universal (the agent handles well-specified SQL queries reliably but struggles with ambiguous requirements or novel data structures). Human capacity is high (the operator is a data engineer with relevant expertise). Institutional constraints are minimal. The AOCM assigns Tier III oversight: the agent acts autonomously on execution of confirmed queries but requires human approval before schema modification or external API calls. Explanations take the form of step-by-step reasoning traces, surfaced at checkpoints rather than after every action. The trust band is monitored via the operator's override rate, and the system is flagged for over-trust if the operator approves all agent recommendations without reviewing execution logs.

Content Moderation at Scale (High Criticality, Growing Regulatory Pressure)

Considering a major social media platform's content moderation pipeline. Task criticality is high for specific categories, including child safety, and terrorist incitement; and moderate for others, including as spam and hate speech. AI competency is strong for clear-cut violations but degrades sharply for context-dependent cases. Human capacity is constrained by volume and psychological harm. The AOCM assigns differentiated tier configurations: Tier IV for clear-cut, high-confidence AI removals (human monitoring via random audit); Tier I for child safety content (mandatory human review before action); Tier II for borderline cases (AI flags, human decides). Institutional constraints are operationalized through documented accountability matrices specifying which human roles are responsible for each decision category. Moderator wellbeing is incorporated as a constraint on human capacity, triggering capacity recalculation when moderator distress indicators exceed thresholds.

Preliminary Simulation of AOCM Propositions

To assess the internal coherence and preliminary feasibility of the AOCM, the six propositions were tested against the three illustrative cases in this study using a cross-case activation matrix. For each case, every proposition was evaluated on three criteria: whether the proposition's precondition was present in the case context; whether the predicted configuration aligned with the case's optimal oversight assignment; and whether any proposition failed to activate or produced an inconsistent prediction. Table 6 presents the activation matrix. Cells are colour-coded, with green indicating the proposition activated and its prediction was fully consistent with the case configuration; amber indicates partial activation or a boundary condition that qualifies the prediction; red indicates a gap where the proposition did not activate or produced an inconsistent result.

The simulation reveals that five of six propositions are fully supported across all three cases. The single gap is Proposition P6 in Case 2 (LLM Coding Agent). Because the deployment involves a single expert operator with no institutional accountability framework, the accountability mapping required by P6 is implicit rather than formally assigned. This gap is theoretically important as it suggests that the AOCM's P6 requires a minimum level of organizational formalization to activate, and that low-constraint deployments present a specific oversight vulnerability even when technical configurations are otherwise sound. Future empirical research should operationalize the institutional constraints variable to establish the threshold below which P6's accountability mapping becomes infeasible.

Table 6 AOCM Proposition Simulation: Cross-Case Activation Matrix (Three Illustrative Scenarios)

Proposition	**Proposition Name**	**Formal Prediction**	**Case 1 Clinical AI**	**Case 2 LLM Coding Agent**	**Case 3 Content Moderation**
P1	Task Criticality→ HITL Tier	Higher criticality→ lower HITL tier (more direct human control)	**High** Tier II: radiologist reviews all; sub-centimetre cases	**Moderate** Tier III: agent acts autonomously on confirmed queries only	**Differentiated** Tier I for child safety; Tier IV for spam; Tier II for borderline cases

ISSN No. 2454-6194 | DOI: 10.51584/IJRIAS |Volume XI Issue III March 2026

			specifically prompted		
P2	AI Competency Boundary→ Intervention Threshold	Intervention rate proportional to probability of out-of-distribution operation	**Well-bounded** Sub-centimetre nodule detection triggers radiologist-specific prompt	**Wide but partial** Schema changes and novel data structures trigger mandatory approval gate	**Context-dependent** Context-sensitive violation cases trigger escalation to human review
P3	Cognitive Load→ Explanation Depth	Higher cognitive load→ shallower, more actionable explanations	**High (volume + fatigue)** Brief confidence flags; SHAP plots excluded	**Low (expert user)** Step-by-step reasoning traces surfaced at decision checkpoints	**High (scale + distress)** Category flags only; detailed reasoning chains withheld
P4	Trust Calibration→ Performance	Performance maximized at calibrated trust; degraded by both extremes	**Automation bias risk** Override rate monitored; rate below 2% triggers bias alert	**Over-trust risk** Blanket approval without log review flagged as over-trust signal	**Dual risk** High-confidence removals risk complacency; borderline cases risk aversion
P5	Feedback Loop→ Trust Drift Prevention	Continuous feedback prevents vigilance decrement and algorithm aversion	**Override rate tracking** Rate above 25% triggers reliability review; below 2% triggers alert	**Log review monitoring** Override rate and log engagement tracked per operator session	**Audit sampling** Random Tier IV audits; moderator wellbeing metrics feed the feedback loop
P6	Institutional Constraints→ Accountability Mapping	Explicit role-based accountability assignment predicts better safety outcomes	**FDA and liability constraints** Radiologist named as accountable; AI output is advisory only	**Minimal constraints** Accountability implicit in operator role; no formal assignment exists	**Platform policy framework** Role-specific accountability matrix defined per content category
Proposition Activation Summary: Case 1 (Clinical AI): 5/6 fully supported, 1/6 partial. Case 2 (LLM Coding Agent): 4/6 fully supported, 1/6 partial, 1/6 gap identified (P6). Case 3 (Content Moderation): 5/6 fully supported, 1/6 partial.					

DISCUSSION

Structural Tensions in the HITL Literature

The synthesis in this study reveals four structural tensions that recur across sectors and framework types, reflecting deep features of human-AI collaboration that cannot be engineered away but must be managed.

The explainability-performance tradeoff, in which more interpretable systems tend to sacrifice predictive accuracy, remains the most extensively discussed tension in the corpus. While inherently interpretable models (Arrieta et al., 2020; Rudin, 2019) offer a partial solution for specific problem types, the tradeoff persists in high-dimensional perceptual domains, precisely those where AI offers the greatest marginal value over human judgment. The AOCM addresses this by treating explanation depth as a variable calibrated to human cognitive capacity, not a universal maximum: high-capacity operators in low-volume contexts benefit from rich

ISSN No. 2454-6194 | DOI: 10.51584/IJRIAS |Volume XI Issue III March 2026

explanations, while high-volume operators under time pressure benefit from compressed, actionable summaries.

The autonomy-accountability gap, in which higher degrees of AI agency diffuse and obscure human responsibility, reflects the fundamental legal and organizational challenge of deploying systems that act without being actors in the legal sense. Current accountability frameworks, including the EU AI Act's provider-deployer liability structure, struggle to attribute causality in multi-agent pipelines where AI actions span multiple organizations. The AOCM's accountability mapping proposition (P6) addresses this by requiring explicit role-based accountability assignment as a precondition for any oversight configuration, but acknowledges that this is a design target rather than an achievable condition in all deployments.

The over-trust/under-trust duality, in which miscalibrated trust degrades human-AI complementarity in both directions, is the tension with the most direct implications for system performance. The empirical literature consistently finds that trust is miscalibrated in the absence of deliberate calibration mechanisms (Hemmer et al., 2021; Zhang et al., 2020), and that miscalibration is costly in both directions. The AOCM's trust calibration proposition (P4) formalizes this as a symmetric performance loss function, motivating the use of feedback loops (P5) that detect and correct drift in both directions.

The participation-effectiveness paradox, in which nominally inclusive design processes often fail to alter power structures that govern AI deployment (Madaio et al., 2020; Sloane et al., 2020), highlights the limits of purely technical oversight frameworks: if the organizational and power structures in which AI systems are deployed are not also addressed, technically sound oversight mechanisms will be circumvented or rendered ineffective. The AOCM incorporates institutional constraints as a first-class input, but cannot resolve the underlying political and organizational challenges that this tension reflects.

Conflicting Evidence and Unresolved Debates

The synthesis presented in this study emphasizes convergent patterns, a necessary step for framework derivation. However, several areas of genuine disagreement and weak evidence in the corpus warrant explicit acknowledgment, as they qualify the confidence with which the AOCM’s propositions should be applied.

The most consequential conflict in the explainability literature concerns when explanations help versus hurt. Bansal et al. (2021) and Hemmer et al. (2021) demonstrate that explanations degrade team performance when they do not faithfully represent the AI’s actual reasoning process, producing a false mental model that is worse than no explanation at all. D. Wang et al. (2019) find that global explanations reduce automation bias more reliably than local ones. Yet Lai et al. (2021) show that explanations improve accuracy only when calibrated to the AI’s actual reliability, and that generic explanations applied uniformly can actively reduce decision quality. These findings are not fully reconcilable as no single explanation design strategy consistently improves calibration across tasks, populations, and AI systems. The AOCM’s treatment of explanation depth as a cognitive-load function (P3) captures part of this variability but does not resolve the underlying question of explanation fidelity.

The trust calibration literature contains a second unresolved debate, on whether algorithm appreciation or algorithm aversion is the more common baseline disposition, and whether either is stable across contexts. Logg et al. (2019) and Logg and Minson (2022) document strong appreciation effects in opaque domains, directly contradicting the dominant aversion narrative in the human factors literature (Dietvorst & Bharti, 2020; Parasuraman & Manzey, 2010). The discrepancy appears domain-dependent, but no study to date has mapped trust dispositions systematically across a representative sample of professional contexts. The AOCM treats trust band classification as an empirical output of the feedback loop rather than an assumed baseline, which sidesteps this debate methodologically but does not resolve it theoretically.

A third area of weak and contested evidence concerns HITL effectiveness in high-volume automated pipelines, particularly content moderation and fraud detection. Several studies report that human review at scale produces lower accuracy than full automation, because cognitive fatigue and high base rates of non-violations erode human discriminability below AI performance thresholds (Parasuraman & Manzey, 2010; Sloane et al., 2020).

ISSN No. 2454-6194 | DOI: 10.51584/IJRIAS |Volume XI Issue III March 2026

This creates a direct tension with the normative assumption underlying Tier I and Tier II oversight configurations, that human involvement improves outcomes. The evidence suggests this assumption holds only under conditions of manageable volume, adequate time per case, and meaningful case heterogeneity. Where these conditions are absent, the case for human review rests on accountability and legitimacy grounds rather than accuracy grounds, a distinction that regulatory frameworks rarely make explicit and that the AOCM's current formulation does not fully capture.

Implications for System Designers

System designers operationalizing the AOCM should prioritize three design principles derived from the synthesis. First, make AI competency boundaries legible. The single most consistent predictor of calibrated trust in the study's corpus is whether operators can identify when the AI is likely to err. This requires documentation of training distribution boundaries, explicit uncertainty quantification (Gal & Ghahramani, 2016), and communication of known failure modes in user-accessible language. Second, design for cognitive load: oversight mechanisms that impose attentional demands disproportionate to the oversight benefit will be abandoned in practice. Third, close feedback loops: the temporal gap between AI decisions and outcome feedback is a primary driver of trust miscalibration; shortening this gap, through simulation, rapid outcome notification, or periodic performance debriefing, is among the highest-leverage interventions available.

Implications for Policymakers and Regulators

The AOCM's formalizations have direct implications for regulatory operationalization. The EU AI Act's requirement for human oversight by design in high-risk AI systems is a necessary but under-specified mandate because it does not define what depth of oversight is required, how oversight quality should be measured, or what feedback mechanisms are required to maintain oversight effectiveness over time. The AOCM's six propositions provide a basis for operationalizing these requirements in conformity assessment frameworks.

Adedokun et al.'s (2026) The 5-P Strategic and V.A.L.U.E. frameworks function in relation to this study as a foundational approach. They address the 'why' of human oversight, studying the ontological stakes of ceding human agency to AI, while this study addresses the 'how', that is, the architectural conditions under which meaningful oversight is achievable. The V.A.L.U.E. framework and the AOCM framework are not competing but nested. The V.A.L.U.E. framework's value-preservation goals provide the normative rationale that the AOCM's design propositions seek to operationalize.

The NIST AI RMF's GOVERN function, which requires organizations to define accountability structures, maps directly onto AOCM Proposition P6. However, NIST's framework does not address the dynamic recalibration of oversight as AI systems evolve post-deployment, a gap that the AOCM's continuous feedback loop addresses. Therefore, it is recommended that future iterations of these frameworks incorporate explicit requirements for ongoing oversight calibration, not merely initial oversight design.

Limitations

This review is subject to limitations that constrain the scope of its conclusions. First, the literature search was restricted to English-language publications, potentially excluding significant contributions from non-English-speaking research communities, where both AI deployment and governance research are highly active. Second, publication bias may have inflated effect sizes in the empirical literature, as studies showing null effects of explainability or HITL interventions may be less likely to be published. Third, the AOCM's six propositions, while grounded in the synthesized evidence, are theoretical propositions requiring empirical validation. In addition, the synthesis relied on thematic rather than quantitative meta-analysis, limiting the precision of cross-study comparisons.

Lastly, the AOCM itself embeds normative assumptions, particularly that meaningful oversight is achievable within existing organizational and regulatory structures, that may not hold in highly adversarial environments or in settings where the humans nominally in the loop lack genuine authority to intervene. Addressing these structural conditions requires political and organizational change that no technical framework can substitute for.

CONCLUSION

This systematic review has synthesized 187 peer-reviewed studies on human-in-the-loop frameworks, oversight mechanisms, and trust calibration in agentic AI systems published between 2020 and 2026. The analysis confirms that the delegation frontier, the gap between what AI can do and what humans can safely delegate, is not merely a matter of AI capability but reflects deep structural features of human cognition, organizational accountability, and sociotechnical system design.

The four-tier HITL taxonomy proposed in the study provides a common vocabulary for characterizing oversight architectures across sectors, grounding comparisons that have previously been difficult due to terminological fragmentation in the field. The five classes of oversight mechanisms identified in the study are explainability, intervention, audit and accountability, training, and regulatory mechanisms, which represent the primary design levers available to practitioners, each with documented strengths and failure modes.

The Adaptive Oversight Calibration Model (AOCM) introduced represents this review's primary conceptual contribution. The AOCM operationalizes meaningful oversight in a way that is both theoretically grounded and empirically testable, formalizing it as a dynamic function of task criticality, AI competency boundaries, human cognitive capacity, and institutional constraints, updated continuously through performance feedback. The model's six formal propositions provide a research agenda for the next wave of empirical work on human-AI teaming and a practical framework for policymakers seeking to operationalize oversight requirements in regulation.

The stakes of getting this right are high. Agentic AI systems are being deployed at scale in healthcare, criminal justice, financial markets, and information infrastructure domains where AI errors can cause irreversible harm at speed and at scale unprecedented in human history. The question is no longer whether AI will be involved in these decisions, but under what conditions, with what oversight, and accountable to whom. This review provides the conceptual tools to begin answering those questions.

REFERENCES

1. Adedokun, S. A., Adeyemo, I. A., Adedokun, D. A., Ogunkan, S. K., & Ogunniyi, O. K. (2026). Artificial Intelligence and the Essence of Humanity: Strategic Frameworks for Utilizing Technology and Preserving Values in an Automated Era. Journal of Science Innovation and Technology Research, 10(9), 196–213. https://doi.org/10.70382/ajsitr.v10i9.069
2. Agwunobi, Z. (2026). A Legal Technology and Digital Trust Infrastructure Framework for Bridging Digital Trust Between Artificial Intelligence and Human Intelligence. International Journal of Scientific Research in Science, Engineering and Technology, 13(2), 55–79. https://doi.org/10.32628/IJSRSET261358
3. Alfredo, R., Echeverria, V., Jin, Y., Yan, L., Swiecki, Z., Gašević, D., & Martinez-Maldonado, R. (2024). Human-centred learning analytics and AI in education: A systematic literature review. Computers and Education: Artificial Intelligence, 6, 100215. https://doi.org/10.1016/j.caeai.2024.100215
4. Amodei, D., & Hernandez, D. (2022). AI and compute: Revisiting the exponential. Anthropic Technical Report.
5. Amodei, D., Olah, C., Steinhardt, J., Christiano, P., Schulman, J., & Mané, D. (2016). Concrete Problems in AI Safety (arXiv:1606.06565). arXiv. https://doi.org/10.48550/arXiv.1606.06565
6. Arrieta, A. B., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., Garcia, S., Gil-Lopez, S., Molina, D., Benjamins, R., Chatila, R., & Herrera, F. (2020). Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. Information Fusion, 58, 82–115. https://doi.org/10.1016/j.inffus.2019.12.012
7. Bai, Y., Kadavath, S., Kundu, S., Askell, A., Kernion, J., Jones, A., Chen, A., Goldie, A., Mirhoseini, A., McKinnon, C., Chen, C., Olsson, C., Olah, C., Hernandez, D., Drain, D., Ganguli, D., Li, D., Tran-Johnson, E., Perez, E., … Kaplan, J. (2022). Constitutional AI: Harmlessness from AI Feedback (Version 1). arXiv. https://doi.org/10.48550/ARXIV.2212.08073

8. Bansal, G., Wu, T., Zhou, J., Fok, R., Nushi, B., Kamar, E., Ribeiro, M. T., & Weld, D. (2021). Does the Whole Exceed its Parts? The Effect of AI Explanations on Complementary Team Performance. Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems, 1–16. https://doi.org/10.1145/3411764.3445717
9. Bommasani, R., Hudson, D. A., Adeli, E., Altman, R., Arora, S., Arx, S. von, Bernstein, M. S., Bohg, J., Bosselut, A., Brunskill, E., Brynjolfsson, E., Buch, S., Card, D., Castellon, R., Chatterji, N., Chen, A., Creel, K., Davis, J. Q., Demszky, D., … Liang, P. (2022). On the Opportunities and Risks of Foundation Models (arXiv:2108.07258). arXiv. https://doi.org/10.48550/arXiv.2108.07258
10. Cabitza, F., Campagner, A., Ferrari, D., & Ciucci, D. (2021). The need to rehumanize clinical AI. Artificial Intelligence in Medicine, 118, 102121. https://doi.org/10.1016/j.artmed.2021.102121
11. Cabrera, Á. A., Perer, A., & Hong, J. I. (2023). Improving Human-AI Collaboration With Descriptions of AI Behavior. Proceedings of the ACM on Human-Computer Interaction, 7(CSCW1), 1–21. https://doi.org/10.1145/3579612
12. Cai, C. J., Winter, S., Steiner, D., Wilcox, L., & Terry, M. (2021). Onboarding Materials as Cross-functional Boundary Objects for Developing AI Assistant. ACM Computing Surveys, (CHI 2021 Japan), 1–11. https://doi.org/10.1145/1122445.1122456
13. Choudhury, S. (2023). Speeding up to fall behind: A critical review of human-machine teaming. Frontiers in Neuroergonomics, 4, 1093982. https://doi.org/10.3389/fnrgo.2023.1093982
14. Daigle, K. & GitHub Staff. (2023, November 8). Octoverse: The state of open source and rise of AI in 2023. The GitHub. https://github.blog/news-insights/research/the-state-of-open-source-and-ai/
15. Dietvorst, B. J., & Bartels, D. M. (2022). Consumers Object to Algorithms Making Morally Relevant Tradeoffs Because of Algorithms' Consequentialist Decision Strategies. Journal of Consumer Psychology, 32(3), 406–424. https://doi.org/10.1002/jcpy.1266
16. Dietvorst, B. J., & Bharti, S. (2020). People Reject Algorithms in Uncertain Decision Domains Because They Have Diminishing Sensitivity to Forecasting Error. Psychological Science, 31(10), 1302–1314. https://doi.org/10.1177/0956797620948841
17. Dietvorst, B. J., Simmons, J. P., & Massey, C. (2015). Algorithm aversion: People erroneously avoid algorithms after seeing them err. Journal of Experimental Psychology: General, 144(1), 114–126. https://doi.org/10.1037/xge0000033
18. Dressel, J., & Farid, H. (2018). The accuracy, fairness, and limits of predicting recidivism. Science Advances, 4(1), eaao5580. https://doi.org/10.1126/sciadv.aao5580
19. Ehsan, U., Liao, Q. V., Muller, M., Riedl, M. O., & Weisz, J. D. (2021). Expanding Explainability: Towards Social Transparency in AI systems. Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems, 1–19. https://doi.org/10.1145/3411764.3445188
20. Endsley, M. R. (2023). Situation awareness in future autonomous systems. Human Factors, 65(1), 99–111. https://doi.org/10.1177/00187208211059252
21. European Union. (2024). The AI Act Explorer: EU Artificial Intelligence Act. Future of Life Institute. https://artificialintelligenceact.eu/ai-act-explorer/
22. Fogliato, R., Chappidi, S., Lungren, M., Fisher, P., Wilson, D., Fitzke, M., Parkinson, M., Horvitz, E., Inkpen, K., & Nushi, B. (2022). Who Goes First? Influences of Human-AI Workflow on Decision Making in Clinical Imaging. 2022 ACM Conference on Fairness Accountability and Transparency, 1362–1374. https://doi.org/10.1145/3531146.3533193
23. Gal, Y., & Ghahramani, Z. (2016). Dropout as a Bayesian Approximation: Representing Model Uncertainty in Deep Learning (Version 6). arXiv. https://doi.org/10.48550/ARXIV.1506.02142
24. Glaser, B. G., & Strauss, A. L. (1967). The discovery of grounded theory: Strategies for qualitative research. Aldine.
25. Gomes, D. (2024). A Comprehensive Study of Advancements in Intelligent Tutoring Systems Through Artificial Intelligent Education Platforms: In F. T. Moreira & R. O. Teles (Eds.), Advances in Educational Technologies and Instructional Design (pp. 213–244). IGI Global. https://doi.org/10.4018/979-8-3693-6170-2.ch008
26. Hadfield-Menell, D., & Hadfield, G. K. (2019). Incomplete Contracting and AI Alignment. Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society, 417–422. https://doi.org/10.1145/3306618.3314250

27. Hemmer, P., Schemmer, M., Vössing, M., & Kühl, N. (2021). Human-AI Complementarity in Hybrid Intelligence Systems: A Structured Literature Review. PACIS 2021 Proceedings, 78. https://aisel.aisnet.org/pacis2021/78
28. Jansen, D. (2021). The International Spillovers of the 2010 U.S. Flash Crash. Journal of Money, Credit and Banking, 53(6), 1573–1586. https://doi.org/10.1111/jmcb.12790
29. Jerry, B., Moreno, L., & Martínez, P. (2026). Human Oversight-by-Design for Accessible Generative IUIs. https://doi.org/10.48550/ARXIV.2602.13745
30. Karinshak, E., Liu, S. X., Park, J. S., & Hancock, J. T. (2023). Working With AI to Persuade: Examining a Large Language Model's Ability to Generate Pro-Vaccination Messages. Proceedings of the ACM on Human-Computer Interaction, 7(CSCW1), 1–29. https://doi.org/10.1145/3579592
31. Kazim, T., & Tomlinson, J. (2023). Automation Bias and the Principles of Judicial Review. Judicial Review, 28(1), 9–16. https://doi.org/10.1080/10854681.2023.2189405
32. Lai, V., Chen, C., Liao, Q. V., Smith-Renner, A., & Tan, C. (2021). Towards a Science of Human-AI Decision Making: A Survey of Empirical Studies (arXiv:2112.11471). arXiv. https://doi.org/10.48550/arXiv.2112.11471
33. Lee, J. D., & See, K. A. (2004). Trust in Automation: Designing for Appropriate Reliance. Human Factors: The Journal of the Human Factors and Ergonomics Society, 46(1), 50–80. https://doi.org/10.1518/hfes.46.1.50.30392
34. Li, Q., & McAdams, D. (2025). Interactive machine learning framework enabling affordable and accurate prototyping for supporting decision-making. Proceedings of the Design Society, 5, 2131–2140. https://doi.org/10.1017/pds.2025.10227
35. Logg, J. M., & Minson, J. A. (2022). Algorithm appreciation and aversion: Why people should use AI more and how to help them do it. Current Directions in Psychological Science, 31(6), 518–524. https://doi.org/10.1177/09637214221117483
36. Logg, J. M., Minson, J. A., & Moore, D. A. (2019). Algorithm appreciation: People prefer algorithmic to human judgment. Organizational Behavior and Human Decision Processes, 151, 90–103. https://doi.org/10.1016/j.obhdp.2018.12.005
37. Madaio, M. A., Stark, L., Wortman Vaughan, J., & Wallach, H. (2020). Co-Designing Checklists to Understand Organizational Challenges and Opportunities around Fairness in AI. Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems, 1–14. https://doi.org/10.1145/3313831.3376445
38. Mosier, K. L., Skitka, L. J., Heers, S., & Burdick, M. (1998). Automation Bias: Decision Making and Performance in High-Tech Cockpits. The International Journal of Aviation Psychology, 8(1), 47–63. https://doi.org/10.1207/s15327108ijap0801_3
39. Munro, R. (with Safari, an O'Reilly Media Company). (2021). Human-in-the-Loop Machine Learning (1st edition). Manning Publications.
40. Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C. L., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., Schulman, J., Hilton, J., Kelton, F., Miller, L., Simens, M., Askell, A., Welinder, P., Christiano, P., Leike, J., & Lowe, R. (2022). Training language models to follow instructions with human feedback (Version 1). arXiv. https://doi.org/10.48550/ARXIV.2203.02155
41. Page, M. J., McKenzie, J. E., Bossuyt, P. M., Boutron, I., Hoffmann, T. C., Mulrow, C. D., Shamseer, L., Tetzlaff, J. M., Akl, E. A., Brennan, S. E., Chou, R., Glanville, J., Grimshaw, J. M., Hróbjartsson, A., Lalu, M. M., Li, T., Loder, E. W., Mayo-Wilson, E., McDonald, S., … Moher, D. (2021). The PRISMA 2020 statement: An updated guideline for reporting systematic reviews. BMJ, n71. https://doi.org/10.1136/bmj.n71
42. Parasuraman, R., & Manzey, D. H. (2010). Complacency and Bias in Human Use of Automation: An Attentional Integration. Human Factors: The Journal of the Human Factors and Ergonomics Society, 52(3), 381–410. https://doi.org/10.1177/0018720810376055
43. Parasuraman, R., Sheridan, T. B., & Wickens, C. D. (2000). A model for types and levels of human interaction with automation. IEEE Transactions on Systems, Man, and Cybernetics - Part A: Systems and Humans, 30(3), 286–297. https://doi.org/10.1109/3468.844354
44. Pawson, R. (2006). Evidence-based policy: A realist perspective. Sage.
45. Rahwan, I. (2021). Society-in-the-loop: Programming the algorithmic social contract. Ethics and Information Technology, 20(1), 5–14. https://doi.org/10.1007/s10676-017-9430-8

46. Reich, T., Kaju, A., & Maglio, S. J. (2023). How to overcome algorithm aversion: Learning from mistakes. Journal of Consumer Psychology, 33(2), 285–302. https://doi.org/10.1002/jcpy.1313
47. Richardson, L. S., Fidock, J., & Gunawan, I. (2025). Systematic Literature Review of Levels of Automation (Autonomy) Taxonomy: Critiques and Recommendations. International Journal of Human–Computer Interaction, 41(24), 15824–15843. https://doi.org/10.1080/10447318.2025.2502978
48. Romeo, G., & Conti, D. (2026). Exploring automation bias in human–AI collaboration: A review and implications for explainable AI. AI & SOCIETY, 41(1), 259–278. https://doi.org/10.1007/s00146-025-02422-7
49. Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. Nature Machine Intelligence, 1(5), 206–215. https://doi.org/10.1038/s42256-019-0048-x
50. Russell, S. J., & Norvig, P. (2021). Artificial Intelligence: A Modern Approach. Pearson.
51. Shani, I. & GitHub Staff. (2023, June 13). Survey reveals AI's impact on the developer experience. The GitHub. https://github.blog/news-insights/research/survey-reveals-ais-impact-on-the-developer-experience/
52. Shneiderman, B. (2020). Human-Centered Artificial Intelligence: Reliable, Safe & Trustworthy. International Journal of Human–Computer Interaction, 36(6), 495–504. https://doi.org/10.1080/10447318.2020.1741118
53. Shneiderman, B. (2022). Human-Centered AI (1st ed.). Oxford University PressOxford. https://doi.org/10.1093/oso/9780192845290.001.0001
54. Shortliffe, E. H., & Sepúlveda, M. J. (2022). Clinical Decision Support in the Era of Artificial Intelligence. JAMA, 320(21), 2199. https://doi.org/10.1001/jama.2018.17163
55. Sloane, M., Moss, E., Awomolo, O., & Forlano, L. (2020). Participation is not a Design Fix for Machine Learning (Version 3). arXiv. https://doi.org/10.48550/ARXIV.2007.02423
56. Tabassi, E. (2023). Artificial Intelligence Risk Management Framework (AI RMF 1.0) (NIST AI 100-1; p. NIST AI 100-1). National Institute of Standards and Technology (U.S.). https://doi.org/10.6028/NIST.AI.100-1
57. Thomas, J., & Harden, A. (2008). Methods for the thematic synthesis of qualitative research in systematic reviews. BMC Medical Research Methodology, 8(1), 45. https://doi.org/10.1186/1471-2288-8-45
58. Topol, E. J. (2019). High-performance medicine: The convergence of human and artificial intelligence. Nature Medicine, 25(1), 44–56. https://doi.org/10.1038/s41591-018-0300-7
59. Trist, E., & Bamforth, K. (1951). Some social and psychological consequences of the longwall method of coal-getting. Human Relations, 4(1), 3–38.
60. Vogl, R. (Ed.). (2021). Research handbook on big data law. Edward Elgar Publishing Limited.
61. Wachter, S., Mittelstadt, B., & Russell, C. (2021). Why fairness cannot be automated: Bridging the gap between EU non-discrimination law and AI. Computer Law & Security Review, 41, 105567. https://doi.org/10.1016/j.clsr.2021.105567
62. Wang, D., Yang, Q., Abdul, A., & Lim, B. Y. (2019). Designing Theory-Driven User-Centric Explainable AI. Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems, CHI '19, 1–15. https://doi.org/10.1145/3290605.3300831
63. Wang, S., Wang, F., Zhu, Z., Wang, J., Tran, T., & Du, Z. (2024). Artificial intelligence in education: A systematic literature review. Expert Systems with Applications, 252, 124167. https://doi.org/10.1016/j.eswa.2024.124167
64. World Health Organization. (2021). Ethics and Governance of Artificial Intelligence for Health: WHO Guidance (1st ed). World Health Organization. https://www.who.int/publications/i/item/9789240029200
65. Zhang, Y., Liao, Q. V., & Bellamy, R. K. E. (2020). Effect of confidence and explanation on accuracy and trust calibration in AI-assisted decision making. Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency, 295–305. https://doi.org/10.1145/3351095.3372852
66. Ziegler, D. M., Stiennon, N., Wu, J., Brown, T. B., Radford, A., Amodei, D., Christiano, P., & Irving, G. (2020). Fine-Tuning Language Models from Human Preferences (arXiv:1909.08593). arXiv. https://doi.org/10.48550/arXiv.1909.08593