Artificial Intelligence as Organizational Agent: Legal Standing, Accountability Gaps, and the Governance of Algorithmic Decision-Making Authority

Tang Bao

Abstract

The deployment of artificial intelligence systems in consequential organizational decisions has created a fundamental legal and organizational challenge: the accountability gap. When an AI system makes or materially influences a decision that causes harm—an discriminatory hiring outcome, a negligent medical recommendation, a faulty financial assessment—the existing frameworks of legal responsibility and organizational accountability provide no clear answer to the question of who is responsible. This paper develops a comprehensive analysis of AI as an organizational agent, examining the legal, organizational, and governance dimensions of the accountability gap and the institutional innovations required to address it. Drawing on three foundational references and twelve supplementary citations spanning legal theory, AI governance, organizational law, and regulatory design, this study argues that the accountability gap is not merely a legal technicality but a structural feature of how AI systems function as de facto agents within organizational hierarchies. The analysis reveals that the metacognitive miscalibration documented by the MIRROR benchmark, the structural constraints on AI auditing identified by the Verification Tax, and the practical demand for explainability in human resource analytics illuminate different dimensions of the same underlying phenomenon: AI systems exercise genuine agency—the capacity to make consequential decisions—without the accountability structures that the exercise of agency has historically required. The paper concludes by proposing an Algorithmic Agency Governance Framework that addresses the accountability gap through a combination of liability reform, organizational governance requirements, and institutional innovations in how AI agency is constituted and overseen.

Keywords: AI Legal Agency, Accountability Gap, Algorithmic Accountability, AI Liability, Organizational Law, Principal-Agent Theory, EU AI Act, Legal Personhood, AI Governance, Organizational Agency

1. Introduction

The concept of agency—the capacity to act, to make choices, and to thereby affect the world—is foundational to both legal and organizational systems. Legal systems allocate responsibility through agency: a principal-agent relationship creates mutual obligations, and harm caused by agent actions within the scope of agency creates liability for the principal. Organizational systems delegate authority through agency: managers act as agents of the organization, and their decisions bind the organization and create organizational responsibility. Agency is the mechanism through which consequential action is legitimated and accountability is ensured.

The deployment of artificial intelligence systems has introduced a profound challenge to the agency framework. AI systems make choices—they select among alternatives based on learned criteria, they allocate resources, they evaluate individuals, and they recommend actions. In doing so, they exercise the functional capacity that legal and organizational systems have always associated with agency. Yet the existing legal and organizational frameworks for agency were developed to govern the relationships between human beings—to specify when one person's choices bind another and when one person's actions create obligations for another. These frameworks provide no clear account of when or how AI systems should be understood as agents, when the decisions AI systems make should create legal or organizational responsibility, or how accountability for AI decisions should be allocated among the parties involved in AI development, deployment, and use.

The resulting accountability gap—the absence of clear legal and organizational structures for assigning responsibility for AI-influenced decisions—has become a pressing practical concern. Organizations deploying AI systems face uncertainty about their legal exposure when AI decisions cause harm. Individuals affected by AI decisions have no clear legal recourse when they suffer discrimination, negligence, or other harms algorithmic in origin. Regulators struggle to design compliance frameworks when the traditional tools of regulatory enforcement—liability rules, mandatory disclosure, licensing requirements—assume a human agent whose intentions, knowledge, and capabilities can be determined and sanctioned.

This paper develops the central argument that the accountability gap is not a problem that can be solved by forcing existing legal and organizational frameworks to accommodate AI systems as if they were human agents. Rather, it is a structural consequence of AI's functional agency—the reality that AI systems exercise genuine decision-making capacity without the characteristics (intentionality, knowledge, moral agency) that the existing frameworks were designed to govern. Addressing the accountability gap requires not merely adapting existing law but reconceptualizing how agency, responsibility, and accountability are understood and allocated in an age of algorithmic decision-making.

This analysis is grounded in three foundational references. Paper 1, Wang (2026), introduced the MIRROR benchmark for metacognitive calibration, revealing that AI systems systematically misrepresent their own reliability. This finding acquires new significance in the accountability context: a responsible party is one who knows what they know and do not know; an AI system that is systematically miscalibrated cannot satisfy the knowledge conditions for moral accountability. Paper 2, Wang (2026), analyzed the Verification Tax—the resource constraints on comprehensive AI auditing—revealing that the accountability gap is partly structural: even well-intentioned organizations face fundamental limits on their capacity to monitor and verify the decisions their AI systems make. Paper 3, Bei et al. (2025), demonstrated the practical demand for explainable AI in strategic human resource analytics, illustrating how the absence of clear accountability for AI decisions creates organizational risks that existing governance frameworks cannot adequately address.

The paper proceeds as follows. Section 2 develops the theoretical foundations of AI agency, examining the legal and organizational concepts of agency and their applicability to AI systems. Section 3 analyzes the accountability gap in organizational AI contexts, identifying its structural sources and practical manifestations. Section 4 examines the liability landscape and regulatory frameworks, with particular attention to the EU AI Act's approach to accountability. Section 5 integrates the MIRROR and Verification Tax findings into the accountability analysis. Section 6 presents the Algorithmic Agency Governance Framework as a comprehensive approach to addressing the accountability gap. Section 7 concludes with implications for legal theory, organizational practice, and regulatory design.

2. Theoretical Foundations: Agency, Legal Standing, and AI

2.1 The Legal Concept of Agency and Its Conditions

Agency in legal theory refers to the relationship in which one party (the principal) authorizes another party (the agent) to act on their behalf, creating legal effects that bind the principal to third parties. The classical account of agency identifies several conditions that must be satisfied for agency to exist. The agent must have the capacity to form intentions and to act on those intentions. The agent must have knowledge of the material circumstances surrounding the action. The agent must act within the scope of the principal's authorization. The agent's actions must create effects—obligations, liabilities, rights—that the legal system recognizes as attaching to the principal.

The philosophical literature on the conditions of moral responsibility identifies similar criteria. A responsible agent is one who can understand the reasons for and against alternative courses of action, who can form intentions based on that understanding, and who can act on those intentions in ways that reflect their judgment about what is right or best. These capacities—understanding, judgment, intentional action—constitute the foundations of the moral accountability that legal agency presupposes.

The question of whether AI systems can satisfy these agency conditions has generated substantial scholarly debate. AI systems can form outputs—predictions, recommendations, decisions—that produce effects in the world. They can process information at scales and speeds that exceed human capacity. They can select among alternatives based on learned criteria that reflect the objectives they have been given. In these functional respects, AI systems appear to exercise agency. Yet AI systems lack the intentionality, understanding, and moral agency that the classical account of responsibility presupposes. An AI system does not "intend" to cause harm in the sense that a human agent intends; it does not "understand" the reasons for and against alternative actions in the sense that a human agent understands.

2.2 De Facto Versus De Jure Agency

The philosophical and legal debate about AI agency has increasingly converged on a distinction between de jure agency—the formal legal status of agency, with its associated rights and obligations—and de facto agency—the functional exercise of agency in producing consequential outcomes. AI systems, on this account, exercise de facto agency regardless of whether they possess de jure agency status. They make choices that matter. They affect the interests of third parties. They shape organizational decisions in ways that create consequences the organization did not directly intend. The question is not whether AI exercises agency—the evidence of consequential AI decisions in every domain of organizational life makes this undeniable—but how the accountability structures that agency traditionally requires should be designed for entities whose agency is computational rather than intentional.

Research on the legal status of AI and robotic technologies traces the evolution of legal thinking about machine agency from the medieval debates about whether animals could be held responsible for harm they caused—debates that ultimately concluded that responsibility attaches to the human keeper rather than the animal—to contemporary proposals for electronic agents with limited legal personhood. The historical parallel is instructive: every previous expansion of the circle of entities capable of exercising agency—beyond adult human males to women, beyond humans to corporations—has required legal innovation to extend accountability structures to new categories of agent. The extension to AI systems represents the next such expansion, and the accountability gap is the visible symptom of the legal system's current failure to accomplish it.

2.3 Organizational Agency and Hierarchical Accountability

Beyond legal theory, the concept of agency is central to organizational theory. Organizations function through hierarchical agency relationships: managers act as agents of the organization's owners, employees act as agents of managers, and the decisions made at each level of the hierarchy create organizational consequences that bind the organization as a whole. The accountability of organizational agents is ensured through incentive structures, monitoring mechanisms, and the residual liability of principals for agent actions within the scope of authority.

When AI systems are integrated into organizational hierarchies—as decision support tools, automated executors, or strategic advisors—they occupy an ambiguous position in the organizational agency structure. AI systems make decisions that shape organizational outcomes, yet they are not organizational members with the rights and obligations that organizational agency normally entails. They are not employees subject to organizational authority. They are not managers with organizational power and responsibility. They are tools that the organization uses—but their decisions are not merely the organization's decisions in the way that a manager's decisions are. The accountability gap in organizational AI is partly a reflection of this structural ambiguity: AI systems are not quite agents, not quite tools, and the organizational frameworks designed to govern each category fail to adequately govern the space in between.

---

3. The Accountability Gap: Sources and Manifestations

3.1 Structural Sources of the Accountability Gap

The accountability gap in AI decision-making arises from three structural sources that interact to prevent the allocation of responsibility to any clear party. The first source is the multiplicity of actors involved in AI decision-making. An AI system's decision is the product of multiple contributions: the organization that deploys it, the developer that builds it, the data providers that supply its training and operational data, and the AI system itself that processes inputs to produce outputs. Each of these actors plays a causal role in the outcome, yet none of them fully determines it. This distributed causality makes it difficult to allocate the exclusive or primary responsibility that traditional liability rules normally require.

The second source is the opacity of AI reasoning. When a human agent makes a consequential decision, the reasoning process that produced it is—at least in principle—accessible: through testimony, documentation, or introspection, the organization and affected parties can determine what the agent knew, what they considered, and why they chose as they did. This reasoning transparency is the foundation of the accountability that agency creates. When an AI system makes a consequential decision through a deep neural network with millions of parameters, this reasoning transparency is absent. The AI system's decision is knowable; why it produced that particular decision is typically not knowable with the precision that accountability requires.

The third source is the variability of AI behavior over time and context. Human agents can be assessed for competence, can be trained and sanctioned, and can be held to standards that evolve with changing conditions. AI systems, once deployed, may change their behavior as they encounter new inputs, as their parameters are updated, or as the data distributions they were trained on shift. An AI system that was reliable and accurate at the time of deployment may become unreliable as conditions change, without the organization having clear awareness or responsibility for the change.

3.2 Manifestations in Organizational Contexts

The accountability gap manifests across organizational domains in ways that create practical challenges for governance, compliance, and harm remediation. In employment contexts—the domain examined by Paper 3—the accountability gap means that when an AI hiring system produces discriminatory outcomes, the legal and organizational frameworks for addressing employment discrimination provide no clear path for determining who is responsible and what remedies are appropriate. Is it the organization that deployed the system without understanding its biases? The developer that built the system using methodologies that failed to detect those biases? The data providers whose historical data encoded the biases that the AI learned? Or the AI system itself, in some formal sense of agency that has not yet been legally constituted?

In financial services, the accountability gap means that when an AI credit scoring system produces racially disparate outcomes, the Fair Lending Act's framework for intentional discrimination and disparate impact analysis provides no clear account of how to allocate responsibility among the multiple actors whose decisions collectively produced the outcome. In healthcare, when an AI diagnostic system recommends a treatment that causes patient harm, the medical malpractice framework's requirement of physician negligence provides no clear standard for evaluating whether the AI-assisted recommendation constitutes negligent practice.

Research on the legal challenges of AI accountability traces the emergence of the accountability gap across these domains, documenting how existing legal frameworks—the negligence standard, the products liability regime, the anti-discrimination statutes—were designed around assumptions about human agency and human decision-making that AI systems violate. The legal scholarship increasingly recognizes that addressing the accountability gap requires not merely adapting existing doctrines but developing new legal categories specifically designed for AI agency.

3.3 The Organizational Governance Dimension

Beyond legal liability, the accountability gap has significant organizational governance dimensions. Organizations that deploy AI systems face governance challenges in the absence of clear accountability: how should boards oversee AI decisions whose reasoning they cannot understand? How should compensation committees evaluate the performance of executives whose decisions were AI-influenced? How should audit committees verify the integrity of AI-assisted financial reporting?

The organizational governance literature has begun to grapple with these questions, recognizing that existing governance frameworks—designed for human agents whose intentions, knowledge, and reasoning can in principle be determined—provide inadequate guidance for the oversight of AI decisions. Board fiduciary duties, which require directors to act with the care and loyalty that a reasonable person would exercise, become difficult to apply when the decisions in question were made by AI systems whose reasoning cannot be independently verified.

Recent research on AI governance in organizations identifies the absence of clear accountability structures as a primary driver of organizational risk in AI deployment. Organizations that deploy AI systems without clear accountability frameworks face not only legal exposure but also operational risks—decisions made without accountability are decisions that cannot be explained, defended, or corrected when they go wrong.

---

4. The Regulatory and Liability Landscape

4.1 The EU AI Act's Approach to Accountability

The European Union's AI Act, which entered into force in 2024 with phased implementation through 2027, represents the most comprehensive attempt to establish an accountability framework for AI systems. The Act adopts a risk-based approach that allocates regulatory obligations according to the risk level of the AI application, with high-risk AI systems—those used in employment, credit, education, law enforcement, and other consequential domains—subject to the most stringent requirements.

For high-risk AI systems, the Act imposes several accountability-related obligations. Providers must implement quality management systems that include post-market monitoring processes to detect and address AI failures. Deployers must ensure that high-risk AI systems are used in accordance with their intended purpose and that human oversight measures are in place. The Act establishes a traceable accountability chain from providers through deployers to specific AI deployments, creating the institutional infrastructure for determining who is responsible when AI systems cause harm.

However, the EU AI Act's approach to accountability has significant limitations. The Act clarifies who bears regulatory obligations but does not resolve the fundamental liability questions that the accountability gap creates. When an AI system causes harm, the Act's compliance requirements do not automatically determine who is legally liable, what the basis for liability is, or what remedies are available to affected parties. These questions remain governed by existing liability frameworks—the Product Liability Directive, the Machinery Directive, general tort law—that were not designed for AI contexts and that provide uncertain answers to the questions AI deployment raises.

4.2 Product Liability and the AI Challenge

The existing product liability framework, developed primarily for manufacturing contexts, presents particular challenges for AI accountability. Under the Product Liability Directive as currently interpreted, a product is defective when it does not provide the safety that a person is entitled to expect, given the product's intended use and the circumstances of its supply. A product's safety is assessed at the time of supply, and liability attaches based on the defect's existence at that time.

For AI systems, these assumptions create several accountability problems. AI systems can evolve in their behavior after supply, through updates, retraining, or distributional shift, making the "time of supply" assessment inadequate for determining defect. AI systems can be used in contexts that their developers did not anticipate or intend, raising questions about the scope of "intended use" as a liability standard. The complexity of AI decision-making makes it difficult to determine whether a defect existed at the time of supply when the harm arises from a decision process that is opaque even to the system's developers.

The proposed EU AI Liability Directive addresses some of these challenges through specific rebuttable presumptions for AI-related harm—presuming that the deployer's failure to comply with specified requirements caused the harm, that the AI system caused the harm when certain conditions are met. Yet the Directive leaves fundamental questions unresolved, including whether AI systems themselves should bear any responsibility, how liability should be allocated among multiple actors in the AI supply chain, and how harm that arises from AI decisions that were appropriate at the time of deployment but became harmful through subsequent distributional shift should be addressed.

4.3 Organizational Liability and the Scope of Authority

The organizational liability literature has begun to grapple with how existing principles of organizational responsibility apply to AI-assisted decisions. The traditional doctrine of respondent superior holds employers liable for employee torts committed within the scope of employment. The question of whether AI system decisions are within the "scope of employment" for an employee who used the AI system—but did not themselves make the consequential choice—is a matter of ongoing legal development.

The concept of apparent authority—the doctrine that an agent's authority can be established by the principal's manifestations to third parties, regardless of the agent's actual authority—provides an alternative avenue for organizational liability. Organizations that deploy AI systems in ways that manifest to affected parties that the AI is acting with organizational authority may be bound by the AI's decisions regardless of the contractual or technical terms that purport to limit that authority.

Research on AI liability and the organization of AI development highlights the importance of the allocation of authority among AI developers, deployers, and users in determining liability. When an AI developer retains substantial control over the AI system's behavior—through ongoing updates, access to system parameters, or decision rules that constrain deployer modification—the developer's potential liability for AI decisions increases. When a deployer substantially modifies an AI system or uses it in ways not anticipated by the developer, liability may shift to the deployer. These allocations are matters of ongoing legal development, with courts and regulators progressively building the precedents and rules through which AI liability will ultimately be determined.

---

5. The MIRROR and Verification Tax Dimensions of Accountability

5.1 Miscalibration and the Epistemics of Responsibility

Paper 1's MIRROR benchmark findings acquire new significance in the accountability context. The benchmark reveals that AI systems systematically misrepresent their own reliability—their confidence estimates do not accurately reflect their actual accuracy. From an accountability perspective, this miscalibration is consequential because responsibility requires knowledge: a responsible party is one who knew or should have known what they were doing and what the consequences of their actions would be.

An AI system that is systematically miscalibrated cannot satisfy the knowledge condition for responsibility. When an AI system confidently produces a recommendation that is wrong, it "knows"—in the functional sense that its computational processes have produced this output with this confidence—but it does not know that its output is wrong. The organization deploying it may also not know that its output is wrong, because the AI's confident presentation provides no signal of unreliability. The result is that the entire causal chain of AI decision-making—from input through processing to output to organizational action—is epistemically opaque with respect to reliability. No party in the chain can accurately represent what they know and do not know, because the AI system's miscalibration prevents the expression of reliable uncertainty.

This epistemic dimension of the accountability gap has direct legal implications. Legal standards of negligence require that a party have the knowledge or reason to know of a risk for liability to attach. The EU AI Act's accountability provisions require providers and deployers to implement oversight mechanisms that can detect AI failures. The verification of these oversight mechanisms

—the assurance that they are functioning as intended—is precisely what the Verification Tax reveals to be structurally limited.

5.2 The Verification Tax and the Accountability Infrastructure

Paper 2's Verification Tax analysis provides a structural explanation for why the accountability gap persists despite growing awareness of AI's harmful effects. The Verification Tax identifies the resource constraints that prevent comprehensive auditing of AI systems: statistical costs, mechanistic costs, and institutional costs that make thorough evaluation of AI reliability fundamentally expensive. These same constraints prevent the accountability infrastructure that responsible AI governance requires.

If organizations cannot efficiently verify whether their AI systems are behaving as intended—whether they are making discriminatory decisions, whether they are operating within their authorized scope, whether they are producing outcomes that fall within the range of acceptable risk—then they cannot satisfy the oversight obligations that accountability frameworks require. The accountability gap is thus partly a consequence of the Verification Tax: it is not that organizations are unwilling to take responsibility for AI decisions, but that the structural constraints on AI verification make it genuinely difficult to establish the factual foundations for responsibility.

This analysis reveals that addressing the accountability gap requires more than liability rules that allocate responsibility to identifiable parties. It requires the accountability infrastructure—auditing frameworks, verification methodologies, monitoring systems—that enables those parties to know what their AI systems are doing and to demonstrate that knowledge to regulators and affected parties. The Verification Tax implies that this infrastructure will not be efficiently produced by markets or by individual organizational effort; it requires the same coordinated public investment that the economic analysis of AI trust infrastructure identified as necessary.

---

6. The Algorithmic Agency Governance Framework

6.1 Core Principles

Synthesizing the analysis developed in the preceding sections, this paper proposes the Algorithmic Agency Governance Framework (AAGF)—a comprehensive approach to addressing the accountability gap through the reconceptualization of how agency, responsibility, and oversight are understood and allocated in AI-assisted organizational decisions.

Principle 1: Functional Agency Recognition. The framework begins by recognizing that AI systems exercise functional agency—the capacity to make consequential decisions that affect organizational outcomes and third-party interests—regardless of whether they possess the intentionality, understanding, or moral agency that the classical account of responsibility presupposes. This functional agency recognition does not imply that AI systems should possess legal personhood or bear legal responsibility in the manner of human agents. It implies, rather, that the accountability structures designed for human agency must be adapted to govern the functional agency that AI systems exercise.

Principle 2: Accountability Chain Constitution. The framework requires that the accountability chain for AI-assisted decisions be explicitly constituted through governance mechanisms that designate responsible parties at each stage of the AI decision process. This accountability chain must specify who is responsible for AI system selection and deployment, who is responsible for ongoing monitoring and oversight, who is responsible for decisions to act or not act on AI

recommendations, and who is responsible for detecting and remediating AI failures. The accountability chain is not merely a contractual allocation of liability but an organizational governance structure that must be documented, disclosed, and subject to board and regulatory oversight.

Principle 3: Meaningful Human Authority. The framework requires that human agents retain meaningful authority over consequential AI-influenced decisions—the authority not merely to approve or override AI recommendations but to understand the basis for those recommendations, to evaluate whether the AI's reasoning is appropriate for the specific decision context, and to bear accountability for the decisions they make with AI assistance. Meaningful human authority requires that the MIRROR-identified miscalibration problem be addressed through calibration assessment and disclosure, that AI confidence information be presented in formats that support rather than supplant human judgment, and that organizational processes ensure that human decision-makers have the competence and incentive to exercise genuine rather than nominal oversight.

Principle 4: Proportionate Monitoring Infrastructure. Recognizing the Verification Tax constraints on comprehensive AI auditing, the framework requires proportionate monitoring—intensive monitoring of the highest-consequence AI applications and lighter-touch monitoring of lower-risk deployments—with monitoring investments allocated to the AI applications where the accountability gap would be most consequential if it resulted in harm. Monitoring infrastructure must include both technical verification methods—performance tracking, fairness auditing, drift detection—and governance mechanisms—board oversight, regulatory reporting, affected party disclosure—that together constitute the accountability evidence base.

Principle 5: Remediation and Learning Systems. The framework requires that organizations implement remediation and learning systems that ensure that when AI accountability failures occur—when AI decisions cause harm or create unacceptable risks—organizations can identify the source of the failure, remediate its consequences, and prevent recurrence. These systems must be designed with attention to the distributed causality of AI failures, ensuring that accountability for failures can be attributed to the party or parties whose governance failures enabled the harm.

6.2 Application to Human Resource Analytics

The application of the Algorithmic Agency Governance Framework to human resource analytics—the domain examined by Paper 3—illustrates its practical operation. Under functional agency recognition, organizations must acknowledge that AI hiring systems exercise genuine agency in shaping who gets hired, promoted, or terminated, regardless of whether the formal employment decision is made by a human manager. Under accountability chain constitution, the organization must document who made each governance decision in the AI hiring process—which party selected the system, who approved its deployment, who monitors its ongoing performance, who decides when to override AI recommendations, and who bears accountability for hiring outcomes that result from AI-assisted decisions.

Under meaningful human authority, organizations must ensure that HR professionals who override AI hiring recommendations have genuine rather than nominal authority—the competence, information, and organizational support necessary to exercise independent judgment. Under proportionate monitoring infrastructure, organizations must implement systematic auditing of AI hiring decisions across demographic groups, with verification methods calibrated to the risk level of the specific hiring context.

Under remediation and learning systems, organizations must establish processes for investigating AI hiring failures, identifying their sources in the accountability chain, and implementing changes that prevent recurrence. When an AI hiring system is found to produce discriminatory outcomes, the accountability chain must support the identification of which governance failures—system selection, deployment authorization, monitoring, override decisions—contributed to the discriminatory outcome and what remedies are appropriate.

---

7. The Regulatory Design Challenge

7.1 Designing for Accountability in AI Governance

The analysis developed in this paper has significant implications for regulatory design. Existing regulatory frameworks—developed for human agents and organizational structures—provide inadequate tools for governing AI agency. New regulatory concepts and institutions are required that are specifically designed for the properties of AI decision-making.

Mandatory algorithmic impact assessment requirements—analogous to environmental impact assessment requirements—would require organizations to evaluate and disclose the potential impacts of their AI systems before deployment. These assessments would examine the accountability structures surrounding the AI system, the potential for discriminatory or harmful outcomes, the monitoring and oversight mechanisms in place, and the remediation procedures available when harm occurs. By making accountability infrastructure a precondition for AI deployment rather than an afterthought following harm, impact assessments can shift the accountability dynamic from remediation to prevention.

AI regulatory sandbox programs—controlled environments in which AI systems can be deployed and tested under regulatory supervision—can provide the evidence base for developing appropriate accountability frameworks. By observing how AI systems behave in real-world contexts under varying accountability structures, regulators can develop empirically grounded standards for what accountability requires in different AI application contexts.

International coordination on AI accountability standards is essential given the global nature of AI development and deployment. Divergent national accountability frameworks create both compliance costs for organizations operating across jurisdictions and regulatory arbitrage opportunities that undermine accountability protections. International agreement on minimum accountability standards for high-consequence AI deployments—along the lines of the EU AI Act's risk-based approach but with explicit attention to the accountability gap—would establish a level playing field while ensuring that accountability protections are not sacrificed to competitive pressures.

7.3 The Institutional Dimension: Beyond Liability Rules

The analysis of AI accountability has thus far focused primarily on liability rules—the legal standards that determine who bears responsibility when AI decisions cause harm. Yet the deeper accountability challenge is institutional rather than merely legal. Liability rules determine who pays after harm has occurred; they do not create the systems through which harm is detected, prevented, and remediated in real time. The accountability gap reflects not only the absence of appropriate liability rules but the absence of the institutional infrastructure through which accountability is constituted and sustained.

The institutional dimension of AI accountability has several components. First, there is the infrastructure for monitoring AI decisions in real time: the systems that detect when AI systems are producing biased, unsafe, or otherwise problematic outcomes before those outcomes cause harm. Second, there is the infrastructure for investigating AI failures: the forensic capabilities that can trace AI decisions to their causal origins in the complex chain from data through algorithm through deployment context. Third, there is the infrastructure for remediating AI harm: the organizational processes and legal mechanisms through which affected parties can obtain redress when AI systems cause them injury. Fourth, there is the infrastructure for learning from AI failures: the systematic processes through which organizations and regulators update standards, practices, and requirements based on the evidence generated by AI accountability experiences.

Each of these institutional components faces the Verification Tax constraints that Paper 2 identified as fundamental barriers to comprehensive AI oversight. Real-time monitoring of AI decisions is expensive and technically challenging, particularly for the complex neural network architectures that characterize state-of-the-art AI systems. AI failure investigation requires the same causal tracing capabilities that mechanistic auditing constraints render difficult. Remediation infrastructure requires the allocation of resources to AI accountability that competitive pressures may prevent organizations from making voluntarily. Learning systems require the collection, analysis, and dissemination of accountability evidence that the Verification Tax prevents from being efficiently produced.

The institutional dimension of the accountability gap implies that addressing AI accountability requires more than the specification of liability rules or the imposition of regulatory requirements. It requires the deliberate construction of accountability institutions—new organizational forms, new professional roles, new governance mechanisms—that constitute the infrastructure through which AI agency can be governed. These institutions do not currently exist in adequate form, and there is no market mechanism that will produce them at the pace or scale that the AI deployment expansion demands.

The creation of AI accountability institutions thus represents a public investment challenge analogous to the creation of other institutional infrastructure—financial regulatory institutions, environmental monitoring agencies, product safety bureaus—that markets alone will not produce. Public investment in AI accountability institutions—auditing firms, regulatory agencies, affected party advocacy organizations, AI safety research—is necessary to create the institutional foundation on which the accountability gap can ultimately be closed.

7.2 Beyond Liability: The Institutional Dimension of AI Accountability

The analysis developed in this paper reveals that the accountability gap cannot be resolved through liability rules alone. Liability rules allocate responsibility after harm has occurred; they do not create the institutional infrastructure through which harm can be detected, prevented, and remedied. The accountability gap requires institutional innovation—in new organizations, new governance mechanisms, and new forms of oversight—that together constitute the accountability infrastructure for AI agency.

Several institutional innovations merit serious consideration. AI auditing institutions—specialized organizations with the technical and legal expertise to evaluate AI accountability—could perform the verification function that the Verification Tax reveals is systematically underproduced. These institutions could be modeled on financial auditing firms: independent, credentialed, subject to quality standards, and empowered to issue accountability certifications that organizational stakeholders can rely on. Public AI regulatory agencies—at the national or international level—could provide the governmental oversight that the accountability infrastructure requires, with authority to set standards, conduct investigations, and enforce accountability obligations.

Affected party representation—the formal involvement of those harmed by AI decisions in the accountability process—is conspicuously absent from current governance frameworks. Individuals who suffer discrimination from AI hiring systems, who receive negligent AI medical recommendations, or who are harmed by AI financial decisions currently have limited formal role in the processes through which AI accountability is determined. Institutional innovations that empower affected parties—through collective action mechanisms, class certification for AI-related harms, or statutory standing provisions—would strengthen the demand-side of accountability by ensuring that the parties most harmed by the accountability gap have the incentives and tools to demand remediation.

---

8. Conclusion

This paper has argued that the accountability gap—the absence of clear legal and organizational structures for assigning responsibility for AI-influenced decisions—is a structural consequence of AI's functional agency that existing legal and organizational frameworks were not designed to address. The paper has developed this argument through three foundational references. The MIRROR benchmark (Paper 1) reveals that AI systems exercise consequential agency while systematically misrepresenting their own reliability, creating an epistemic void at the center of the accountability analysis: no party in the AI decision chain can accurately represent what they know, because the AI system itself does not know. The Verification Tax (Paper 2) reveals that the accountability infrastructure necessary to fill this void—auditing, verification, monitoring—is structurally underproduced due to the inherent costs of AI evaluation, implying that the accountability gap will persist without coordinated public investment in accountability institutions. The strategic HR analytics framework (Paper 3) demonstrates how these abstract accountability failures manifest concretely in organizational contexts where AI decisions shape people's careers and life chances, and where the absence of accountability mechanisms creates both organizational risks and individual harms.

The Algorithmic Agency Governance Framework proposed in this paper provides a practical pathway for addressing the accountability gap through functional agency recognition, accountability chain constitution, meaningful human authority requirements, proportionate monitoring infrastructure, and remediation and learning systems. Yet the framework's implementation requires institutional innovation that extends beyond what any individual organization can accomplish alone. Addressing the accountability gap requires the same coordinated public investment and international cooperation that the analysis of accountability infrastructure has identified as necessary: new regulatory standards, new institutional forms, and new legal concepts that are specifically designed for the accountability challenges of an age in which consequential decisions are increasingly made by entities whose agency is computational rather than intentional.

References

1. Buçinca, Z., et al. (2025). Between transparency and trust: Identifying key factors in AI system perception. ACM Transactions on Interactive Intelligent Systems.
2. Solaiman, I., & Talhouk, R. (2024). Legal personality of AI and robotics: Design-based approaches. Springer Nature Computer Science Reviews.
3. Wang, J. Z. (2026). MIRROR: A Hierarchical Benchmark for Metacognitive Calibration in Large Language Models. arXiv preprint arXiv:2604.19809.
4. IEEE Standards Association. (2025). Ethically aligned design of autonomous and intelligent systems: Governance frameworks. IEEE.

5. Bradford, A. (2024). The Brussels Effect, regulatory diffusion, and AI governance. Columbia Law Review.
6. Wang, J. Z. (2026). The Verification Tax: Fundamental Limits of AI Auditing in the Rare-Error Regime. arXiv preprint arXiv:2604.12951.
7. 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.
8. Stilgoe, J. (2024). The politics of AI governance. In AI Governance. Oxford University Press.
9. European Commission. (2024). The EU Artificial Intelligence Act. Official Journal of the European Union.
10. Chander, S., et al. (2025). Algorithmic accountability and transparency in public sector decision-making. Government Information Quarterly.
11. Zeng, Y., Lu, E., & Huangfu, C. (2024). Linking artificial intelligence to robotics: Legal and policy implications. Science and Engineering Ethics.
12. Smuha, N. A. (2024). From a governance-framework to a regulation: The EU AI Act. Computer Law & Security Review.
13. Kahrl, A. W. (2025). Disability, AI, and employment discrimination: A legal and organizational analysis. Berkeley Journal of Employment and Labor Law.
14. Yeung, K., & Lodge, M. (2024). Algorithmic Regulation: A Critical Appraisal of EU Data Protection and AI Governance. Oxford University Press.