Explainable AI Governance in Organizational Decision-Making: Metacognitive Calibration, Algorithmic Accountability, and the Future of Human-AI Collaboration

Hannah Kelly

Abstract

The rapid integration of artificial intelligence into organizational life has created a governance crisis that spans strategic management, human resource analytics, and the fundamental nature of human-machine collaboration. As AI systems assume an increasingly consequential role in decisions that shape careers, resource allocation, and strategic direction, the capacity of these systems to explain their reasoning—and the capacity of organizations to audit and govern them—has become a central challenge for both research and practice. This paper develops a comprehensive framework for explainable AI governance in organizational decision-making contexts, drawing upon four foundational references and eleven additional citations spanning metacognitive calibration, AI auditing theory, large language model governance, algorithmic management, and human-AI collaboration science. The analysis reveals that current AI governance frameworks are undermined by a fundamental tension: the most capable AI systems are also the most opaque, and the most consequential organizational decisions are precisely those for which transparent, auditable AI is most urgently needed. The paper introduces a governance architecture grounded in metacognitive calibration principles, algorithmic accountability mechanisms, and adaptive trust frameworks that together provide a pathway toward responsible AI deployment in organizations. The framework addresses the distinct requirements of strategic business decision support, human resource analytics, and cross-functional AI oversight, while engaging with the resource constraints identified by the AI auditing literature.

Keywords: Explainable AI, AI Governance, Metacognitive Calibration, Algorithmic Management, Organizational Trust, Human-AI Collaboration, LLM Governance, HR Analytics, AI Auditing, EU AI Act

1. Introduction

The deployment of artificial intelligence within organizations has reached an inflection point. What began as experimental applications in isolated functions has evolved into a systematic transformation of how organizations make decisions, supervise workforces, allocate resources, and formulate strategy. As of 2025, 61% of US companies use AI-powered analytics to measure employee productivity or behavior, a statistic that reflects the breadth of algorithmic influence across organizational life. Yet this diffusion of AI into organizational processes has far outpaced the development of governance frameworks capable of ensuring that AI systems operate responsibly, accountably, and in alignment with organizational values and stakeholder interests.

The governance challenge is not merely technical but fundamentally organizational. AI systems in enterprise settings make recommendations that affect hiring and promotion decisions, performance evaluations, strategic investments, resource distributions, and customer interactions. These decisions carry significant consequences for individuals and organizations alike, yet the AI systems making or informing these recommendations are frequently opaque. The challenge of explainability—making the reasoning of AI systems transparent and interpretable to human decision-makers—is therefore not an abstract research question but a pressing practical imperative.

This paper addresses a central research question: how can organizations govern AI systems in ways that enable the beneficial use of AI capabilities while ensuring accountability, transparency, and appropriate human oversight? This question spans multiple dimensions: the technical challenge of generating explanations for complex AI models, the organizational challenge of integrating explainability into decision workflows, the regulatory challenge of compliance with emerging frameworks such as the EU AI Act, and the cognitive challenge of enabling human decision-makers to appropriately calibrate their trust in AI recommendations.

The analysis draws upon four foundational references. Wang (2026) introduced the MIRROR benchmark for metacognitive calibration in large language models, revealing that state-of-the-art AI systems are systematically miscalibrated in ways that undermine trustworthy deployment. Wang (2026) further developed the concept of the Verification Tax—the fundamental resource costs that constrain comprehensive AI auditing in high-stakes settings. Jiang et al. (2026) demonstrated the integration of BERT with simulation algorithms for business decision support, illustrating how interpretable AI architectures can be designed for organizational applications. Bei et al. (2025) presented a framework for strategic human resource analytics with explainable AI, revealing both the potential and the governance challenges of AI in people management.

The paper proceeds as follows. Section 2 provides background on AI in organizational decision-making and the emergence of algorithmic management. Section 3 reviews the landscape of explainable AI in management and governance contexts. Section 4 examines metacognitive calibration and its implications for organizational trust in AI. Section 5 analyzes the Verification Tax and its implications for AI governance design. Section 6 reviews LLM governance in organizational settings. Section 7 examines the application of explainable AI to human resource analytics. Section 8 presents a unified framework for explainable AI governance in organizations. Section 9 discusses regulatory compliance, with particular attention to the EU AI Act. Section 10 concludes with limitations and future research directions.

---

2. Background: AI in Organizational Decision-Making

2.1 From Automation to Algorithmic Management

The integration of AI into organizational life has evolved through distinct phases, from early automation of routine tasks to the contemporary phenomenon of algorithmic management—the use of AI systems to perform managerial functions including task allocation, performance monitoring, and decision-making. Algorithmic management represents a qualitatively different relationship between technology and organization than simple automation. Where automation substitutes machines for human labor in executing predetermined processes, algorithmic management introduces AI systems into the decision-making processes that previously constituted the core of managerial work.

The diffusion of algorithmic management across industry sectors has been documented in recent research, which shows that data-driven supervision has moved from the platform economy into mainstream management, reshaping autonomy, accountability, and everyday decision-making. This diffusion reflects a broader structural shift in which AI systems are not merely tools that human managers use but active agents in organizational governance. The implications of this shift are profound: when AI systems make or substantially influence managerial decisions, the questions of who is accountable for those decisions and how their reasoning can be scrutinized become central concerns for organizational theory and practice.

The literature on algorithmic management identifies several distinctive features of AI-mediated organizational governance. First, AI systems can process and integrate vastly larger volumes of information than human managers, enabling decisions that are more data-informed but potentially less interpretable. Second, AI systems can operate continuously and consistently, applying the same decision criteria across all cases without the variability that characterizes human judgment—but also without the contextual flexibility that experienced managers bring. Third, AI systems can make decisions at speeds and scales that exceed human capacity, creating organizational processes that are more efficient but also more opaque and harder to intervene in.

2.2 AI as a Strategic and Operational Decision Partner

Beyond workforce management, AI systems are increasingly deployed as partners in strategic and operational decision-making. Large language models have proven capable of assisting with many aspects of organizational decision-making, including collecting information from databases, synthesizing relevant evidence, and brainstorming possible courses of action ahead of major decisions. The emergence of LLM-augmented decision support systems represents a significant evolution in how organizations access and process information for decision-making.

Research on LLM-augmented algorithmic management has proposed governance-oriented architectures that integrate LLMs with explainable organizational decision systems. These architectures recognize that while LLMs can bridge the gap between quantitative signals and human-readable explanations, their deployment in governance-critical applications requires careful attention to transparency, accountability, and the management of model limitations. The governance implications are significant: an LLM that provides strategic recommendations without transparent reasoning creates accountability gaps that organizations cannot easily close.

Recent work has critically evaluated the quality of LLM-generated strategic advice, finding that while these systems can produce polished and superficially coherent recommendations, they frequently lack the depth, contextual nuance, and critical judgment that characterize high-quality strategic analysis. This finding has direct implications for organizational governance: the apparent fluency and confidence of LLM outputs can create a veneer of rigor that masks underlying analytical weaknesses, potentially leading decision-makers to over-rely on AI recommendations that are less reliable than they appear.

---

3. Explainable AI in Management and Governance Contexts

3.1 The Case for Interpretability and Explainability

The argument for explainable AI in organizational contexts rests on multiple foundations: ethical accountability, regulatory compliance, practical utility, and organizational learning. Ethically, stakeholders—including employees, customers, investors, and the public—have legitimate interests in understanding how consequential decisions that affect them are reached. When an AI

system influences a hiring decision, a loan approval, or a strategic investment recommendation, the affected parties have reasonable expectations of understanding the basis for that decision.

A recent survey of the case for interpretability and explainability in AI systems argues that interpretability is not a fixed property of an algorithm but a socio-technical practice grounded in context, purpose, and thoughtful design. When supported by the right processes and infrastructure, interpretability becomes a powerful tool for improving accountability, fostering trust, and guiding the responsible adoption of machine learning systems across industries. This framing emphasizes that explainability is not merely a technical requirement but an organizational design challenge that requires attention to workflows, incentive structures, and governance mechanisms alongside algorithmic transparency.

From a regulatory perspective, the EU AI Act represents the most comprehensive attempt to codify explainability requirements into law. The Act categorizes AI systems by risk level and imposes strict compliance obligations on high-risk AI providers and deployers. For organizational decision-makers, the Act's requirements for transparency, human oversight, and risk management create both compliance obligations and a framework for thinking systematically about what explainable AI governance should achieve.

3.2 Explainability Methods and Their Organizational Applications

The technical methods available for generating AI explanations include a broad spectrum ranging from inherently interpretable models to post-hoc explanation techniques applied to complex neural networks. Inherently interpretable models—including decision trees, rule-based systems, and sparse linear models—provide transparent reasoning that is directly readable by human decision-makers. However, these models frequently sacrifice predictive accuracy for interpretability, creating a tension between model fidelity and organizational transparency.

Post-hoc explanation methods, including SHAP (SHapley Additive exPlanations) and LIME (Local Interpretable Model-agnostic Explanations), have emerged as practical tools for making complex model predictions interpretable. SHAP values—grounded in cooperative game theory—quantify the contribution of each input feature to a specific prediction, providing a principled basis for feature-level explanation. Research on SHAP applications in business contexts shows that these tools help explain how models make decisions, increasing transparency and trust, particularly in regulated industries.

The practical application of explainability methods in organizational settings requires attention to how explanations are consumed and used. A meta-analysis of explainable AI methods in clinical decision support systems found that the usability of explanations depends critically on how they are presented, the prior knowledge of the recipient, and the decision context in which they are used. Similar considerations apply in organizational settings: an explanation that is technically accurate but cognitively overwhelming fails to serve the goal of enabling informed human judgment.

---

4. Metacognitive Calibration and Organizational Trust in AI

4.1 The MIRROR Benchmark and Its Organizational Implications

Paper 1, Wang (2026), introduced the MIRROR benchmark for evaluating metacognitive calibration in large language models. Metacognitive calibration—the capacity of an AI system to accurately assess and communicate the boundaries of its own knowledge and competence—represents a critical yet underexplored dimension of AI governance in organizational settings. The MIRROR framework evaluates metacognitive calibration across three dimensions: global calibration, local calibration, and metacognitive sensitivity. Global calibration assesses the overall alignment between model confidence and empirical accuracy. Local calibration evaluates confidence accuracy for specific subsets of instances. Metacognitive sensitivity—the most operationally relevant dimension—assesses the model's ability to distinguish between its own correct and incorrect responses.

The MIRROR findings reveal that even the most capable state-of-the-art LLMs exhibit systematic miscalibration, with a consistent tendency toward overconfidence. This finding has direct and significant implications for organizational AI deployment. An AI system that is overconfident in its incorrect recommendations poses substantial risks in organizational contexts where those recommendations inform consequential decisions. The tendency of humans to increase their reliance on AI recommendations when AI provides high confidence ratings—documented in research on trust calibration—amplifies this risk, creating a potential for systematic errors that pass undetected because human oversight is selectively disengaged precisely when it is most needed.

Recent research on metacognitive sensitivity as the key to calibrating trust in AI decision-making reinforces the organizational importance of these findings. The study demonstrates that knowing when to trust AI advice is critically important in the modern world, and that appropriate trust calibration requires AI systems to accurately signal not just their predictions but their confidence in those predictions. When this metacognitive signal is absent or unreliable—as the MIRROR findings suggest it currently is—human decision-makers are left without the informational foundation needed for appropriate trust calibration.

4.2 Trust Calibration in Human-AI Organizational Collaboration

The challenge of trust calibration in organizational settings is not merely a technical problem but a deeply human one. Research on trust and AI weight in organizational management decision-making reveals that humans and AI form natural teams when collaborating on decisions, and that effective human-AI teams require systems that can leverage the unique capabilities of each while overcoming the limitations of both. The framework of collaborative human-AI trust proposes active management of trust as a dynamic process rather than a static property, recognizing that appropriate reliance on AI changes as contexts, tasks, and the AI system's demonstrated reliability evolve.

A particularly insightful perspective comes from research on the evolutionary mismatch between human cognitive adaptations and AI-mediated environments. Humans evolved to calibrate trust through rich social cues—including facial expressions, vocal tone, and body language—that convey the confidence, competence, and reliability of interaction partners. AI systems lack these cues, creating an evolutionary mismatch that helps explain why trust calibration problems are especially pronounced in human-AI teams. The practical implication is that organizational AI systems must compensate for the absence of natural trust signals by providing explicit, reliable metacognitive communication—confidence estimates, uncertainty flags, and confidence intervals—that enable human users to calibrate their reliance appropriately.

Recent work on calibrating trust in generative AI through human-centered testing frameworks with adaptive explainability provides a methodological template for addressing this challenge. The framework embeds trust calibration directly into AI evaluation, positioning explanation strategies as adaptive levers for improving human-AI collaboration. This approach recognizes that trust calibration is not achieved by providing more information but by providing the right information —information that enables users to form accurate mental models of when the AI system is reliable and when it is not.

---

5. The Verification Tax and the Limits of Organizational AI Auditing

5.1 The Verification Tax: Concept and Organizational Implications

Paper 2, Wang (2026), introduces the concept of the Verification Tax—the observation that auditing AI systems for fairness, accuracy, and safety incurs fundamental costs that cannot be eliminated through improved methodology alone. The Verification Tax encompasses three dimensions that are particularly relevant to organizational AI governance: a statistical dimension (the need for large, representative datasets to evaluate AI behavior in rare-event regimes), a mechanistic dimension (the difficulty of tracing how complex models arrive at their decisions), and an institutional dimension (the need for trained auditors, regulatory infrastructure, and standardized evaluation protocols).

The organizational implications of the Verification Tax are significant and multifaceted. First, the resource intensity of comprehensive AI auditing means that organizations cannot realistically audit every AI decision or even every AI model in their portfolio at the same level of rigor. Organizations must make strategic choices about where to concentrate auditing resources—typically on the highest-stakes applications—while accepting that lower-stakes AI applications may operate with less comprehensive oversight. This creates a tiered governance structure in which the intensity of auditing is calibrated to consequence.

Second, the Verification Tax implies that purely retrospective auditing—evaluating AI systems after they have been deployed—is insufficient for ensuring responsible AI governance. The difficulty of detecting AI failures in complex organizational contexts means that proactive governance mechanisms—including algorithmic impact assessments, pre-deployment testing for bias and calibration, and ongoing monitoring for performance drift—must complement retrospective auditing.

Third, the institutional dimension of the Verification Tax highlights that organizations face a governance capacity challenge. The supply of trained AI auditors, ethicists, and governance specialists has not kept pace with the demand created by the rapid proliferation of AI applications. Organizations must therefore invest in building internal AI governance capacity while also developing practical tools that enable non-specialist managers to exercise meaningful oversight of AI systems.

5.2 Governance Design Under Verification Constraints

The Verification Tax analysis suggests several design principles for organizational AI governance. First, governance architectures should prioritize inherently interpretable models for the highest-stakes decisions, reserving complex neural networks for applications where their predictive advantages are essential and where the cost of reduced interpretability is acceptable. Second, organizations should invest in monitoring systems that can detect AI performance degradation

and distributional shift in real time, enabling targeted auditing rather than comprehensive but impractical constant oversight. Third, governance frameworks should be designed with verification efficiency in mind, employing risk-based approaches that concentrate rigorous evaluation on the applications where failures would be most consequential.

The engineering of explainable AI systems for GDPR-aligned decision transparency provides a modular framework for continuous compliance that partially addresses the Verification Tax by embedding governance requirements into the system design process rather than treating them as add-on evaluations. The modular approach enables incremental compliance verification rather than requiring comprehensive system-level audits, reducing the Verification Tax by making auditing more targeted and efficient.

---

6. LLM Governance in Organizational Settings

6.1 The Emergence of LLM Governance

The deployment of large language models in organizational settings has created new governance challenges that extend beyond those posed by traditional machine learning models. LLMs are characterized by emergent capabilities that are difficult to predict from training data alone, by outputs that are fluent and confident but potentially unreliable, and by decision influence that can be diffuse and difficult to attribute to specific organizational actors.

Research on LLM governance identifies the essential principle that organizations cannot govern what they cannot see. Transparency requires documenting model training data sources, intended use cases, decision pathways, and limitations—and making model behavior interpretable to both technical and non-technical stakeholders. This principle has operational implications for how organizations procure, deploy, and monitor LLM-based systems.

A comprehensive review of large language models for intelligent decision support in inventory and supply chain operations finds that LLMs are predominantly positioned as orchestration and explanatory layers operating alongside structured components such as optimization solvers, simulation engines, and digital twins, rather than as autonomous decision-makers. Governance and organizational readiness are identified as critical success factors, suggesting that the organizational challenges of LLM deployment are at least as significant as the technical challenges.

The concept of enterprise LLM governance as a set of controls determining how LLMs are allowed to behave inside an organization emphasizes that governance is not merely about controlling AI but about creating the organizational structures and processes within which AI can be safely and productively used. This framing aligns with the broader literature on algorithmic management, which recognizes that AI governance is fundamentally an organizational design challenge.

6.2 Risks and Failure Modes of LLMs in Organizations

The risks associated with LLM deployment in organizational settings are multiple and interconnected. Hallucination—the generation of plausible but incorrect or fabricated outputs—represents a fundamental reliability challenge for organizational LLM applications. When LLMs produce authoritative-sounding recommendations that are factually incorrect or analytically flawed, the apparent confidence of the output can mask the underlying unreliability.

Research on the fidelity of LLM-generated strategic advice found that LLMs tend to produce what has been termed "trendslop"—superficially plausible but analytically shallow recommendations that follow prevailing narratives without critical evaluation. This failure mode is particularly concerning in strategic decision contexts, where the cost of poorly considered recommendations

can be substantial.

The calibration challenge identified by the MIRROR benchmark is especially acute for LLMs, which are trained to produce fluent, confident outputs regardless of the reliability of the underlying knowledge. An LLM that is uncertain about a factual matter will nonetheless produce a confident, articulate response—precisely the behavior pattern that the MIRROR framework identifies as a metacognitive calibration failure. In organizational settings, this behavior can lead to the inappropriate confidence in AI-generated strategic recommendations that the literature on AI decision support has documented.

---

7. Explainable AI in Human Resource Analytics

7.1 Strategic HR Analytics and the XAI Imperative

Paper 4, Bei et al. (2025), presents a framework for strategic human resource analytics with explainable AI, demonstrating how AI-driven analytics can be integrated with human decision-making in ways that balance analytical power with interpretive transparency. The application of explainable AI to HR analytics is particularly instructive because HR decisions are among the most consequential for individuals—affecting employment, compensation, career development, and professional trajectory—and because HR contexts present distinctive challenges for algorithmic fairness and accountability.

The integration of explainable AI into HR analytics addresses several governance imperatives. First, explainable AI enables HR professionals to understand and verify the reasoning behind algorithmic recommendations, supporting informed human judgment rather than uncritical automation. Second, explainability provides an audit trail that enables organizations to demonstrate the basis for HR decisions to affected employees, regulatory bodies, and courts. Third, explainability enables the identification and correction of algorithmic bias—an increasingly important governance requirement as HR analytics systems are deployed across diverse workforces.

The governance framework for strategic HR analytics with explainable AI recognizes that the value of AI in HR contexts depends not just on predictive accuracy but on the capacity of AI systems to support human decision-makers in ways that are consistent with organizational values, legal requirements, and employee expectations of fairness.

7.2 Algorithmic Management and Workforce Governance

Beyond strategic analytics, the broader phenomenon of algorithmic management raises governance challenges that extend across the employment relationship. Research on algorithmic management and the future of human work shows that AI-driven supervision has moved from the platform economy into mainstream management, reshaping autonomy, accountability, and everyday decision-making in ways that affect workers across industries and occupational categories.

The governance of algorithmic management requires attention to multiple dimensions. Transparency requirements—increasingly mandated by regulation and normative expectation—call for disclosure of how AI systems are used in workforce management. Accountability mechanisms require clarity about who is responsible when algorithmic decisions cause harm. Fairness considerations require that AI systems do not systematically disadvantage protected groups through biased training data or algorithmic design.

The EU AI Act addresses algorithmic management directly, classifying AI systems used in employment and worker management as high-risk applications subject to stringent compliance requirements. These requirements include data governance provisions ensuring that training datasets are relevant, sufficiently representative, and free from bias—as well as human oversight provisions requiring that humans remain meaningfully in the loop for consequential employment decisions.

Recent research on fairness and trust in AI decision-making emphasizes the critical role of human involvement and outcome favorability in shaping employee perceptions of AI fairness. The study finds that real-time transparency and visible evidence of human involvement in decision-making are key determinants of whether employees perceive AI-driven HR decisions as fair and trustworthy. These findings suggest that organizational governance frameworks for HR AI must attend not just to technical fairness but to the experiential dimensions of how AI-driven decisions are perceived and received by employees.

---

8. A Framework for Explainable AI Governance in Organizations

8.1 The GReAT Governance Architecture

Synthesizing the perspectives developed across the preceding sections, this paper proposes the Governance of Responsible AI in Organizations (GReAT) framework—a comprehensive architecture for explainable AI governance that integrates metacognitive calibration principles, algorithmic accountability mechanisms, adaptive trust frameworks, and regulatory compliance processes.

The GReAT framework comprises four interconnected governance layers, each addressing a distinct dimension of organizational AI responsibility.

Layer 1: Metacognitive Calibration and Uncertainty Communication. The foundational layer addresses the challenge identified by the MIRROR benchmark: that AI systems must be capable of accurately assessing and communicating their own knowledge boundaries. This layer mandates the integration of metacognitive calibration assessment into the AI procurement and deployment process, requiring that organizations evaluate not just the predictive accuracy of candidate AI systems but also their calibration properties. Systems that exhibit significant miscalibration should not be deployed in consequential decision-making contexts without calibration remediation or human oversight safeguards. The layer also requires that deployed AI systems communicate uncertainty to users through calibrated confidence estimates, explicit abstention on uncertain predictions, and accessible uncertainty visualization.

Layer 2: Algorithmic Accountability and Audit Trail Management. The second layer addresses the Verification Tax challenge by implementing a tiered accountability architecture that concentrates intensive auditing on the highest-stakes applications while maintaining proportionate oversight of lower-stakes systems. This layer requires that all AI systems maintain comprehensive audit trails—documenting inputs, outputs, decision context, and human oversight actions—that can support both internal quality assurance and external regulatory review. The layer incorporates the modular compliance engineering approach, embedding accountability requirements into AI system design rather than treating them as post-hoc evaluations.

Layer 3: Organizational Trust Architecture and Human-AI Collaboration Design. The third layer addresses the organizational challenge of enabling appropriate trust calibration in human-AI teams. Drawing on the collaborative human-AI trust literature and the adaptive explainability framework, this layer mandates that organizations design AI interfaces and workflows that

support active trust management—enabling human users to understand when and why AI recommendations are reliable, and to adjust their reliance on AI accordingly. The layer also requires organizational processes for training employees in human-AI collaboration and for monitoring the effectiveness of trust calibration in practice.

Layer 4: Regulatory Compliance and Continuous Governance. The fourth layer addresses the regulatory dimension of AI governance, with particular attention to the EU AI Act's risk-based framework. This layer requires organizations to classify AI applications by risk level, implement appropriate compliance measures for each tier, and establish ongoing monitoring and reporting processes that demonstrate continued compliance. The layer recognizes that the regulatory landscape is evolving and requires governance architectures that are adaptable to new requirements.

8.2 Application to Strategic Decision Support

For strategic business decision support—the domain addressed by Paper 3, Jiang et al. (2026)—the GReAT framework operates as follows. AI systems used for strategic analysis must be evaluated for metacognitive calibration before deployment, and must provide explicit uncertainty estimates alongside recommendations. Audit trails must document the data sources, models, and reasoning steps that inform strategic recommendations. Interface design must enable decision-makers to understand the basis for AI recommendations and to adjust their reliance on AI based on contextual factors including the novelty and complexity of the decision. Governance processes must ensure compliance with applicable regulatory requirements while enabling the agility needed for effective strategic decision-making.

8.3 Application to HR Analytics and Workforce Management

For human resource analytics—the domain addressed by Paper 4, Bei et al. (2025)—the GReAT framework applies additional HR-specific governance requirements. AI systems used in hiring, performance evaluation, compensation, and promotion decisions must meet the highest calibration and explainability standards, reflecting the significant consequences of these decisions for individuals and the legal requirements applicable to employment-related AI. The framework mandates human involvement in all consequential HR decisions, with AI serving as a recommendation engine rather than an autonomous decision-maker. Employee-facing transparency requirements ensure that individuals affected by AI-driven HR decisions can understand the basis for those decisions and contest them if necessary.

---

9. Regulatory Compliance: The EU AI Act and Organizational Accountability

9.1 The EU AI Act's Governance Requirements

The EU AI Act represents the most comprehensive regulatory framework for AI governance currently in force, with enforceable obligations for general-purpose AI models taking effect in August 2025. The Act's risk-based approach categorizes AI systems into four tiers—unacceptable risk, high risk, limited risk, and minimal risk—with progressively more stringent requirements as risk increases. AI systems used in employment and worker management, including algorithmic management and HR analytics applications, are classified as high-risk, subject to the most stringent compliance requirements.

For organizational AI governance, the Act creates several specific obligations. High-risk AI providers must implement comprehensive data governance systems ensuring that training data is relevant, representative, and free from bias. They must maintain technical documentation sufficient to enable post-market surveillance and regulatory auditing. They must implement human oversight measures ensuring that AI systems are subject to meaningful human control. And they must register high-risk AI systems in a publicly accessible EU database before deployment.

The Act also introduces specific transparency requirements for AI systems that interact directly with natural persons, including obligations to inform users that they are interacting with an AI system. These transparency requirements are relevant for organizational AI deployments including chatbots, automated customer service systems, and AI-powered management tools.

9.2 Organizational Preparedness and Compliance Pathways

Preparing for EU AI Act compliance requires organizations to take a comprehensive inventory of their AI applications, classify these applications by risk level, assess their current compliance status, and implement remediation measures where necessary. The concept of coordination transparency—governing distributed agency in AI systems—provides a useful framework for thinking about how organizational governance structures must evolve to manage AI as a distributed organizational capability rather than a set of discrete tool deployments.

Research on coordination transparency argues that governance must move from explaining outputs after the fact to steering coordination in real time. This represents a fundamental shift in the temporal orientation of AI governance: from retrospective accountability to proactive oversight. For organizations, this shift implies that governance structures must be integrated into the AI deployment process from the outset, rather than applied as an afterthought after AI systems are already in operation.

The latest wave of EU AI Act obligations, which took effect in 2025, requires organizations to prepare documentation for transparency compliance, establish internal compliance structures aligned with the AI Act's governance framework, and monitor the activities of the AI Office and national authorities as regulatory guidance evolves. Organizations that have not already begun this process face significant compliance risk, particularly for high-risk AI applications in employment and worker management.

---

10. Limitations and Future Research Directions

This paper is subject to several limitations that should be acknowledged. First, the GReAT framework proposed in this paper is conceptual: its practical implementation and empirical validation in organizational settings remain future work. The framework synthesizes insights from multiple research communities—AI technical research, organizational behavior, law, and regulation—and the operational translation of these insights requires further collaborative development. Second, the paper's analysis of regulatory compliance focuses primarily on the EU AI Act, which, while influential, is not the only regulatory framework affecting organizational AI deployment. Comparative analysis of regulatory approaches across jurisdictions would strengthen the governance analysis. Third, the paper's treatment of organizational trust and human-AI collaboration draws on research that is still evolving, and the practical guidance offered in this area should be understood as provisional pending further empirical validation. Fourth, the metacognitive calibration findings of the MIRROR benchmark, while significant, are based on current-generation AI models that will evolve rapidly; governance frameworks must be designed to adapt to changing model capabilities rather than assuming static technical properties.

Several priority areas for future research emerge from this analysis. First, empirical studies of organizational AI governance in practice are needed to validate the GReAT framework and identify practical barriers to its implementation. Second, the development of practical metacognitive calibration tools for organizational use—metrics, dashboards, and decision aids that can be used by non-specialist managers—represents an important applied research direction. Third, longitudinal studies of trust calibration in human-AI organizational teams would provide empirical foundations for the adaptive trust frameworks proposed in this paper. Fourth, comparative regulatory research examining how different jurisdictions approach the governance of organizational AI would provide an evidence base for regulatory design. Fifth, the intersection of algorithmic management and employee well-being—with attention to the psychological and social impacts of AI-mediated management—deserves deeper investigation than the current literature provides.

11. Conclusion

This paper has argued that explainable AI governance in organizational settings requires a multidimensional framework that integrates technical explainability, metacognitive calibration, algorithmic accountability, and adaptive trust architecture. The four foundational references have provided essential anchor points throughout this analysis. The MIRROR benchmark findings demonstrate that organizational trust in AI cannot be taken for granted and must be actively engineered through calibration-aware system design. The Verification Tax analysis reveals that comprehensive AI auditing faces fundamental resource constraints that require governance architectures to be designed for verification efficiency alongside analytical power. The BERT-ISAC integration demonstrates that interpretable AI architectures for business decision support are both achievable and necessary for responsible deployment. The strategic HR analytics framework illustrates how explainability principles can be applied to one of the most consequential domains of organizational AI application.

The practical implication is that organizations must invest in AI governance as a core organizational capability rather than a compliance exercise. This requires building internal governance capacity, selecting and deploying AI systems with governance requirements in mind, designing human-AI collaboration workflows that support appropriate trust calibration, and establishing ongoing monitoring and accountability processes that can adapt to evolving regulatory requirements and technological capabilities. The alternative—treating AI governance as an afterthought—is not viable in an organizational environment where AI decisions increasingly shape consequential outcomes for individuals, communities, and the organization itself.

References

1. Wang, J. Z. (2026). MIRROR: A Hierarchical Benchmark for Metacognitive Calibration in Large Language Models. arXiv preprint arXiv:2604.19809.
2. Gu, A., & Dao, T. (2023). Mamba: Linear-Time Sequence Modeling with Selective State Spaces. arXiv preprint arXiv:2312.00752.
3. Wang, J. Z. (2026). The Verification Tax: Fundamental Limits of AI Auditing in the Rare-Error Regime. arXiv preprint arXiv:2604.12951.
4. Jiang, A., Huang, J., & Yun, X. (2026). Design and empirical research of simulation algorithms for business decision support with BERT and ISAC integration. In International Conference on Cloud Computing, Performance Computing, and Deep Learning.
5. Bei, J., Liu, Z., Huang, J., Wang, X., & Yang, P. (2025). Strategic Human Resource Analytics with Explainable Artificial Intelligence. In Proceedings of the 2025 6th International Conference on

Computer Science and Management Technology.
6. Lee, M. H., et al. (2025). Metacognitive sensitivity: The key to calibrating trust and optimal decision making with AI. PMC, National Institutes of Health.
7. Arrieta, A. B., et al. (2025). Transparent AI: The case for interpretability and explainability. arXiv preprint arXiv:2507.23535.
8. Tredence. (2026). What Is LLM Governance? Managing Large Language Models Responsibly. Tredence Inc.
9. Frontiers in Organizational Psychology. (2025). Trust and AI weight: Human-AI collaboration in organizational management decision-making. Frontiers in Psychology, 12, 1419403.
10. Schein, C., et al. (2025). Collaborative human-AI trust (CHAI-T): A process framework for active management of trust in human-AI collaboration. ScienceDirect.
11. Jarrahi, M. H., et al. (2021). Algorithmic management in a work context. Big Data & Society, 8(2), 1-15.
12. European Union. (2024). The EU Artificial Intelligence Act. Official Journal of the European Union.
13. ModelOp. (2025). EU AI Act: Summary and Compliance Requirements. ModelOp Inc.
14. Settanni, G., et al. (2025). Applying artificial intelligence to clinical decision support in mental health: What have we learned? Journal of Medical Internet Research, 26, e53089.
15. Alation. (2025). Explainable AI Governance: Frameworks for Trust, Transparency and Compliance. Alation Inc.